Multiobjective Optimization in Models of Synthetic and Natural Living Systems

Luís F Seoane

TESI DOCTORAL UPF / ANY 2016

DIRECTOR DE LA TESI
Ricard Solé Departament Department of Experimental and Health Sciences

For all those that have been obliterated by the machine despite their effort and merits.

Acknowledgments

A walk on part in the war

Michael Crichton’s sequel of *Jurassic Park* starts with a talk by Ian Malcolm at the chapel of a former convent in Canyon Rd, Santa Fe. Ian Malcolm is the cool scientist wearing a leather jacket and sun glasses. He is the one talking about chaos theory and predicting that the dinosaurs – all of them female – might be able to adapt and reproduce because, you know, “life finds a way”[1]. The convent had been repurposed into the Santa Fe Institute (SFI) a few years ago. In it, Malcolm explains his theory about life at the edge of chaos. Malcolm is a fictional character, but *Life at the Edge of Chaos* is for real and almost a dogma within the walls of the institute: “Complex systems tend to locate themselves at a place we call ‘the edge of chaos’. We imagine the edge of chaos as a place where there is enough innovation to keep a living system vibrant, and enough stability to keep it from collapsing into anarchy. It is a zone of conflict and upheaval, where the old and the new are constantly at War.”[2]

The convent in Canyon Rd was the first physical location of the Institute. Now it has been moved to the outskirts of Santa Fe, in a hill at Hyde Park Rd. I arrived there for the first time in early May 2013 together with Ricard Solé, Sergi Valverde, and Max Carbonell; friends, supervisors, and work colleagues at Pompeu Fabra’s Complex Systems Lab. I can’t remember anything from those first days. I was inebriated with

[1] https://www.youtube.com/watch?v=SkWeMvrNiOM

[2] Crichton, M., 1997. *The lost world*. Random House.

emotion and height. At 2000 meters above the sea level the sky shines bluer, the warm air is warmer and every afternoon feels like a watercolor painting. The Institute is entrenched on the rough brown sand and surely it is alive and has got roots growing underneath its glass and adobe foundations[3]. My lips and knuckles dried, I had a mathematical fever and a vision of equations until I woke up again half a week later – and I woke up to never be the same.

The Institute had been established in the mid 80s. A group of scientists led by George Cowan started an intellectual journey that few understand even today: a multidisciplinary research institution to defy reductionism and bypass traditional academic hierarchies. Reductionism is the most successful school of thought in science. It consists in breaking a system down to its smallest components and analyze that bottom level to attain knowledge about the original scale. But let's take a look at the science looming in this XXI century: a hyperconnected world with unpredictable emerging behaviors in economics or politics; the unsolved questions in biology about the origin of life, the development of complex structures such as organs and tissues from tiny little cells; behind all that coordination we like to guess large scale computations implemented by abstract networks of genes and molecules that *dialog* with each other; meanwhile, those mysteries that are the human language and intelligence, or the propagation, mutation, and persistence of culture. All of this still demands an explanation. And all these are problems that, we think, “cannot be solved looking at their parts with each time a better resolution”[4]. The Institute's approach to nature is systemic: the key is not in the components, but in the interactions among them. The smaller details might not be relevant. Instead, we seek the surprise that materializes as very simple rules co-evolve exploring the unexpected.

My mates from the Pompeu Fabra University left after three weeks. I

[3]The city of Santa Fe imposes that all its buildings imitate the shape and color of adobe as much as possible. The institute indulged a little bit and eventually became a cubist box of glass and colored concrete that offers an austere impression.

[4]http://santafe.edu/support/the-history/30years/ (http://santafe.edu/support/the-history/)

stayed for another month to engage in the summer school of the Institute. I cried the first and the last days. The SFI summer school is a sacred place in space-time. We were lodged in the sober dorms at St. John's College, also at the outskirts of Santa Fe and also grounded against a hill with the same deep earth and breathtaking blue sky. During the night, the coyote would howl behind the gaunt bushes of the dessert and during the day, our heads were burning pregnant with ideas. Each Friday we would visit the Institute and we would mess around the minuscule library where Cormac McCarthy had written The Road[5]. We would discuss impossible theories on the glass walls of the diaphanous inner yard, or all over the windows from which New Mexico appeared infinite and pink, thrashed by that summer's savage fires.

Every waking minute and person were sublime and deserve to be remembered forever. But if I have to pick up a moment of that summer I choose UC Berkley's John Harte's lecture. The slides were quite schematic. I followed them with ease from the back of the hall. One by one, very slowly. Outside, a huge dry storm unfolded with bolts and thunders that lightened the room here and now and projected the colors of the faraway fires against the thick, hanging curtains. Little by little the equations and the questions so little by little the whole theory and the experimental data. When I realized it, I had just attended to the explanation of ecological systems in equilibrium. I mean: The Theory of Ecological Systems in Equilibrium! And I felt a joy that, I though, must be the same that any kid would have felt half a century ago listening to Maxwell invent electromagnetism. And I cried again.

The Santa Fe Institute are the interactions between its people. There are around one hundred *external faculty* – among them Ricard Solé, my PhD supervisor and my link to that institution – and around fifty resident researchers. There are talks and discussions during lunch and at three o'clock tee is served. Then, people gather to debate new science and ideas; that abstruse mathematical detail or that AI in the computer. Around there were Bernat Corominas-Murtra – who was briefly my comrade in Barcelona – and his colleagues Rudolph Hanel and Stephan Thurner, who are

[5]McCarthy, C., 2009. *The Road*. Pan Macmillan.

busy bringing Statistical Mechanics to a new stage. There was also Steve Lansing, who figured out Bali's religion to discover a socioecologic system in perfect order since centuries ago[6]. And many more, but above all there was Murray. A delicate, venerable elder; gentle and smiling. I could shake his hand but I barely did it out of fear and respect. A living legend of the XX-th century, the very last giant, Murray Gell-Mann: discoverer of the quarks and the scientist who named them. One of the founders of the Institute and a personification of its multidisciplinar spirit. After being awarded the Nobel prize in physics he turned his attention to human language to find the ultimate root of all tongues.

Stuart Kauffman, one of the first guardians of the young Institute and its manifesto, was not there. I remember with devotion when I learned about his work for the first time. It was through his book *At Home in the Universe*[7] right when I arrived to my new home in Barcelona. After reading it, my way of being and the way I look at the world had changed forever. Kauffman brought me an unexpected light to the problem of the origins of life. His autocathalitic cycles had been getting assembled in a chemical background for hundreds and thousands of years while the earth wasted its geological ages, little by little – just as John Harte's ecology, once again little by little, just like the good science – quite little by little inert matter got self-assembled, almost surely following those Maths that Kauffman had shown to us, catalizing here and there a spark of molecules that flickered, then faded away; inorganic matter got pregnant, pregnant with the incipient until it couldn't bear it anymore and it discharged as an unavoidable tide, exaggerated, and organic. Self- organized life came over the earth to stay now and forever.

I went back to Santa Fe that November invited by Simon DeDeo with the remote mission of applying Statistical Mechanics to social systems –

[6]Lansing, S., 2012. *Perfect Order: Recognizing complexity in Bali*. Princeton University Press.

[7]Kauffman, S., 1995. *At home in the universe: The search for the laws of self-organization and complexity*. Oxford University Press.

just as Isaac Asimov's Hari Seldon[8], and as many researchers are trying to do all over the world. We believe that it is possible. We believe that the Institute is the best place to try it. The heath of the summer had faded and towards the end of that month if snowed as if we were in the North of Europe. The excitement from June had yielded to a quiet calm, a fallow that some day will sprout. During that second trip I could enjoy more of the singular city that is Santa Fe and the very odd culture that survived there as the Spanish empire broke apart. Many there believe themselves the heirs of that despotic and murderous *grandeur* and they still await the comeback of their ancestors avoiding mixing up with the aborigines or the wetbacks, whom they despise. Their roots have already been engraved by fire on that American arid soil. Looking back at certain names – Camino la Tierra, Senda de los Espectros, El Cielo del Oeste – these are just petty backstreets or meanders in the road, wastelands where many people live in disgrace, impoverished and obliterated by a time that doesn't exist anymore.

Summertime epilogue

This happens to me every time that I'm going back home. It happened when I moved from Berlin for a stage at the science department of the University of Granada, and it happened again when I left Granada back for Berlin. (Time is so finite!) I still project my self, opening new lines towards the future. Each paper, each research article that I come across is a dormant idea – a revolution perhaps? – that rests within the ingenuous pages. (I wanna embrace them all and time is so finite!) I try, but every day I must set aside – in a promise that I will return – many of them that I cannot possibly read. A pile grows at the background of my desk. Each night before going to sleep I glide my tired hand over the articles to measure them; I lay my face on them, smell the ink and feel all that science beating inside. And one day, of course, I must move out. Leave Berlin, leave Granada; after two months, leave Santa Fe.

[8] Asimov, I. *Prelude to Foundation*; *Foundation*; *Foundation and Empire*; *Second Foundation*. Random House LLC.

I tried to pack all my papers in the suitcase. I brought them from New Mexico to Washington, where I met my boyfriend. Then we went to New York and I stored the suitcase, closed, in the little room we had rented in East Harlem. Kids die there; many every day, the press said. There the summer was humid and worn-out, and the guts of the Big Apple roared with an organic hum made of concrete, screws, and nuts. I intended to continue our trip with the suitcase filled up. My boyfriend convinced me to open it. Then pretended to be angry, but I know he was laughing inside. Among the souvenirs, the books and postcards, the dirty clothes, some baseball balls... there they were, all those papers accumulated in Santa Fe. Ricard Solé's works on cellular computing, or Sergi Valverde's on technological evolution. The first chapter of Dayan & Abbott's Theoretical Neuroscience, Axelrod's model of cultural dissemination, Jim Crutchfield's epsilon machines and other names of weird theories born for the first time at the Santa Fe Institute. I took them out to my regret, approximately two kilos, and I knew I had to let them go. As we came out of the building at the Latin Harlem, I dropped them piece by piece on the filthy, sacred corners of New York. I left them behind. Now that I'm gone perhaps a gust of wind passes by and scatters them all over. And in that clutter, maybe one of the kids that die every night can trap them – trap those papers, all the equations and science they contain – and fly away with them.

Afterwords to *A walk on part in the war*

The previous text is a love letter to a world that I always wished that it existed. It is a translation from Spanish of an article published at the journal of the Alumni association of the Pedro Barrié de la Maza Foundation[9]. Like many other elements in this thesis, the title of the article is taken from Pink Floyd. A verse of *Wish You Were Here*[10] captures what these four years felt to me: a walk on part on the war, at the forefront of science, exploring a territory unknown to us, maybe right, maybe wrong, risking to make a petty contribution, yet daring to explore what others

[9](Number **17**, pp.68-71 (2014).

[10]“Did you exchange a walk on part on the war by a lead road in a cage?”

deemed uninteresting, proposing ideas that many think are mistaken and lead nowhere. Science is a wager and we bet blindly. It might take years to know the outcome and we might lose all at once. To foster science, a world is needed where discussion is free, where others put their resources and prestige on the line so you can take a chance, where the struggle is aimed at that cloudy $10\,\%$ of problems that are worth considering; right at the frontier where new meaning is being born, as this thesis proposes, by the Darwinian dynamics that can wipe all of your effort in one fell swoop. A world that I always wished that it existed.

And it does, and I was generously introduced to it by the supervisor of this PhD Thesis, Ricard Solé. I hope I could make justice to all what he meant to me during these years – both at the personal and professional levels, if I could trace the line between the two. To him I owe a myriad opportunities and a vision of the world that is precise, original, revolutionary at times. I also owe him several moral values that might be dangerous to keep as the world moves in darker ages: an unshakable compromise with science and honesty above every thing, with a very especial care for the people that follow his lead on this walk on part in the war. I wish I could make justice to this, but I can’t and I will never can. Let the little text above, that *walk on part in the war* be my last humble attempt.

These words must be extended to my colleagues at the Complex Systems Lab. I think all of us know we are participating of something singular and we often pay the price for it: little sleep hours, the quirky relationship of our country with science, etc. Against every odd, the Complex Systems Lab feels like that little corner of the Santa Fe Institute that they forgot to bring to New Mexico. In creating that feeling, Sergi Valverde must be acknowledged first. Besides his inspiring scientific work and persona, I was lucky enough to share trips to Santa Fe and Arizona with him where we could explore this weird world of ours. In one or another of these trips to Santa Fe, Max Carbonell, Salva Durán, and Raúl Montañez were by my side, dodging shots in the desert, seeking bears or coyotes in the night, enjoying and suffering our cholesterol rich diet. I didn’t have the chance to travel with Aina Ollé, Josep Sardanyes, Carlos Rodríguez-

Caso, Nuria Conde, Ben Shirt-Ediss, Eva García, Jordi Piñero, Adriano Bonforti, or Javier Macia. But life is long and wide and I look forward to the next chance. I admire your tenacity, impressive originality, and hard work; and I'm eager to discover the science that you are bringing to the world. I didn't have the chance to travel with Dr. Amor either, but we will be meeting soon at the best place possible. Also, I cannot forget Bernat Corominas-Murtra, a great inspiration since the day I arrived to the Complex Systems Lab.

All these people, in one way or another, have contributed to the unique intellectual place that we came upon. They will need to excuse that I keep these words shorter than they deserve, but time is pressing and finite. To all of them, companions in this walk on part in the war for a long or short time, I dedicate the text above too.

As I move away from the Complex Systems Lab I cannot forget those who helped me along the way, starting by Jorge Mira who trusted my criterion since very early. I am glad we keep our research line open and look forward to what the future can bring.

Without Benjamin Blankertz it would not have been possible to make the machine that reconstructs images from the mind. That is an impressive piece of science fiction come truth.

I always look back to my time in Granada, were I left a great part of me. Thanks to everybody there, but specially to Miguel Ángel Muñoz, Samuel Johnson, Sebastiano de Franciscis, Jordi Hidalgo, Virginia Domínguez, and Paula Villa, I have become a much better scientist and person. I look forward to a future in which you all are present, in one way or another.

I wish to thank Sergio Espeso for his friendship during these last months that I have spent in Barcelona. I hope we meet again in the future! I'm also eager to see your research and what you bring into this world. I do not doubt that it will be magnificent!

Finally, as I approach my new scientific home, I must thank Max Tegmark for the amazing opportunity that he has put in front of me. I will work as hard as possible to get the most out of it!

Europa

Three weeks before this PhD thesis was due my friends and flatmates Marc Marin and Pedro García booked me a flight to Berlin and shipped me away for 24 hours. That was a terrible day for a vacation. I was having a hard time putting together my research during the last years and my writing was not at its best. I think I couldn't have finished this thesis without that trip! I am glad that Marc came with me, and sorry that Pedro did not – and I promise we will fix that some day.

Coming to Berlin for a last time made me remember this funny impression that I have about Europe. I imagine it as a chess board, with my friends occupying distant squares: Javi in Poland, Gonzalo in Brussels, Julita in Barcelona, Javiño in Madrid, Zoran in Croatia, Irati and Ane in Berlin, Borja and Edurne in Frankfurt, Conchita and Fran back in the Canary Islands, ... From time to time we move around. Marc gets his van towards Berlin, Gonzalo comes back from Brussels, I get a train direction nowhere. There were those days in the past when many of my friends were crossing this worn-out chess board at the same time, and every kind of adventures would happen to them. I would picture this huge game unfolding and those anonymous heroes of the XXI century arriving safely to our meeting point, often Berlin or Barcelona. And the magnitude of the game was such that I couldn't believe how big we had become, how insignificant we had made time and space. During those 24 hours I had that feeling back and I was glad to meet some of those old friends (Irati and Ane) and some new ones (Marcos). Again, this thesis wouldn't have been possible if Marc and Pedro would not have kidnapped me, and therefore I will remain forever thankful.

Two weeks after that I asked them to produce the cover of this thesis. One night I left Pedro working on it and I went to the bed. I dreamed three different dreams in which Pedro was coming over with alternative versions of the cover. In the second dream, the cover looked suspiciously similar to the actual one, but a strange grid was printed on it, dividing the drawing in little squares. “What is that”, I would ask in my dream. Pedro would cut out one of those squares and hand it over to me saying:

“there you go, the ticket for your flight”. Now I understand, that dream was a prophecy of our 24 hours away in Berlin, only time is twisted and past and future is not what it looks like. However these last weeks went by, time got back into a straight line again, and Pedro, Marc, and Irene completed the astounding cover of this PhD Thesis in a few days. There is no gratitude that can pay such an amazing work! For it I'll be forever indebted!

Remember that space and time across the chess board that we could make vanish? Every now and then those distances across Europe become huge again. Barcelona is unfairly apart from Brussels, and hence I'm unfairly apart from Gonzalo Heredia. To him I owe the happiest and most magical years of my life. Those kind of moments that once they have happened they exist and will persist forever. I cannot forget a single one of them, from the day we met until the terrible instant that we moved away from each other, whenever that was. Now we sail on our own, discovering the world that we are so eager about. I hope we meet again, and we can tell each other endless tales of people and cities. Even if we don't walk side by side anymore, you are forever my fellow traveler in the greatest adventure possible.

Everywhere in Europe if you look up to the sky there's a Milky way that tells the direction home. I can never be lost. Just up my head, and there it is, as magic powders or a destiny engraved in the Universe. The day I deposit this PhD Thesis I will take (I will have taken) a train that follows that shiny path towards the west. People complain about trains in Spain, especially this one. It takes 13 hours to make ~ 1000 kilometers. It stops everywhere it can and zigzags along the way visiting impossible towns as the night passes by. I think those complaining people didn't understand what this train is about. Between 6 or 7 in the morning it reaches Galicia, the direction home that the Milky way is telling us about. It enters from the south, stepping on the rails of the first train that ever entered in Galicia (that Iron Madonna that Curros Enríquez sang about more than a century

ago[11]). The tracks are deep within the valley of the Miño river, and so close to the water that at times it feels like we are floating and that the train is sustained only because the people inside keep dreaming. We ride covered by a lush glade and the first lights of the morning fall over the valley and reach inside the train in straight rays that wake us up from our drowsiness. You get dizzy with the intense yellow and green and the babbling of the train and the water. You dream, keep dreaming so the train doesn't sink. Or not, because in this mystical world there is little distinction between sleeping and being awake. And then you get home.

I get home, and I get to thank my mother for all her effort during the years, all her faith in my research and all her patience with me and with my absences. The work that I present in this PhD Thesis is hers as much as mine. To my mother I owe a chance in this world – and what a great one!! Every step I take is possible only thanks to her, and every time I wish I were back home to tell her, to share the wonders of the universe that she missed so that I could get this chance. Now my adventure continues, again far away from her, and I wish more than never that distance does not exist and that time could be stopped for a while or forever. Besides her, as always, my grandmother. Together with my grandfather, they were my home and my childhood. They were my inspiration and the example of how a person has to be in this life. And to them I will keep coming now and forever with endless gratitude.

Because of time constraints I cannot say everything that I need to say about my family. Especially to my cousins David, Manuel, Pili, Lino, Mechi, Cristina, Lucas, Teo, Tubío, and Belén... I will be to them forever indebted for making up such great persons and making my trip back home so desirable. Of course, my aunts and uncles Chiruca, Beli, Marcelino, Che, Paco, and Elena are always present in my thoughts.

[11] Curros-Enríquez, M. Na chegada a Ourense da primeira locomotora.

Abstract

This thesis tackles a series of phenomena across diverse fields from the point of view of optimization. The problems studied often bring together diverse conflicting constraints that cannot be fully satisfied. Out of these conflicts emerges a notable complexity. Pareto optimality theory deals with the optimization of simultaneous targets and was consequently used as a preferred framework. In this thesis, it is shown how mathematical properties of Pareto optimal designs affect very diverse complex systems in universal ways. The implications for the different phenomena studied are discussed. Especial attention was paid to the optimization of complex networks, the optimality of communication codes and human language, and the tradeoff between complex biological structures versus the replication efficiency required by Darwinian dynamics.

Resumen

Esta tesis estudia una serie de fenómenos de diferentes campos desde el punto de vista de la optimización. Los problemas estudiados suelen imponer a sus soluciones una serie de requisitos que se contradicen entre sí, y que, por lo tanto, no pueden ser del todo satisfechos. De este tipo de conflictos emerge una complejidad notable. La teoría de Pareto optimalidad trata, en efecto, sobre la optimización de varios objetivos contradictorios. Este es el marco teórico seleccionado para estudiar la mayoría de los problemas. En esta tesis se muestra cómo las propiedades matemáticas de los diseños Pareto óptimos afectan de manera universal a sistemas complejos muy diversos. Las implicaciones en cada fenómeno concreto son discutidas. Se presta especial atención a la optimización de redes complejas, a las características óptimas de posibles códigos de comunicación, y al compromiso que emerge entre estructuras biológicas complejas frente a la eficiencia en la replicación impuesta por la selección Darwiniana.

Preface

Many works in complex systems start out with a clarification of what such systems are. When attempting to publish papers with multidisciplinary research, it is very usual that referees request an introduction of what complexity is; oftentimes we end up with repeating commonplaces about the field. From time to time an expert shows up and says: “No one really knows what complex systems are. Each one has its own definition. No general agreement exists among scientists on a measure of complexity. There isn’t a complex systems science really!”

We acknowledge the difficulty of unifying or even defining complex systems, but we also reject complex systems as a dogmatic platform from where to study a delimited series of phenomena. Instead, we believe that *intriguing* aspects of nature lay out there demanding an explanation – whether they comply or not with some minimal requirements of computational (or whatever rule of) complexity. We think that good scientific questions abound, and that we (already two decades into the XXI century) are equipped with a series of tools and concepts that can answer them despite the different fields that these questions originate in.

Hence, this thesis shall not be about complex systems. Instead, it is about a series of astounding (and not fully understood yet) aspects of Nature that still today capture our imagination and curiosity. Some of the issues that we attempted to tackle during these last years are: i) How can we bring together computation theory and biology? How can we cope with (or benefit from) the limits imposed by the physical substrate of computation? Is it possible to come up with universally efficient design patterns? ii) How does all of this come together into a cell, a paramount compu-

tational and functional unit in biology? Precisely, can we come up with a sound computational theory of cell differentiation? How do different optimality constraints determine the coding of such a complex structure? iii) What is human language? How does it emerge? How is it different from other communication mechanisms? What is the role of abstract grammatical particles? Are they necessary and/or sufficient for complex communication to arise? Are there any efficiency constraints similar to those found elsewhere? Do they have similar effects on communication systems? iv) How do Darwinian dynamics incorporate the information of the environment? Does this distinguish between meaningful and meaningless bits of information?

Needless to say, our inquiries in some of these questions could not be included in this PhD thesis.

We do not care whether the underlying systems and models investigated are complex or not. We care that they present relevant, unsolved questions. We do find a series of recurrent topics and this is what prompts us to wonder whether similar mathematical descriptions (shall they be complex or not) might tell us something about many of these problems at once. We have chosen to look at these issues through the tradeoffs and conflicting constraints that all these systems seem to present. The chosen paradigm hence is Pareto (or multi objective) optimality (PO or MOO) which functions as a framework to guide our research.

This thesis begins with an introduction to Pareto optimality and to its tools and concepts in chapter 1. We present a few recent examples that illustrate how other researchers have also found PO useful in advancing their understanding about diverse systems.

We found that there is a deep connection between Pareto optimal and thermodynamic systems. As will be discussed below, the Pareto front (which constitutes the solution to MOO problems) is equivalent in most relevant aspects to the Gibbs surface, a geometrical object whose shape and differential geometry dictates the phases, phase transitions, responses to perturbations, and critical points in statistical mechanics. This connection stems from the application of Lagrange multipliers to the Pareto

front, but it had not been noted before in the literature to the best of our knowledge. We made this connection with the Gibbs surface and phase transitions explicit in two papers [282] and [286], and we brought in criticality in a third one [284]. The former two, and the first half of the later paper constitute chapter 2.

Despite the parsimony of this theoretical framework and its sound correspondence with statistical mechanics, we found great opposition (for different and often contradictory reasons) to the ideas in [282]. To bypass these criticisms, and as an application of our theoretical framework, we worked out a well-defined case study of our findings using complex networks [285] by studying their tradeoff between efficiency and economy. This problem is relevant not only to make our point, but also in relation to the engineering of very diverse real systems from railroads to biologically efficient structures. Besides, this research allows us to illustrate a very important idea: that Pareto optimality must *always* drive a series of systems towards a critical state [284]. All this constitutes the matter of chapter 3, which ends with a brief review of Pareto optimal systems found in the literature from the perspective introduced in chapter 2.

Chapter 4 tackles efficiency aspects of communication, with special emphasis on human language. For the first half of the chapter we made an extensive literature review ([301]) to support that, against our naive intuition, ambiguity might play a crucial role in achieving communicative efficiency. This idea is further researched in a (still unpublished) paper [288] that contextualizes human language within a morphospace of all possible communication codes. This constitutes the second half of chapter 4.

A second attempt at understanding human language is made in chapter 5. This alternative view considers a systematic approach based on the tradeoff between the information that we can capture about a phenomenon (language as a symbolic dynamics in this case) and the simplicity of the model that we use. To do this we need a few concepts from symbolic series analysis that we introduce in that same chapter. This research is in its early stages and will be continued by this author in the future.

The relevance of the tradeoff between ‘efficient inference’ and ‘sim-

plicity' goes beyond mathematical modeling, as discussed in chapter 5. In [287] we argue that biological living systems are the ultimate *builders of models* about the real world; and that while comprehensive models offer a clear advantage for survival, Darwinian dynamics introduces a conflict because fast replication is highly rewarded. This makes up chapter 6, which is our effort to bring together Darwinian evolution and information theory (an effort shared with many other authors). We find interesting insights about information theoretical limits and potential triggers of complex life, as well as an important connection with philosophical ideas about the origin of meaningful information through Darwinian evolution.

Because of the diversity of the work developed during these years, the hypotheses of this thesis are not presented at its beginning, but before each chapter so that they can be better contextualized within each corresponding topic. We wrap everything together in chapter 7, where we discuss briefly the main ideas presented and where we assess whether the different hypothesis have been resolved and how.

The list of papers elaborated for this PhD thesis is the following:

- Seoane, L. F. and Solé, R., 2016. Information theory, predictability, and the emergence of complex life. Submitted.

- Seoane, L. F. and Solé, R., 2016. Exploring the morphospace of communication codes. In preparation.

- Seoane, L.F. and Solé, R., 2015. Phase transitions in Pareto optimal complex networks. *Phys. Rev. E*, **92**(3), p.032807.

  Available at:
  http://journals.aps.org/pre/pdf/10.1103/PhysRevE.92.032807

- Solé, R.V. and Seoane, L.F., 2015. Ambiguity in language networks. *Linguist. Rev.*, **32**(1), pp.5-35.

  (Available at:

http://www.degruyter.com/view/j/tlir.2015.32.issue-1/tlr-2014-0014/tlr-2014-0014.xml)

- Seoane, L. F. and Solé, R., 2015. Systems poised to criticality through Pareto selective forces. arXiv preprint arXiv:1510.08697.

  Available at:
  http://arxiv.org/abs/1510.08697

- Seoane, L. F., Solé, R., 2015. Multiobjective optimization and phase transitions. In *Proceedings of ECCS 2014*, ch.22.

  Available at:
  http://arxiv.org/abs/1509.04644

- Seoane, L. F. and Solé, R., 2013. A multiobjective optimization approach to statistical mechanics. arXiv preprint arXiv:1310.6372.

  Available at:
  http://arxiv.org/abs/1310.6372

Over the years, many other problems were explored beyond the ones presented here. An effort was made to model stem cells and cell differentiation through stochastic Boolean networks. This framework was also used to analyze the intriguing phenomenon of alternative splicing by which a same gene can produce different proteins playing with the combination of transcribed exons and introns. These problems were abandoned but interesting material was produced that might come again under focus in the future.

An important part of this author's work was in studying the complexity of circuits for distributed computation ([257]) within the Syncom ERC project. This resulted in an unpublished proof of the minimal number of components needed to implement Boolean circuits of given input size when distributed computation is used. The order of magnitude of this minimum number turns out to grow similarly to the one for integrated

circuits proved by Shannon. The design of efficient distributed circuits partly motivated our interest in MOO problems. We used this framework to implement a software platform for the systematic design of Pareto optimal distributed computation circuits. This process led us to interesting questions (and tentative answers) in other fields. Notwithstanding, work related to Pareto optimal circuit design is in an advanced stage and shall result in published papers. Within the same ERC project, this author's task was also to find alternative computational strategies. Some approaches were explored concerning classic Artificial Neural Networks (which resulted in a working paper, [283]). Other paradigms with Liquid State Machines [195] and Bayesian inference through spiking neurons [193] were explored as well, but have not been turned into published material.

The Complex Systems Summer School of the Santa Fe Institute was an important meeting point with other colleagues in the field of complexity. Thanks to it, an effort to apply maximum entropy methods to social dynamics was made together with Simon DeDeo and Steve Lansing. This resulted in a lot of unpublished material and the mastery of methods that have been applied in this PhD thesis. Also, the summer school resulted in two interesting projects, one about information flow for clustering in social networks [82] and another one to analyze written history as a complex adaptive system [203].

Further research was conducted at the Berlin Brain-Computer Interface group of the Technische Universität Berlin. Following a proposal by this author, we developed a functional Brain-Computer Interface for image reconstruction using EEG signals. The concept was tested experimentally and resulted in a paper ([278, 279]).

Finally, the dynamics of speakers of different languages within a closed territory have long been a field of inquiry for this author. This research continues to be developed together with colleagues at the Universidade de Santiago de Compostela, at the Universiteit Leiden, and at the Real Academia Galega. During this PhD thesis, this line of work resulted in three papers ([234, 280, 281]) and a fourth one in preparation.

The list of papers completed during these last years but not related to the matter presented in this PhD thesis is the following:

- Seoane, L.F., Parafita, M. C., Casares, H., Monteagudo, H., Mira, J., 2016. Predicting the evolution of heterogeneous language contact situations: the case of Galician-Spanish bilingualism. In preparation.

- Seoane, L. F. and Mira, J., 2016. Modeling the life and death of competing languages from a physical and mathematical perspective. In *Bilingualism and Minority Languages: Current trends and developments*, in press.

- Seoane, L.F., Gabler, S. and Blankertz, B., 2015. Images from the mind: BCI image evolution based on rapid serial visual presentation of polygon primitives. *Brain-Computer Interfaces*, 2(1), pp.40-56.

  Available at:
  http://www.tandfonline.com/doi/abs/10.1080/2326263X.2015.1060819

- Seoane, L.F. and Solé, R.V., 2013. Synthetic biocomputation design using supervised gene regulatory networks. arXiv preprint arXiv:1310.5017.

  Available at:
  http://arxiv.org/abs/1310.5017

- Otero-Espinar, M.V., Seoane, L.F., Nieto, J.J. and Mira, J., 2013. An analytic solution of a model of language competition with bilingualism and interlinguistic similarity. *Physica D*, **264**, pp.17-26.

  Available at:
  http://www.sciencedirect.com/science/article/pii/S0167278913002522

- Massad, D., Omodei, E., Strohecker, C., Xu, Y., Garland, J., Zhang, M. and Seoane, L.F., 2013. Unfolding History: Classification and

analysis of written history as a complex system. In *Proc. of the CSSS*, Santa Fe Institute.

Available at:
http://santafe.edu/media/cms_page_media/500/main.pdf

- Darmon, D., Omodei, E., Flores, C.O., Seoane, L.F., Stadler, K., Wright, J., Garland, J. and Barnett, N., 2013. Detecting communities using information flow in social networks. In *Proc. of the CSSS*, Santa Fe Institute.

  Available at:
  http://bit.ly/1RU5YKC

Contents

Acknowledgments **XVI**

Abstracts **XVII**

List of figures **LIX**

List of tables **LXI**

1. PARETO OPTIMALITY **1**
1.1. Optimization is at the heart of our physical reality 1
1.2. Brief historical overview 4
1.3. Definitions and Methods 8
1.3.1. Pareto optimality 8
1.3.2. Genetic algorithms for Multi Objective Optimization 13
1.3.3. Trait space and morphospaces 15
1.3.4. A recent example in biology 19

2. THE PARETO FRONT AND THE GIBBS SURFACE – PHASES, PHASE TRANSITIONS, AND CRITICALITY **23**
2.1. Collapsing MOO into the simplest SOO problem 24
2.1.1. Concavity/convexity and order parameters 27
2.2. Phase transitions in the Pareto front 28
2.2.1. Second order phase transitions 29
2.2.2. First order phase transitions 30

2.3. Criticality in the Pareto front 32
2.4. Thermodynamics as a multiobjective optimization problem 38
2.4.1. Equivalence between the Gibbs surface and the Pareto front 40
2.4.2. Solving the Ising and Potts models from an MOO perspective 45

3. PHASE TRANSITIONS AND CRITICALITY IN PARETO OPTIMAL SYSTEMS 53
3.1. Phase transitions in Pareto optimal complex networks . . . 54
3.1.1. Overview of the problem 54
3.1.2. Multiobjective optimization of complex networks 57
3.1.3. Outcome of network optimization 59
3.1.4. Discussion of Pareto optimal networks 71
3.2. Systems poised to criticality through Pareto selective forces 73
3.2.1. Descriptions of Pareto optimal sets based on statistical ensembles 74
3.2.2. Pareto selective forces in real life 77
3.3. Reviewing some selected literature with Pareto optimal designs 80
3.3.1. Networks for efficient communication 80
3.3.2. Robust topological networks 83
3.3.3. Protein folding 85

4. EXPLORING THE CRITICALITY OF HUMAN LANGUAGE 91
4.1. Ambiguity in language networks 92
4.1.1. Introduction 92
4.1.2. Scaling in Language 93
4.1.3. Small World Language Networks 98
4.1.4. Ambiguity in Semantic Networks 104
4.1.5. The Least-Effort Language Agenda 108
4.1.6. Ambiguity, Principles of Information Theory, and Least Effort 119
4.1.7. Discussion and Prospects 121

4.2. Exploring the morphospace of communication codes . . . 124
4.2.1. The criticality of least effort communication . . . 125
4.2.2. Sampling the morphospace . . . 128
4.2.3. The morphospace of communication codes. . . . 129
4.2.4. Code archetypes and real languages . . . 143

5. COMPRESSING SYMBOLIC DYNAMICS 149
5.1. Ants and meaningful levels of description . . . 150
5.2. Information theoretical approaches to level identification . 152
5.2.1. The information bottleneck method . . . 156
5.2.2. Conditions for good coarse-grainings of symbolic dynamics . . . 158
5.3. A naive empirical approach to the structure of human language . . . 161
5.3.1. General problem . . . 163
5.3.2. Computational and Statistical Mechanics of human language . . . 165
5.3.3. Discussion . . . 171

6. INFORMATION THEORY, PREDICTABILITY, AND THE EMERGENCE OF COMPLEX LIFE 173
6.1. Introduction . . . 174
6.2. Evolution and Information Theory . . . 177
6.2.1. Messages, channels, and bit guessers . . . 179
6.3. Life complexity is tuned by the predictability-replication tradeoff . . . 186
6.3.1. Guessers isolated in environments of fixed size . . 187
6.3.2. Evolutionary drivers . . . 191
6.4. Discussion . . . 195

7. CONCLUSIONS 199

A. APPENDICES 207
A.1. Analytic and numeric approaches for MOO solving . . . 207
A.1.1. A multiobjective genetic algorithm . . . 207

A.1.2. Smoothing of the front and order parameters . . . 210

List of Figures

1.1. **Leading the way towards Multi Objective Optimization a** Léon Walras [340], **b** Francis Ysidro Edgeworth [102], and **c** Vilfredo Pareto [238]; three historical figures in the introduction of multi objective thinking in economy. 4

1.2. **Curve of indifferent distributions of goods.** An agent must choose between different distributions of goods, in this case represented by coins and sugar in an abstract mathematical space. This one precise agent considers *equal to its interests* any distribution of goods lying along the solid line (open circles) . Similarly, distributions of goods laying along the dashed line (black dots) are considered equal to each other by the agent. Distributions in different lines are not considered equal to each other: they are either better or worse to the agent. The same thing happens with other distributions outside the lines (gray dots). 5

1.3. **Pareto optimality and the Pareto front. a** The abstract set X consists of all feasible objects within our design space. Upon these objects we can measure a series of *target functions* $T_f(X) = \{t_1(x), t_2(x)\}$ that map these objects into *target space* **b**, here a plane. **c** If we seek to minimize t_1 and t_2 simultaneously, a straightforward geometric condition (dominance, shaded area and black dot) discards solutions that are less optimal – i.e. they perform worse in both targets. This same condition does not allow us to resolve situations in which two solutions do not dominate each other – i.e. one of them scores better than the other in one of the targets and worse in the other one. The most honest choice is to keep those mutually non-dominated solutions. This defines the most optimal tradeoff termed *Pareto front* (**b** and **c**, $x_\pi \in Pi$ and thick gray lines), which also delimits the set of feasible designs in target space. 9

1.4. **Pareto optimal folding of proteins.** Protein folding is subjected to several conflicting physical constraints including tensions, torsions, electrostatics and others. Attempting to satisfy all these conditions simultaneously results in a Pareto front that contains protein shapes that are ‘more optimal’ in a sense, as we will review in section 3.3.3 of chapter 3. (Adapted from [78].) 11

1.5. **Pareto optimal models of spiking neurons.** Pareto optimal models of spiking neurons that minimize the error in predicting the spike train firing rate and the so-called accomodation error. Each point corresponds to a possible model. All of them together reconstruct the best tradeoff possible, which emerges when we try to account for conflicting error sources with a limited mathematical representations of the real spike trains. (Adapted from [99].) 12

1.6. **Scores to implement a multiobjective genetic algorithm.** **a** The domination relationship induces layers of non-dominated designs given a sample of X. **b** Alternatively, we can score how many designs within a sample dominate each other. 13

1.7. **Raup's coiling morphospace.** **a** Three parameters are enough to characterize the most relevant aspects of possible shell coilings. **b** Depending on the values of these parameters different shapes are obtained. This renders a *morphospace* where we visualize straight away how different morphologies relate to each other and where real shells are located. (Figures adapted from [323] and [255].) 16

1.8. **Network morphospaces.** **a** A morphospace of complex networks with three axes measuring different traits of hierarchical systems. **b** Graphs can also capture the efficiency of communication systems. A morphospace allows us to relate different topologies to each other according to this efficiency. A series of Pareto fronts stand as a relevant frontier of feasible communication designs. (Figures adapted from [68] and [132]) 17

1.9. **Pareto front in phenotype space.** **a** Optimality with respect to each of the targets decreases with the distance to the archetype optimal for that target. This implies that the Pareto front looks like a straight line when plotted in phenotype space [293, 292]. Designs off this line (crossed phenotype) are not Pareto optimal and get replaced by some better solution. **b**, **c** The straight line generalizes to a polytope with K vertices, where K is the relevant number of archetypes – i.e. of underlying target functions, which remain unknown. **d** Using the software introduced in [141], a data set measuring diverse traits on several species of bats was analyzed [293]. The data consistently collapses into a triangle, indicating that three traits are simultaneously optimized. This also implies that most of phenotype space will be empty due to Pareto optimality. (Figures adapted from [293].) 21

2.1. **From multi to single-objective optimization**. **a** When a linear assumption is made, the problem becomes that of minimizing a global *energy* function $\Omega(\lambda) = \lambda t_1 + (1 - \lambda)t_2$ (see text). For a fixed λ one sole SOO is posed whose solution lies where $\tau_\lambda(\Omega)$ (straight lines with slope $\delta = -\lambda/(1-\lambda)$) matches the tangent of the front. **b** By changing λ, we visit other solutions of the same SOO family. 25

2.2. **A notion of up-down is necessary to define convexity**. Convexity and cavities are defined with respect to the direction of improvement of Ω given λ. This way, the criteria for convexity are consistent disregard of whether we deal with maximization or with problems that mix minimization and maximization. 27

2.3. **Convex Pareto front with a tangent whose slope does not span the whole range** $(-\infty, 0)$**. a** The slope of the Pareto front spans $d \in (-\infty, \delta^*)$. The front *ends abruptly* at its bottom-right. There is a range ($\lambda < \lambda^*$, with $\lambda^* = -\delta^*/(1-\delta^*)$ indicated by the gray fan) for which well defined SOO problems exist, whose solution is persistently the same (open circle). For $\lambda > \lambda^*$ (filled circle) the front is sampled gently as in figure 2.1**b**. Any order parameter θ (inset) does not change if $\lambda < \lambda^*$ because the SOO optimum remains the same. Its derivative is not zero for $\lambda > \lambda^*$. This causes an abrupt shift in $\frac{d\theta}{d\lambda}$ at λ^* while $\theta(\lambda)$ remains continuous. **b** The exact same situation happens if the pathology is found at the top- left of the front. **c** A sharp edge is associated with two discontinuities in the derivative of any order parameter. 29

2.4. **Concave Pareto front, or fronts with concavities. a** Only two solutions are ever SOO global optima in a concave front: one if $\lambda > \lambda^*$ and another if $\lambda < \lambda^*$. For $\lambda = \lambda^*$ both solutions coexist. Any order parameter presents a sharp discontinuity at $\lambda = \lambda^*$. A similar situation happens in **b** and **c**. In the later case, while $\theta(\lambda)$ is not continuous, its derivative is. **d** An *energetic landscape potential* is built through equation 2.1. Plotting this function for all the points of the front in **c** reveals an energetic boundary below which no solutions exist. Pareto suboptimal solutions lay above the boundary and SOO optima at fixed λ sit at the bottom of energy wells. Metastable solutions are associated to local minimums and lead to hysteresis if we change λ back and forth. 31

2.5. **Criticality in second order phase transitions. a** Systems critical over a range of values of the targets (here $[t_1^-, t_1^+]$) are such that t_2 is a linear function of t_1 in that range. **b** Rolling a rigid line over this front yields a first order phase transition. Besides, at the critical point any order parameter diverges. These situations containing both first order and critical elements are often referred to as *hybrid phase transitions* [96, 26]. **a1-3**, **b1-3** In this case, the limit as $t_1^- \to t_1^+$ yields a well known second order critical transition. 35

2.6. **Approaching a critical point, I. a**, **b** As before, a linear range $[t_1^-, t_1^+]$ implies a first order phase transition with a critical point at which any order parameter diverges. This again is often called a hybrid phase transition. **a1-3**, **b1-3** In this case, the limit $t_1^- \to t_1^+$ implies that the phase transition is reduced to a critical point with no phase transition associated. 36

2.7. **Approaching a critical point, II. a** There is an alternative to reach the critical point in figure 2.6**a3**, **b3** without a hybrid phase transition. If three targets are involved, this situation happens whenever a cavity ceases to exist. **b** This situation reminds us of the critical point in liquid-vapor transitions. 37

2.8. **Laws of thermodynamics and the Pareto front. a** According to the second law of thermodynamics, at constant internal energy (vertical dashed line) the microcanonical ensemble is the one that maximizes the entropy (and is hence mapped into the open circle in this example). Implementing this maximization for varying energy yields a function on the $U - S$ space that for thermodynamic systems is usually monotonously increasing with U – more energetic systems usually have more entropy. This guarantees that any two points on this curve are mutually non-dominated. There cannot be any point above this curve, thus the obtained curve must be exactly the Pareto front of the corresponding MOO problem. **b** This curve would not match the front only if the microcanonical entropy were not monotonously increasing with U. This is an odd situation in thermodynamics. These non-increasing stretches would necessarily lay inside a cavity (solid black curve) and would never show up in thermodynamic equilibrium. **c** Such situation can also happen beyond the global maximum of the entropy, which is only reached for $T = 0$ (dashed line). Points of the microcanonical entropy beyond this maximum would require $\partial S/\partial U = 1/T < 0$ (i.e. negative temperatures, dotted line). In both **b** and **c** the entropy of microcanonical ensembles still contains the whole Pareto front and, of course, its convex hull. . . 42

2.9. **Pareto fronts for the Ising and Potts models. a** The front of the mean-field Ising model (thick gray line) is convex but ends abruptly revealing a range $\beta \in (0, 1/Jz]$ (gray fan) for which SOOs arrive to the most entropic solution always. **b** A sample of arbitrary distributions $P = \{p_1, p_2, p_3\}$ (black crosses) for the $q = 3$ Potts model is dominated by its Pareto front whose top-right part is concave. This indicates a first order phase transition. **c** That cavity becomes noticeable when analyzing the slope of the front, which is not monotonously decreasing. . . . 47

2.10. **Pareto front for the mean-field Potts model with different q. a** Pareto fronts of the Potts model for $q = 3, 5, 8, 10$. Although hardly noticeable, all these fronts have got concave stretches towards their upper-right ends. This indicates that all of them undergo first order phase transitions from the most disordered state to a more ordered phase where symmetry has been broken to favor just one of the states. A Pareto front for q dominates the Pareto front for every $q' < q$ indicating that this system will never leave empty one of its available states spontaneously, except in the most ordered state in which all spins are aligned. **b** Inverse temperature at which the mean-field Potts model presents its first order phase transition for $q = 3, ..., 10$. The results match perfectly those from the literature [178]. 51

3.1. **A two dimensional example of Pareto optimality.** **a** $\gamma \in \Gamma$ are all possible connected networks with a given number of nodes. They populate some network morphospace where we seek those graphs minimizing some measurable feature. If we deal with just one fitness function, an energy landscape can be defined and the optima are easily found at the bottom of energy wells. **b** If more than one optimization target are at play, this landscape picture falls apart and we need to adopt a Pareto optimization approach. Then our task is to find a set of Pareto optimal solutions ($\Pi_\Gamma \subset \Gamma$) that minimizes all targets (here t_1 and t_2) simultaneously. These functions map each network $\gamma \in \Gamma$ into $\mathbb{R}^2$. The subset of Pareto optimal solutions is mapped into the Pareto front (thick gray curve). Along this curve it is not possible to improve both t_1 and t_2 at the same time. 56

3.2. **Pareto front of the fully topological problem.** **a** The front (solid gray curve) is a straight line connecting two phases: a star and a clique. The slope of the line $d^c = -1$ determines that at $\lambda^c = 1/2$ a first order phase transition takes place. All networks laying on the front are global SOO optima at that critical value. Among them we find networks produced by attaching links to a star and others radically different from the star and from the clique (note the two graphs marked A and B: only one of them can be produced by attaching edges to the star). **b** All *core graphs* for have been listed for $N \leq 5$. Beyond that, it becomes increasingly difficult to count how many there are or even to tell apart two different ones. 60

3.3. **Partly geometrical problem on nodes scattered over a plane.** **a** The front follows the archetype of the topological problem with two roughly perpendicular stretches that trade off between the clique (top-left), the star, and the MST (bottom-right). Incomplete cliques are reached after a second order phase transition because the Pareto front ends abruptly in its top-left (inset). The other extreme of the front ends smoothly. **b** A sharp edge indicates a second order phase transition with the star graph being optimal for a range $\lambda \in (\lambda^-, \lambda^+)$. **c** Plotting an order parameter as a function of λ reveals both transitions at $\lambda \simeq 0.61$ and at $\lambda^- \simeq 0.01$ and $\lambda^+ \simeq 0.3$ (inset). The star is optimal in the range $\lambda \in (\lambda^-, \lambda^+)$, thus any order parameter is constant in that range. **d** The non-analyticity of the Pareto front is inherited by the energetic landscape also as a sharp edge. The SOO is vividly illustrated thanks to this potential landscape, whose minimum is occupied by one same network for several values of λ. 64

3.4. **Partly geometrical problem on a circle.** **a** Again, the front follows the archetype of the topological problem with two roughly perpendicular stretches that trade off between the clique (top-left), star networks, and the MST. Around the clique it is observed the same phase transition as before. In the cavity it is solved a complex rearrangement. Networks that drop their larger connections first (A) must morph into a star (B), which requires some of these far-reaching edges. Therefore, some Pareto optimal networks are produced that never get to be SOO optima (C), yielding a first order transition. **c** Order parameters as a function of λ reveal the first ($\lambda^c \simeq 0.34$, magnified in the inset) and second ($\lambda^* \simeq 0.59$) order transitions. **d** The landscape potential unveils the mechanisms for local equilibrium and hysteresis associated to first order transitions. At low levels of the control parameter ($\lambda \sim 0.1$) only one minimum exists in the global energy Ω. 67

3.5. **Fully geometrical problem for nodes scattered over a plane.** **a** The front has no accidents. It is completely convex and spans all possible slopes so that each $\lambda \in (0, 1)$ poses an SOO with a different solution. As we roll over the front, the clique gently leads to less connected networks, towards the MST. **b** The absence of phase transitions renders smooth plots of any order parameters. 69

3.6. **Fully geometrical problem on a circle.** **a** The front presents a smooth transition between the clique and the open circle, with no relevant feats except in the extremes of the front. These end up abruptly, as in second order phase transitions. **b** Plotting any order parameters reveals these phase transitions at $\lambda_1^* \simeq 0.28$ and $\lambda_2^* \simeq 0.98$ (inset). 70

3.7. **Simultaneous minimization of average path length and edge density. a** Purely topological graphs lie on a linear Π_Γ. In **b**, **c** the Euclidean distance weights the cost of each link when computing the density of edges. Graph drawings are qualitative, units refer to the systems in [285]. **b** Nodes distributed over a plane. Optima trade between a clique, a star graph, and the Minimum Spaning Tree through two second order transitions. **c** Nodes placed over a circle display a first and a second order transition. . 76

3.8. **Testing MaxEnt models. a-c** p-values for MaxEnt models of Pareto optimal sets $\Pi_{A,B,C}$ as a function of λ and α (α = 0.0125, solid; α = 0.1, dashed; and α = 1 dotted lines; curves have been normalized for comparison). **a** α does not affect the critical ($\tilde{\lambda}_A = \lambda_A^c$) description of Π_A. **b**, **c** Changing α changes the best model, so that an α-invariant, consistent description does not arise for these Pareto optima. $\tilde{\lambda}_{B,C}(\alpha)$ usually do not correspond to relevant parameters in phase space. **d** The best model misses the least information about each data set (Π_A, solid; Π_B, dashed; and Π_C dotted lines; α = 1). This loss is vanishingly small in the critical case. 78

3.9. **Pareto optimal networks for communication.** **a** Pareto front extracted from [132] (reconstructed manually from front 1 in figure 1.8**b**) that entails the optimization of graphs with respect to: i) minimizing diffusion efficiency and ii) maximizing routing efficiency. A large cavity in the front and a smaller one (slightly noticeable at the bottom left) reveals two first order phase transitions that are noted in any order parameter **b**. These transitions result in important rewirings of the graphs to switch between distinct topologies: from almost linear chains to tree-like structures containing clusters in its branches, to networks seemingly compacted around a unique cluster. (Network drawings corresponds to actual topologies found in [132]. Their location along the front and at the order parameter plot are qualitatively correct, but not precise.) **c** An energy landscape illustrates how networks can become local minimums and how phases can coexist at a given λ ($\lambda = 0.4$). **d** Similar analysis are possible for the other fronts in figure 1.8**b**. 82

3.10. **Pareto optimal networks against targeted and random attacks.** **a** Pareto fronts reconstructed from [250]. **b** Pareto front projected onto a 3-D space where a new constraint (average degree of nodes) adds up a new dimension. The corresponding Pareto front presents a critical point. **c** The characteristic double well of first order phase transition appears as we move close and beyond the critical point. 84

3.11. **Pareto optimal Protein folding.** **a** Pareto front for the simultaneous minimization of bonding and non-bonding energy in the 1UTG protein. A prominent cavity is revealed that implies a first order phase transition as the control parameters trading both kinds of energy change. Following established methods from the MOO literature [140] it is found that the best protein structure is around the region marked with a large cross. Such structures lay within the cavity of the front. **b** Protein shapes laying in the convex hull of the Pareto front are visited for different values of the control parameter. Foldings within the cavity (including the one selected in [78]) are bypassed by a first order phase transition. (Protein drawings correspond to 1UTG, but not to the actual foldings at either phase. They are intended as an illustration.) **c-f** This phase transition is physical in nature so that these Energy landscapes have a direct interpretation in terms of necessary work or heat to fold/unfold the protein, or of energy barriers that must be overcome. The geometry of the Pareto front may offer key information about the dynamics as control parameters are varied. 89

4.1. A seemingly universal feature of all known human languages is Zipf’s law, illustrated in Figure **a** from the rank-abundance statistics obtained using Melville’s Moby Dick (see text). Moreover **b** language contains multiple levels of nested complexity, illustrated here by means of an idealized collection of spheres whose size rapidly grows as the objects being considered at one level are combined to obtain those in the next level. Letters and syllabus are the first levels, followed by words and pairs of words and eventually sentences. The diagram actually underestimates the real proportions of combinatorics. 95

4.2. A language network can be build in different ways. The simplest one is considering co-occurrence between words within sentences from written corpuses. Here **a** we have used the first chapter of Paul Auster's “Augie Wren Xmas Tale”, from which we draw our network. Each ball is a different word, whereas an undirected link between two balls indicates that those words appeared one after the other within a sentence in the text. Two parts of the web are zoomed in **b** and **c**. In **b** we observe multiple linear structures and chains associated to particular sentences. Meanwhile, in **c** we can see that some words have a very large number of links with others and are referred to as “hubs”, whereas most words have just one or two connections. LPNs follow scale-free degree distributions, as exemplified in **d**. 100

4.3. **A simple network of semantic relations among lexicalized concepts.** Nodes are concepts and links, semantic relations between concepts. This would correspond to a very small subset of a vast set of words and semantic relationships. Associations between words allow us to navigate the network. Locally, the number of triangles is very large, allowing multiple ties among semantically related words – and contributing to a high clustering, as seen in the text. Moreover, given two words, such as “volcano” and “pain” can be linked through different paths, two of which are illustrated here using thick lines. 106

4.4. **Modeling language evolution through robots or matrices.** In order to model language evolution, one can use a number of artificial systems, including among them robotic, embodied agents **a**. Here two robots (image from the Neurocybernetics group at Osnabrück, see https://ikw.uni-osnabrueck.de/~neurokybernetik/) share a common environment seeded by a number of objects, which they can name. Robots can evolve a rudimentary grammar that goes beyond the simple word inventory that we could expect. Additionally, simple mathematical models can also be used in order to capture essential features of language organization. A model of language can be formulated in terns of a matrix **b** that relates a set of n signals (indicated as $s_1, s_2, ..., s_n$) with a set of m objects or actions of reference ($r_1, ..., r_m$). A simple case with $n = m = 6$ is displayed. A signal is associated to an object using a link connecting them. Here for example signal s_5 is used to refer to object r_4. 110

4.5. **Phase transition in least-effort language.** As we vary λ, equation 4.21 awards different importance to a speaker's or a hearer's requirements of a tongue. Accordingly, we move from a scenario that contents the former to one that pleases the later. But the change is sharp and happens at a very precise value of $\lambda = \lambda_c \equiv 0.5$, in accordance with the description of a first order phase transition. The simulations to generate these plots – a Genetic Algorithm (GA) that proceeded to minimize equation 4.21 with different values of λ – are in good agreement with this numerical critical value. Because of this sudden regime shift we can observe very abrupt changes in some *order parameters* than can be measured in a language: **a** The mutual information between signals and objects (whose average value across the top population of the GA is plotted) says how much information the signals of a language convey about the named world. 114

4.6. **Seeding the morphospace of communication codes. a** Pareto optimal codes generated randomly following different prescriptions always fall in the $y = 1 - x$ line. The different process from left to right (components along the x axis have a 0.1 offset) are: i) one dice process, ii) an iterated dice process, iii) power laws with different exponent, iv) preferential attachment, and v) uniform association between objects and signals. **b** After seeding the morphospace with a series of random processes, it remains almost empty. It was necessary to apply an evolutionary algorithm to achieve a uniform sampling of the morphospace. 127

4.7. **Vocabulary size, polysemy, and synonymy over the morphospace.** **a** Vocabulary size represented the number of languages used among all those available and it is only low near the star graph (in a prominent area labeled B) and along the Pareto front. Most of the morphospace presents large vocabularies. **b** Polysemy is large in the B region and as we complete the matrix A towards the block code. **c** Synonymy increases uniformly as we move apart from the front except for codes within B. This locates them far away from the Pareto optimality condition (zero synonymy). 130

4.8. **Vocabulary size, polysemy, and synonymy along the Pareto front.** **a** Codes along the Pareto front keep a relatively low vocabulary except close to the one-to-one mapping. Also, two branches seem noticeable around the middle of the front, suggesting that similar Pareto optimal values of $H_m(R|S)$ and of $H_n(S)$ can be achieved with differently wired codes. **b** A reduced vocabulary size does not result in a strictly monotonous increase of polysemy as we approach the star code. Instead, languages with similar $H_m(R|S)$ may present different polysemy levels. The range available grows as we approach the maximally ambiguous code. 131

4.9. **Studying network connectivity along the Pareto front.** **a** The largest connected component seems to grow steadily as we proceed to the star code due to the ambiguity building up around a few degenerated signals. **b** This results in a similar plot for $||C_1^R||$. **c** In Pareto optimal S-graphs every connected component has got one node. Here we plot this relative to the number of clusters, which turns out to be L. Hence **d** represents $1/L$, the inverse of vocabulary size. **d** H_C does not reveal a clear pattern. This highlights again both the diversity of the Pareto front and that similar values of equally Pareto optimal target functions can be obtained with fairly different configurations of codes. . 134

4.10. **Studying network connectivity over the morphospace.** Most of the morphospace presents well connected networks except along the Pareto front, where the Pareto optimality condition imposes that the code splits into unconnected clusters. The region B also presents smaller connected components. This seems to be the case because those codes do not use most of the signals available. **b** The R-graph does present large connected components in B, while the S-graph is poorly connected again **c**. **d** H_C peaks near the one-to-one mapping and a stripe with large entropy that runs parallel to the front. It vanishes in the interior of the morphospace. 135

4.11. **Complexity of random walks in language graphs along the Pareto front.** **a** The random walks were designed so that both extremes of the front present the largest entropy possible. (The point corresponding to H_R = 1 for the star graph is hard to see.) For intermediate values of $H_m(R|S)$, codes exist that present very different H_R. The lowest values happen very close to the star code. **b** H_S decays naturally as the vocabulary size diminishes. **c** Low H_{2R} for codes near the star graph is reproduced, as well as the maximums at the extremes of the front. The rest of the codes seem to cluster together around characteristic entropy values, but a clear trend cannot be appreciated. **d** While the 2-signal entropy still decays as the vocabulary size diminishes, different kinds of languages seem to be appreciated in the intermediate regions of the Pareto front. This results in different H_{2S} for a same $H_m(R|S)$. This also indicates that H_{2S} captures some structure in Pareto optimal codes that H_S misses – we will see that this is not the case in the rest of the Pareto front. 138

4.12. **Complexity of random walks in language graphs over the morphospace.** **a** The entropy of single object frequency in random walks presents a non-trivial structure within the morphospace. Two different regions (containing well differentiated A matrices) are capable of breaking the symmetry in the random-walk sampling of objects. **b** Codes in yet another region do the proper to the sampling of signals, but this area does not correspond to any of the previous ones. **c** The entropy of 2-grams is low in a region that correlates roughly with a large entropy of the connected component distribution (H_C). **d** The entropy of 2-grams in signals largely reproduces that of single signals, suggesting that there is not any new structure that H_{2S} can uncover. 140

4.13. **Goodness of power laws in describing the frequency of signal use along the Pareto front.** **a** The goodness-of-Zipf test (which is better if the KS-score is lower) has its lowers value near $H_m(R|S) \sim 0.5$, where we have seeded actual Zipf distributions. Discrepancies shall be attributed to the numerical methods employed. **b** A broad region with $H_m(R|S) \geq 0.5$ presents good fit to power laws. This corresponds to the parts of the Pareto front where actual power laws, preferential attachment, and the dice have been seeded. **c** The exponents around these regions vary between $\gamma \sim 1.1$ and $\gamma \sim 2.5$, with codes seemingly belonging to one of two well differentiated branches. . . . 142

4.14. **Goodness of power laws in describing the frequency of signal use across the morphospace.** **a** The Kolmogorov-Smirnov test reveals a region of the morphospace in which Zipf's law accounts fairly enough for the frequency of signal use. **b** Power law distributions with arbitrary exponent render a better fit in a broad region near the lower half of the front, where good fits to power laws are also observed. **c** The preferred exponent raises from ~ 1.2 to ~ 2.5 as we approach the star code, but most of these exponents do not correspond to good fits. 143

4.15. **Clustering of languages across the morphospace.** k-means clustering using all principal components reveals a consistent structure in the morphospace. Five cluster are shown here. Real languages fall within cluster I, close to the one-to-one mapping proper of animal communication systems. The real matrices are marked: *Adj* for the adjectives, *Adv* for the adverbs, *Noun* for the nouns, and *Verb* for the verb. If certain grammatical words are included (named with an apostrophe: *Noun'* for nouns and *Verb'* for verbs) they move into cluster II and towards the center of the morphospace, relatively close to the Pareto front. **b** All clusters get further segregated in two principal component space. This space appears interrupted by a stripe along which no codes exist. 144

5.1. **Epsilon machines**. In epsilon machines, causal states (represented by a double circle) are visited randomly. At every causal state we choose between the outgoing arrows according to the probabilities indicated. When an arrow is chosen we emit the symbol associated to that arrow. Enclosed within each causal state we find the frequency with which it shows up in the long term limit. The entropy of the probabilities of the causal states is the algorithmic complexity C_μ. **a** Actual ϵ-M of the golden mean process, which consists of all binary strings that do not present two consecutive zeros. **b** ϵ-M inferred from the genetic sequence of a laboratory construct. **c** ϵ-M inferred from the complete genome of *Mycoplasma pneumoniae*. The increasing size of the ϵ-machines is correlated with an increase of algorithmic complexity in this case – this is not necessarily true. 154

5.2. **Information theoretical measurements in symbolic dynamics**. A relevant quantities are the entropy of the future dynamics given the causal states ($h_\mu L$), where h_μ represents the entropy production rate – so this number grows linearly with the distance into the future L. The crypticity (χ^+) relates the causal states of forward and backwards dynamics. More importantly for us, the algorithmic complexity C_μ (equation 5.2) captures the complexity of the symbolic dynamics through the entropy of its causal states. This is one of those complexity measurements that is low for deterministic or maximally random processes, but that peaks somewhere near the *edge of chaos* [175]. 155

5.3. **The information bottleneck method reconstructs the convex hull of the Pareto front**. **a** A simple non-deterministic source (see [316]) was reconstructed by models that predicted 2 steps into the future given 5 into the past. Candidate models had different number of causal states (marked in the figure) and perform differently. They trace the rate-distortion curve, that should correspond to the convex hull of an underlying Pareto front. Phase transitions between models take place, but without the actual front we cannot assess their order. **b** Predictively reversible processes present a rate-distortion curve that is a straight line. If this is the actual front of the process, it corresponds to a critical phenomenon according to the framework in section 2.3 of chapter 2. (Figures adapted from [316].) 157

5.4. **Conditions for good coarse-grainings**. **a** To build a model of a symbolic series we need i) a mapping π that translates the fine grained symbols of alphabet χ into a lumped alphabet χ_R and ii) a mapping ψ that establishes a dynamics in the coarse-grained variables. The original mapping (ϕ) is often unknown. If known, it constitutes the most faithful description of the dynamics over χ, but this is often very costly to simulate. **b** Good mappings π and ψ are such that, once a state has been translated into the coarse-grained symbols, information about the fine grained dynamics cannot give us further information about the coarse grained variables (condition I). This condition implies that ψ enforces Markovian dynamics on χ_R (condition IV) and that the mappings π and ψ commute (condition II). If π is deterministic, some backward implications apply. Markovianity is revealed as the weakest condition. 159

5.5. **Interactions between spins and word classes**. **a** A first crude model with spins encloses more information than we need for the kind of calculations that we wish to do right now. **b** A reduced version of that model gives us an *interaction energy* between words or classes of words. These potentials capture some non-trivial features of English syntax – e.g. the existential ‘there’ in “there is” or modal verbs (marked E and M respectively) have a lower interaction energy if they are followed by verbs. Interjections present fairly large interaction energy with any other word, perhaps as a consequence of their independence within sentences. 169

5.6. **Pareto optimal maximum entropy models of human language**. Among all the models that we try out, we prefer those Pareto optimal in energy minimization and entropy maximization. **a** These reveal a hierarchy of models in which different word classes group up at different levels. The clustering reveals a series of grammatical classes that belong together owing to the statistical properties of the symbolic dynamics, such as possessives and determiners which appear near to adjectives. **b** A first approximation to the Pareto front of the problem. A more accurate one is likely to display cavities along the front corresponding to jumps in the complexity. 170

6.1. **Predictive agents and environmental complexity. a** An agent G interacts with an external environment E that is modeled as a string of random bits. These bits take value 0 with probability p and value 1 otherwise. The agent tries to guess a sequence of n bits at some cost, with a reward bestowed for each correctly guessed bit. The persistence and replication of the agent can only be granted if the balance between reward and cost is positive ($\rho_E^G > 0$). **b** For a machine attempting to guess n bits, an algorithmic description of its behavior is shown as a flow graph. Each loop in the computation involves scanning a random subset of the environment $B = (b_1, ..., b_n) \subset E$ by comparing each $b_i \in B$ to a proposed guess w_i. **c** A mean field approach to certain kind of 1-guesser (modeled in the text through equations 6.1, 6.2, and 6.3) in environments of infinite size renders a boundary between survival ($\rho_E^G > 0$) and death ($\rho_E^G < 0$) as a function of the cost-reward ratio (α) and of relevant parameters for the 1-guesser model (p in this case). Note that for $\alpha < 0.5$ every 1-guesser survives for free. 175

6.2. **Information and evolution through natural selection.** **a** A simple diagram of the underlying evolution of a population of bit guessers. The survival and replication of a given agent G is indicated by branching whereas failure to survive is indicated with an empty circle as and endpoint. **b** The propagation of a successful replicator can be understood in terms of a Shannon-like transmission process from one generation to the next. 179

6.3. **From a generative model to inference about the world.** A diagrammatic representation of the algorithmic logic of the bit guessing machine. Our n-guesser contains a generative model (represented by a pool of words) from which it draws guesses about the environment. If a bit is successfully inferred, the chosen conjecture is pursued further by comparing a new bit. Otherwise, the inference is reset. 183

6.4. **Probability of correctly guessing a bit in environment ensembles of constant size.** $\bar{p}^n_m$, average probability that n-guessers correctly guess 1 bit in m-environments for $n = 1$ (crosses, solid line), $n = 2$ (squares, dashed line), $n = 5$ (pluses, dotted line), and $n = 10$ (triangles, dot-dashed line). $\bar{p}^1_m$ can be computed analytically (solid line in the main plot) and marks an average, lower predictability boundary for all guessers. In the inset, the data has been smoothed and compared to a given value of α, represented by a horizontal line. At the intersection between this line and $\bar{p}^n_m$ we find $\bar{m}^n(\alpha)$, the environment size at which n-agents guess just enough bits as to survive given α. Notice that n guessers are evaluated only in environments of size $m \geq n$. 188

6.5. **Mutual information and entropy.** Guessers with $n = 1$ (crosses), $n = 2$ (squares), $n = 5$ (pluses), and $n = 10$ (triangles) are presented. **a** $I(G : E_m)$ and $\langle I(G : E)\rangle_{E_m}$ (inset) quantify the different sources of information that allow more complex guessers to thrive in environments in which simpler life is not possible. **b** The entropy of a guesser's message given its environment seems roughly constant in these experiments despite the growing environment size. This suggests an intrinsic measure of complexity for guessers. Larger guessers look more random even if they might carry more meaningful information about their environment. The thick black line represents the average entropy of the environments (which approaches $log(2)$) against which the entropy of the guessers can be compared. 190

6.6. **Dynamics around $\bar{m}^n(\alpha)$.** Again, guessers with $n = 1$ (solid line), $n = 2$ (dashed line), $n = 5$ (dotted line), and $n = 10$ (dot-dashed line). **a** $P^n(m, \alpha)$ tells us how often do we find n-guessers in m-environments when they are allowed to roam constrained only by their survival function ρ^n_m. The central value $\hat{m}^n$ of $P^n(m, \alpha)$ must converge to $\bar{m}^n(\alpha)$ and oscillations around it depend (through N_g and N_e) on how often do we evaluate the guessers in each environment. **b** Average $\hat{m}^n$ for $n = 1, 2, 5, 10$ and standard deviation of $P^n(m, \alpha)$ for $n = 1, 10$. Deviations are not presented for $n = 2, 5$ for clarity. The inset represents a zoom in into the main plot. 192

6.7. **Evolutionary drivers: competition.** Coexisting replicators will affect each other's environments in non-trivial ways which may often result in competition. We implement a dynamics in which 1-, 2-, 3-, and 4-guessers exclusively occupy a finite number of environments (transmission channels) of a given size (fixed m). The 100 available slots are randomly occupied at $t = 0$ and granted to the best replicators as the dynamics proceed. We show $P_m(n, t = 10\,000)$ for $m = 5, \ldots, 39$ and **a** $\alpha = 0.6$, **b** $\alpha = 0.65$. The most abundant guesser at $t = 10\,000$ is shown for **c** $\alpha \in (0.5, 1)$ and **d** $\alpha \in (0.6, 0.7)$. Once m is fixed, there is an upper value of α above which no guesser survives and all 100 available slots remain empty. Competition and the replication-predictability tradeoff segregate guessers according to the complexity of the environment – i.e. of the transmission channel. Coexistence of different guessers seems possible (e.g. $m = 15$ in **b**), but it cannot be guaranteed that the dynamics have converged to a steady distribution. 194

6.8. **Evolutionary drivers: exhausted resources.** Rather than monopolizing channel slots (as in figure 6.5), we can also conceive individual bits as valuable, finite resources that get exhausted whenever they are correctly *guessed*. Then a successful replicator can spoil its own environment and new conditions might apply to where life is possible. **a** Average reward obtained by 1-, 2-, 5-, and 10-guessers in environments of different sizes when bits get exhausted with efficiency $\beta = 1$ whenever they are correctly guessed. **b** Given $\alpha = 0.575$ and $\alpha = 0.59$, 1- and 2-guessers can survive within upper and lower environment sizes. If the environment is too small, resources get consumed quickly and cannot sustain the replicators. In message transmission language, the guessers crowd their own channel. If the environment is too large, unpredictability takes over for these simple replicators and they perish. 195

List of Tables

5.1. **Grammatical classes present in our corpora**. 167

Chapter 1

PARETO OPTIMALITY

This chapter is a brief introduction to Pareto Optimality and to the different fields where it has historically been applied. Pareto (or Multi Objective) Optimization (PO or MOO) is the chosen paradigm to study a series of diverse systems, from graphs and networks to linguistics to the elaboration of efficient mathematical models about our physical world. All these and other relevant systems often attend to several constraints at the same time. We hypothesize that certain universal features might be captured by the underlying mathematics of a Pareto optimization process alone. We further hypothesize that, despite this unifying framework, those universal features have different realizations depending on the nature of each of the systems investigated. Accordingly, specific hypothesis will be stated in each of the following chapters, always stressing how they emanate from the central picture.

1.1. Optimization is at the heart of our physical reality

Darwinian selection defines a main thread in the dynamics of natural evolution. The concept of fitness underlies the idea that biological change is a search process in a multidimensional space where selection plays the

role of an optimization algorithm [85]. But if we pay attention we find optimal traces of optimality everywhere, well before biological processes even set up. We can see this in physics in the first place, where minimization and maximization show up in almost every theory. All systems relax to their state of minimal energy. Given a temperature, mixtures in thermodynamic equilibrium tend to their maximum entropy configuration [205]. Mechanical systems follow *least effort* geodesics across space in classic mechanics [185] or across space-time if Einstein's relativity steps in [103] – and thus light travels following its shortest path.

In all these cases a single *action* is minimized, which implies optimal Lagrangian or Hamiltonian dynamics. These are "well behaved" examples that allow us to compress our knowledge of the world within well posed functions. Several conflicting ingredients might contribute to those optimal *actions* or *energies* (think, e.g., in the potential and kinetic terms). But they enter the formulas in a definite way often dictated by first principles. Theoretical insights follow from the minimization or maximization of a relatively simple equation. Classic *derivatives* and *variational analysis* already suffice to draw a good picture of our physical world... just often enough.

We can feel Nature's urge for optimality percolating up to other sciences, even if the maths get more complicated. Chemistry is well grounded in thermodynamics, a theory of optimization par excellence as we will see later. Empirical hints of optimal dissipative systems can be found in the shape of valleys and the drainage networks of river basins [260, 262]. These remind us of vascular systems, which arguably follow optimality principles in order to distribute blood and and deliver nutrients to every cell in an organism at a minimum loss [125, 218, 345]. Optimization is such a leitmotif that it sparked heated debate about the algorithmic nature of Darwinian evolution through natural selection, mostly concerning *what targets*, if any, did evolution optimize [84]. Misreading this theory led to peculiar (not necessarily wrong) interpretations of reality, such as the Panglossianism view that we live on the best of all possible worlds [335]; but also to terrible mistakes such as the Eugenic Nazi programs.

Back to our scientific inquiry, human-made structures present traces

of optimality too. From communication networks [7, 23, 173, 277], to engineered designs [25] among many others; artificial systems seem organized around a few utilitarian principles, even if such optimality is often not planned nor enforced by us humans in any conscious way. It feels as if a Maxwellian demon sits behind a wealth of social processes, trading every bit of suboptimal craft for a more efficient configuration when we are not looking. Both these and the biological processes just mentioned are far from the mathematical simplicity of physical laws. It is unlikely that they will be described anytime soon as the extrema of a simple yet comprehensive Lagrangian. New objects might be needed to shed some light on this more elaborated kind of optimality.

Most complex systems result either from evolutionary or design processes where multiple constraints must be simultaneously satisfied [272]. This includes living as well as technological and economic systems [218, 344, 345, 125, 126, 77, 241, 47, 25, 143, 76, 23], in which very often the costs of implementing a task are confronted to its efficiency. More complicated scenarios might involve other conflicting traits as well. The parallelisms are further highlighted by the many engineering problems that are addressed through computer-based Darwinian processes, or by the suggestion that biology *is* engineering fueled by natural selection [87]. Solutions to such multiobjective optimizations entail a compromise. We do not expect to find global optima that satisfy all the needs at once. Instead, a tradeoff – a collection of solutions that satisfy the different constraints with varying degree – may be closer to what we are looking for. When trying to meet ends, there are worse and better tradeoffs, and a first task is to identify the best possible one.

As outlined above, optimization stands as a unifying principle that brings together observations from distant fields. Pareto optimality is the mathematical theory that deals with simultaneous, conflicting optimization targets. It is the framework that identifies that *best tradeoff possible*, called the Pareto front. We propose that this object, which we introduce rigorously in section 1.3.1, better contextualizes and guides us in exploring the complexity of very diverse systems. Because Pareto optimality is an abstract theory for optimization that largely ignores the details of

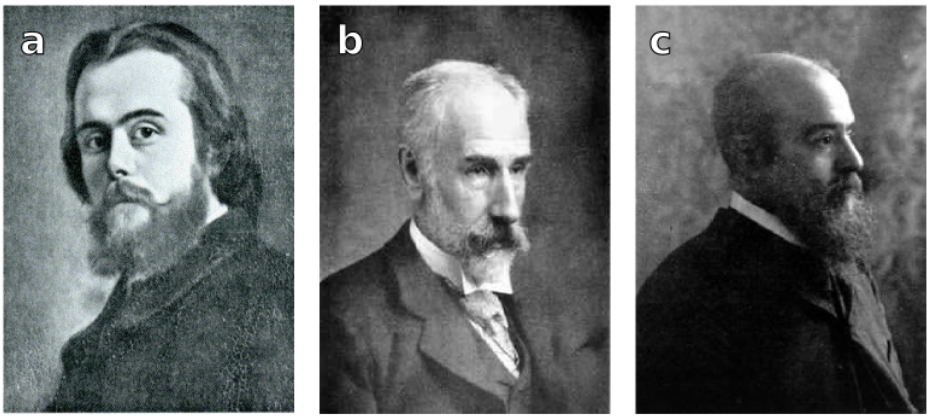


Figure 1.1: **Leading the way towards Multi Objective Optimization** **a** Léon Walras [340], **b** Francis Ysidro Edgeworth [102], and **c** Vilfredo Pareto [238]; three historical figures in the introduction of multi objective thinking in economy.

the constraints, we also conjecture that certain universal features might emanate from its mathematical structure alone.

1.2. Brief historical overview

During the XIX century, Léon Walras [340] (figure 1.1**a**) laid the mathematical basis to think about market equilibrium. He assumed a utilitarian approach: a unit would exist that measures the *utility* of goods – in other words, it is possible to bring together all the elements of a market into one single equation that sums up their values and that can be used to analyze that market's equilibrium. This strong hypothesis was often not agreed upon. Even other authors contributing to the same theory were strongly opposed to the existence of a "unit of labour, or suffering, or enjoyment" [169, 302]. However, this assumption allowed interesting developments that later became the core of economic welfare the-

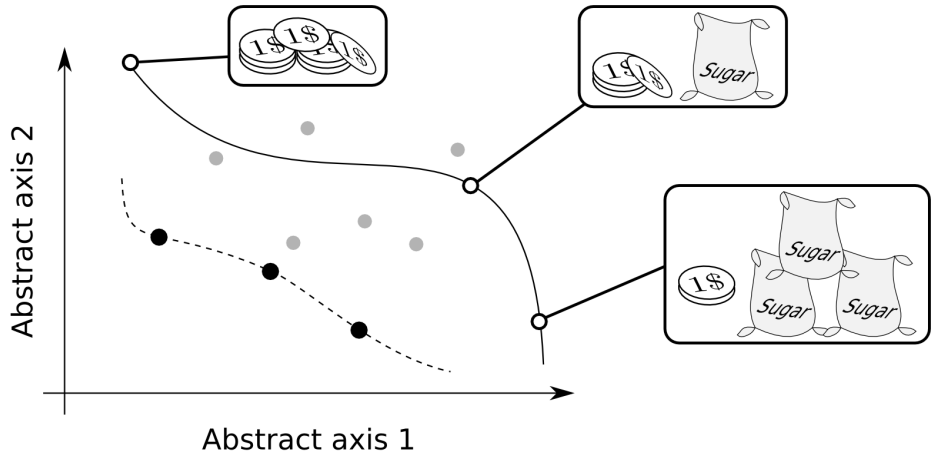


Figure 1.2: **Curve of indifferent distributions of goods.** An agent must choose between different distributions of goods, in this case represented by coins and sugar in an abstract mathematical space. This one precise agent considers *equal to its interests* any distribution of goods lying along the solid line (open circles) . Similarly, distributions of goods laying along the dashed line (black dots) are considered equal to each other by the agent. Distributions in different lines are not considered equal to each other: they are either better or worse to the agent. The same thing happens with other distributions outside the lines (gray dots).

ory. His important works were published during the 1870's in French, which explains why they were largely ignored by American mainstream economists until their translation after 1940. Meanwhile, two European scholars took up Walras's work and moved from a utilitarian perspective into a truly multiobjective one.

A first contribution was made by Francis Ysidro Edgeworth [102] (figure 1.1**b**), an Anglo-Irish economist. Son of an Irish philosopher and a Catalan political refugee, Edgeworth became Professor at Oxford university where he pursued Walras's ideas further. He introduced, among other concepts, the *indifference curve* and the *Edgeworth box*. The for-

mer, still assuming the existence of an utilitarian function, are abstract curves depicting different bundles of goods between which a consumer is indifferent (figure 1.2). So, given two consumer criteria P and π, Edgeworth talks about an abstract "point (x, y) such that in whatever direction that we take an infinitely small step, P and π do not increase together but that, while one increases, the other decreases." [101, 90]. This is the first truly multiobjective guideline for decision making, even if it refers only to neutral – not necessarily optimal – tradeoffs.

Vilfredo Pareto (1848-1923, born Wilfried Fritz Pareto to an Italian family exiled in Paris [238], figure 1.1**c**) was a civil engineer with interests in diverse fields. He only turned to economics (to a great success) late in life. He is famous not only for the concept of Pareto efficiency that we care about here, but also for other contributions such as the Pareto $(80 - 20)$ principle stating that the 80% of wealth is owned by the 20% of the population, or the relevant Pareto distribution (power law) so common in complex systems and that we will find in the next chapters too. He turned Edgeworth's box into a widely used tool to visualize Pareto optima that is taught in every Economics introductory course. His interest in social analysis brought in the concept of *elite* to the social sciences, and he had a view that democracy is just not possible because every social dynamic is indeed subjected to economic and political forces even within democratic regimes.

In the words of Benoit Mandelbrot, Pareto "was fascinated by problems of power and wealth. How do people get it? How is it distributed around society? How do those who have it use it?" [197]. This led to his empirical studies of wealth distribution. More importantly for us, those worries mixed up with optimality principles as he wondered what actions can be carried out that do not worsen the general situation. He builds on Edgeworth's work with an interesting twist: "He [Edgeworth] assumed the existence of utility, and from it he deduces the indifference curves; I instead consider as empirically given the curves of indifference, and I deduce from them all that is necessary for the theory of equilibrium without having recourse to ophelimity" [237, 302]. "Ophelimity" is how Pareto refers to utility throughout his works. He continues: "We will say

that members of a collectivity enjoy *maximum ophelimity* in a certain position when it is impossible to find a way of moving from that position very slightly in such a manner that the ophelimity enjoyed by each of the individuals of that collectivity increases (...)”.

This provides us with a first informal sketch of what Pareto efficiency is: a collection of strategies such that it is not possible to increase everybody's notion of optimality at once. Later on we will provide a sound mathematical basis to this notion. After Pareto's time the concept was ripe in economics, since it was studied in parallel by many authors (e.g. [9, 8]) often without formal reference to Pareto, who was translated into English only in the 1970's [90]. Pareto's ideas resound in multiple elements of economic welfare or game theories. This could include for instance John von Neumann's fundamental *minimax theorem* [336] proving that every finite, zero-sum, two-person game has optimal mixed strategies, and that if there is more than one optimal mixed strategy then there are infinitely many. The same can be said about *Nash equilibrium* [219, 220] defined as those situations in which, when all information about a game is available, no player can do better if she or he unilaterally changes her or his strategy. Nash equilibria are not necessarily Pareto optimal, while they capture a notion of path-dependent optimality. These and other concepts are discussed rigorously in [302], where the history of multi objective optimization is reviewed following the main mathematical contributions.

Economists developed muticriteria optimization based on an agent's interest in finding distributions of goods that satisfy them. It is the clash between agents that leads to tradeoffs. At a more fundamental level we do not need agents. Some independent (physical, biological, or economic) quantities might come in open conflict with each other. Often a set of *target functions* can be established and we might seek the simultaneous optimum to all of them. Hence, applications of MOO in engineering problems followed those of economics [311], with examples as diverse as the validation of hydrologic models [350] or the design of particle accelerators [24]. Some recent, relevant examples will be analyzed in the following section and in chapter 3.

During the last decades several numerical tools have been adapted to find those kind-of-neutral Pareto optimal solutions. Examples of such techniques are *constrained optimization* or *Multi Objective Genetic Algorithms* (MOGAs) [119, 352, 181], of which two examples will be given in section 1.3.2. Because Pareto optima are often degenerate, a lasting problem for engineers and policy makers is that of picking up just one strategy or design. Several methods have been proposed (e.g. locating knees [83, 40], which will come into focus in chapter 3). Even an international society of multi criteria decision making [208] has been created and it concedes the Edgeworth-Pareto medal in a yearly basis. Interestingly, both nature and economists often converge into bet hedging strategies [147] that, under the appropriate circumstances, allow us to transcend the limits reached by Pareto optimal solutions.

1.3. Definitions and Methods

1.3.1. Pareto optimality

Pareto Optimality is a kind of optimization posed on arbitrary sets of objects. As an instance, we may seek the best disposition of pieces to handcraft a motor engine, or the best topology to design a cheap communication network. In these examples, possible arrangements of the pieces or different combinations of links and nodes would constitute *candidate* solutions to the problem at hand. A series of features (such as cost or efficiency) can be measured on each of the candidates, and through these variables we select the best-performing configurations. The key difference with other, more standard optimization tasks is that we want our candidates to perform *better* in several traits at the same time.

More rigorously: Consider a set X of objects upon which an optimization will be enforced (figure 1.3**a**). We refer to all objects $x \in X$ as *feasible*. They constitute the *candidate solutions* or *candidate designs* and represent everything that is physically or mathematically possible within the scope of the given problem. As an instance, for crafting an engine, X is composed of all possible combinations of pieces that produce a physi-

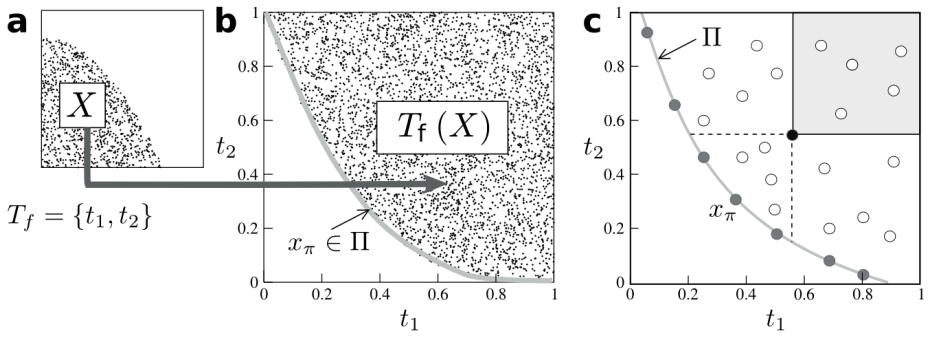


Figure 1.3: **Pareto optimality and the Pareto front. a** The abstract set X consists of all feasible objects within our design space. Upon these objects we can measure a series of *target functions* $T_f(X) = \{t_1(x), t_2(x)\}$ that map these objects into *target space* **b**, here a plane. **c** If we seek to minimize t_1 and t_2 simultaneously, a straightforward geometric condition (dominance, shaded area and black dot) discards solutions that are less optimal – i.e. they perform worse in both targets. This same condition does not allow us to resolve situations in which two solutions do not dominate each other – i.e. one of them scores better than the other in one of the targets and worse in the other one. The most honest choice is to keep those mutually non-dominated solutions. This defines the most optimal tradeoff termed *Pareto front* (**b** and **c**, $x_\pi \in Pi$ and thick gray lines), which also delimits the set of feasible designs in target space.

cally working power drive; while in designing networks, X is composed of every set of nodes and edges that our problem allows. Note that X can get further constrained if we impose, e.g., that all networks must be connected and with a fixed number of nodes. Objects outside X are *not feasible* within the scope of the problem. Insisting on the network example, not connected graphs might be deemed unfeasible.

Among all objects $x \in X$ we wish to find the subset $(x_\pi \in \Pi) \subset X$ that minimizes a set T_f of K given, real valued mathematical features called

target functions:

$$T_f(x) \equiv \{t_k(x); k = 1, ..., K\}. \tag{1.1}$$

Our task is to find those objects that score lower in all $t_k \in T_f$ simultaneously. T_f establishes a mapping between X and $\mathbb{R}^K$. We refer to this space, $\mathbb{R}^K$, as *target space* (figure 1.3**b**). Note that we have adopted minimization without loss of generality. Maximization of certain targets are trivially included by just changing the sign of the affected t_k. We adopt minimization when introducing the different theoretical objects (here and in sections 2.1, 2.2, and 2.3 of chapter 2). Some problems are more intuitively treated if some of the targets are maximized and some examples from the literature deal with mixtures of minimization and maximization. These details are not important since our contributions hold true in every case despite the nature of the optimization.

In *Single Objective Optimization* (SOO) just one global target exists and it is usually trivial to decide between candidate solutions: that with a lower score is selected before others. This way, SOO induces a *global sorting* of X. It might happen that several solutions present the same value of the SOO target; but this is often not the case, and choosing between them is trivial (just flip a coin). Because the value of the global target contains all the information that SOO cares about, this random choice does not affect any important aspect of the problem. In this sense, SOO solutions are *global optima*. Instead, in MOO, feasible designs scoring good in some t_k might score way worse in some $t_{k'}$ and viceversa. It is often difficult to choose between solutions, and picking one over the other affects our problem in relevant ways. The sorting that MOO induces in X can only be partial. We must also renounce to the idea of a global solution and assume that we will at best achieve an *optimal tradeoff* which is composed of all $(x_\pi \in \Pi) \subset X$.

We say that a candidate solution x dominates another y (and we denote it $x \prec y$) if $t_k(x) \leq t_k(y)$ for every $k = 1, ..., K$ and there is at least one k' such that $t_{k'}(x) < t_{k'}(y)$ (figure 1.3**c**). If a solution dominates another it is *objectively better* – i.e. *more optimal* – all targets considered. A solution $x \in X$ will be Pareto optimal if it does not exist any other feasible

solution $y \in X$ such that $y \prec x$. The *Pareto front* ($x_\pi \in \Pi$) is the set of all Pareto optimal solutions of an MOO. This object is mapped into $\mathbb{R}^K$ by T_f yielding a hypersurface $T_f(\Pi)$ of dimension $K-1$ or lower (figure 1.3**b-c**). We shall refer to both Π and $T_f(\Pi)$ as Pareto optimal set or Pareto front indistinctly.

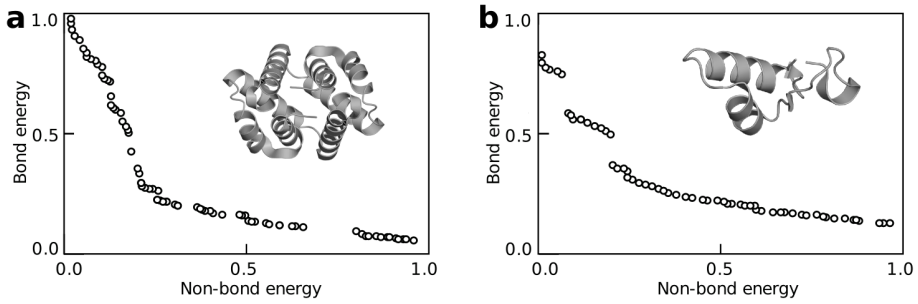


Figure 1.4: **Pareto optimal folding of proteins.** Protein folding is subjected to several conflicting physical constraints including tensions, torsions, electrostatics and others. Attempting to satisfy all these conditions simultaneously results in a Pareto front that contains protein shapes that are ‘more optimal’ in a sense, as we will review in section 3.3.3 of chapter 3. (Adapted from [78].)

In figure 1.3**b-c** we plotted a very smooth front, but more complex ones show up in real examples. Consider a Pareto optimality approach to protein structure prediction [78] (figure 1.4) – a problem that we will examine with care again later adding new insights from the theoretical framework that we will introduce in chapter 2. Atoms in a protein are subjected to both local forces (through their bounds with neighboring atoms) and mean-field forces such as the van der Waals potential emerging from coarse grained, distant atoms. Both contributions should be minimal at equilibrium. The Pareto efficient set of protein structures optimally trades between minimizing local and global forces and produces an effective ensemble of protein conformations.

In building a good model of physical phenomena there are often sev-

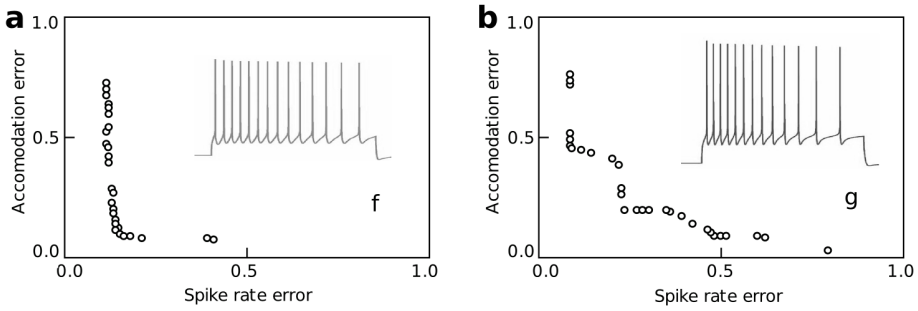


Figure 1.5: **Pareto optimal models of spiking neurons.** Pareto optimal models of spiking neurons that minimize the error in predicting the spike train firing rate and the so-called accomodation error. Each point corresponds to a possible model. All of them together reconstruct the best tradeoff possible, which emerges when we try to account for conflicting error sources with a limited mathematical representations of the real spike trains. (Adapted from [99].)

eral aspects that we would like to capture. The simpler the model, the more likely it is that we cannot account for all these traits at once. This general problem will be discussed in the context of symbolic dynamics in section 5.2 of chapter 5. With a similar philosophy, Druckmann et al. search for models of spiking neurons that simultaneously minimize their divergence with respect to different aspects of real spike trains [99] (figure 1.5). This again produces an ensemble of models whose characteristics might enlighten relevant aspects of spiking neurons. In these two and other MOO problems (see [132, 250, 89] and chapter 3) the fronts display complex shapes, suggesting inhomogeneous accessibility to different alternative Pareto optimal solutions.

1.3.2. Genetic algorithms for Multi Objective Optimization

As it happens in more standard optimization problems, genetic algorithms are of great aid in evolving a population of solutions towards more optimal configurations. The reader is referred to the literature for an exhaustive relation of genetic algorithms in MOO [119, 352, 181]. Here we review two strategies: one that stretches our intuition about tradeoffs but results in either slow or memory consuming algorithms and another one that attains faster simulations and that was consequently employed for the research presented in this thesis. In both cases we will be working with a subset of feasible designs $\hat{X} \subset X$ (a sample of all feasible solutions with size $\|\hat{X}\| = N_{pop}$) that will evolve over time – so we might also note $\hat{X}(t)$, with the initial population $\hat{X}(t = 0) \subset X$ generated by some arbitrary procedure.

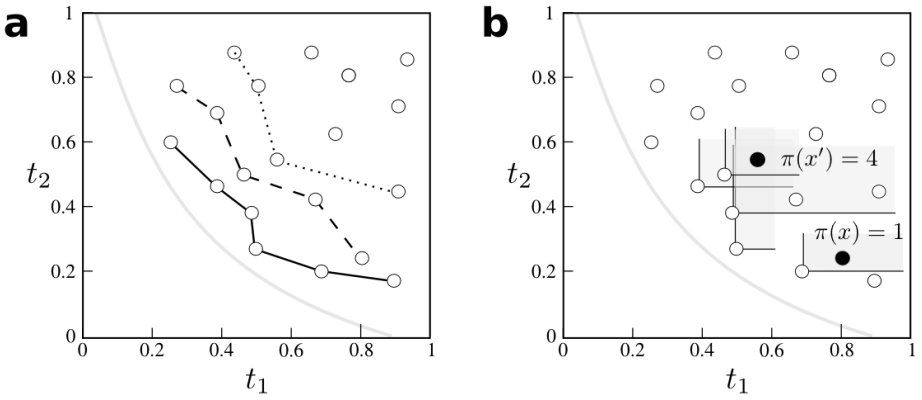


Figure 1.6: **Scores to implement a multiobjective genetic algorithm. a** The domination relationship induces layers of non-dominated designs given a sample of X. **b** Alternatively, we can score how many designs within a sample dominate each other.

For our first genetic algorithm we realize that the dominance relationship induces *layers* within $T_f(\hat{X})$ (figure 1.6**a**), such that in the first layer we find those solutions $(x^0 \in \hat{X}^0) \subset \hat{X}$ that are not dominated by any other element of $\hat{X}$ – i.e., the Pareto optimal set within $\hat{X}$. In the second layer we find those solutions $(x^1 \in \hat{X}^1) \subset (\hat{X} - \hat{X}^0)$ that are not dominated by any other member of $\hat{X}$ once $\hat{X}^0$ has been removed – i.e., again, the Pareto optimal solutions within $(\hat{X} - \hat{X}^0)$. We proceed until every $x \in \hat{X}$ has been assigned to a layer. Then we select a number of layers, say the N_L first ones, whose solutions are preserved and subjected to mutation and crossover to generate $\hat{X}(t + \Delta t)$. We proceed for a fixed number of generations or until the population is stable – i.e. when members from successive generations look alike, which can be quantified.

Those layers induce a hierarchy of tradeoffs at each generation. Each of them consists of solutions that trade between the different targets, but only one of them is optimal within $\hat{X}$. This is precisely the way in which tradeoffs can be less or more optimal than each other, and this is the way in which the Pareto front is the most optimal tradeoff.

For the second approach we realize that the dominance relationship also introduces a score $\pi(x)$ for every $x \in \hat{X}$ (figure 1.6**b**), which is given by the number of other designs $y \in \hat{X}$ that dominate x (i.e. $\pi(x) = ||\{y \in \hat{X} | \, y \prec x\}||$). Pareto optimal solutions $(x_\pi \in \Pi)$ have $\pi(x_\pi) = 0$ in every possible sampling $\hat{X}$ of X, so we proceed to minimize this score as if we were dealing with a regular SOO. The set of non-dominated solutions within $\hat{X}$ (i.e. those solutions $x^0 \in \hat{X}^0$ that belong to the first layer of the previous algorithm) score $\pi(x^0) = 0$ within the sample $\hat{X}$, hence these solutions are always selected by both algorithms. But designs within the l-th layer (with $l \neq 0$) might (and usually will) present diverse dominance scores $(\pi(x_i) \neq \pi(x_j)$ for $x_i, x_j \in \hat{X}^l)$ so that the selected solutions may (and often will) differ between both algorithms.

This second approach penalizes designs that appear near a crowded area in target space, i.e. near an area that is better sampled by $\hat{X}$. This is desirable because we usually seek a complete characterization of the Pareto front. Assigning worst scores to crowded areas promotes diversity

and often leads to a better exploration of X.

After selecting the less dominated designs (say a fixed number N_S of solutions), we proceed with mutation and crossover as usual.

Both algorithms evolve $\hat{X}$ towards Π, but convergence cannot be guaranteed. For the second algorithm we need to evaluate $N_{pop} \cdot (N_{pop} - 1)$ dominance relationships. The first strategy might need less evaluations because once a solution is dominated it falls off $\hat{X}^0$ and we do not need to track it until we are building $\hat{X}^1$ and so forth. As we approach the Pareto front, most solutions are not dominated so the number of evaluations tends to $N_{pop} \cdot (N_{pop} - 1)$ anyway. Furthermore, it would be necessary to implement an efficient tracking of the dominance relationships and how they are modified whenever non-dominated layers are removed – if we fail to do so, several designs need to be evaluated many times. Also, the strategy based on layers does not penalize crowded areas. Finally, the implementation of the second algorithm is much simpler.

Throughout this thesis a series of experiments are reported in which the second strategy has been used. Some hacks exist to reduce the number of dominance evaluations, but those improvements are not crucial for the examples that will be discussed. Different alternatives were explored involving mutation rates, crossover schemes, elite population that does not undergo mutation, etc. These and other details (e.g. population size, number of generations, convergence criteria) are explained along the text or in appendix A.1 when necessary.

1.3.3. Trait space and morphospaces

The outcome of Multi Objective Optimization usually consists of a diverse set with designs that present very different shapes and structures. The richness of such solutions often makes it difficult to represent all their salient features. However, the shape of the Pareto front and how a variety of solutions are spread over it will give us important insights about the underlying complex systems. This graphic summary induced by Pareto optimality conforms a *morphospace*.

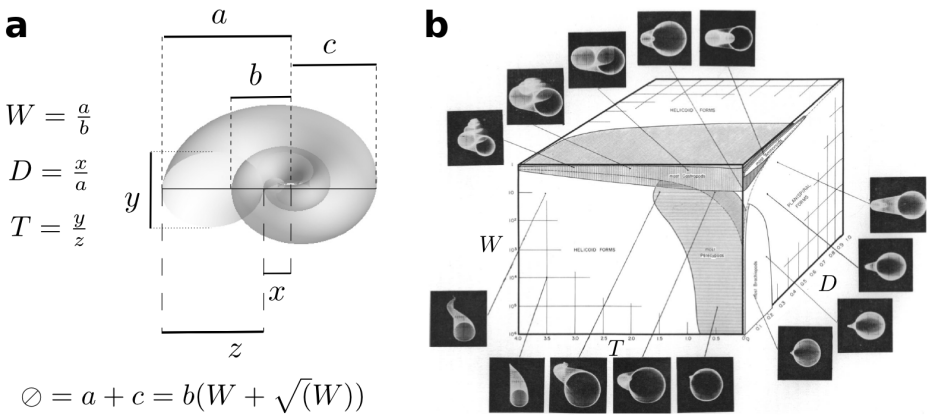


Figure 1.7: **Raup's coiling morphospace.** **a** Three parameters are enough to characterize the most relevant aspects of possible shell coilings. **b** Depending on the values of these parameters different shapes are obtained. This renders a *morphospace* where we visualize straight away how different morphologies relate to each other and where real shells are located. (Figures adapted from [323] and [255].)

Morphospaces were introduced to visualize data about biological structures in paleontology and related disciplines [209]. These spaces consist of specimens mapped onto a geometric space whose axes are given by relevant morphological, physiological, or any other kind of traits. These traits are often selected by the researchers such that the morphospace reveals important information about the organization of the specimens represented.

Raup's studies of shell coiling [255, 256, 323] used three parameters (figure 1.7**a**) to characterize every possible coiling structure. These included the known shapes of earth and sea snails or ammonites, as well as other structures that do not show up in real life. These parameters build up a 3-dimensional morphospace that we see sparsely occupied. Different individuals are located within areas of the geometric space close to morphologically similar specimens. Hence morphospaces help us visu-

alize distances between complex designs. They also suggest meaningful divisions between morphologically distant archetypes so that biological functions can be inferred. As an instance, Raup's classification reveals shapes with low drag for efficient swimming or fast-growing solutions to avoid predation, among others [323].

An interesting finding is that most of the morphospace is empty [255, 256, 268]. This has been found for other morphospaces as well, and Pareto optimality has been proposed as an explanation of this since only optimal designs should show up when evolutionary pressures are relevant [293, 320]. A good question (addressed in [323] for the shell coiling case) is: what are the conflicting traits under simultaneous evolutionary pressure? In other words: how could we read the targets of Pareto optimality off an arbitrary morphospace?

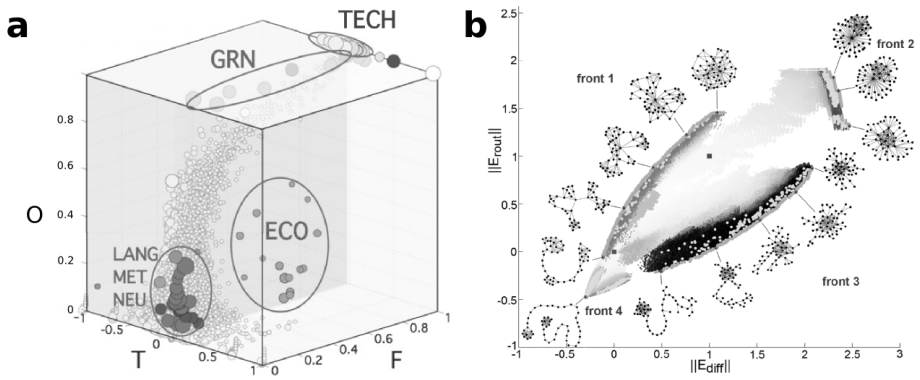


Figure 1.8: **Network morphospaces.** **a** A morphospace of complex networks with three axes measuring different traits of hierarchical systems. **b** Graphs can also capture the efficiency of communication systems. A morphospace allows us to relate different topologies to each other according to this efficiency. A series of Pareto fronts stand as a relevant frontier of feasible communication designs. (Figures adapted from [68] and [132])

Morphospaces can be naturally expanded beyond biological forms,

e.g. to complex networks, which can be sorted according to different relevant traits [12]. As an instance, we can explore three aspects that intuitively define hierarchies [65, 67, 68] : i) that they *look* tree-like, ii) that information has got a clear direction of flow (top-down), and iii) that units within the hierarchy can be ordered. This results again in a 3-D morphospace within which real networks can be located and related to each other. This gives us a very intuitive idea of how hierarchic a network looks like, and where that hierarchy might arise from (figure 1.8**a**). For instance, many naturally occurring graphs such as metabolic, neural, or linguistic networks fall within areas well populated by randomized null models, suggesting that they lack relevant hierarchy or that their hierarchical appearance results from “order for free”. Other graphs such as those representing ecological networks are set well apart of every null model, demanding an explanation for their characteristic, hierarchical organization.

Pareto fronts in target space can be taken as morphospaces too. Our optimization targets become the relevant traits at the axes of the morphospace. Also using networks as models, Joaquín Goñi et al. explore communication efficiency [132]. In their graphs this demands that messages are brought reliably from one node to another. Given a network topology, we can measure the efficiency of diffusion as a means to convey information. Goñi et al. use the average of the inverse of the first-passage time to capture this:

$$E_{diff} = \frac{1}{N(N-1)} \sum_{ij} \frac{1}{t_{ij}}, \tag{1.2}$$

where N is the number of nodes in a network and t_{ij} is the average first-passage time of a random walker between nodes i and j. Diffusion relies on information being public (we do not care that other nodes receive a message that is not intended for them) and abundant (since diffusion is a maximally random process, several copies might be needed to guarantee that our message arrives in time to the desired target). We might not be happy with this, either because we value privacy or economy. Then we

would need to *route* our messages. This is often more costly because individual nodes shall need global information about the topology of the network. In [132], routing efficiency is captured by the average of the inverse shortest path:

$$E_{rout} = \frac{1}{N(N-1)} \sum_{ij} \frac{1}{\phi_{ij}}, \tag{1.3}$$

where ϕ is the length of the shortest path connecting nodes i and j.

We can use an evolutionary algorithm to find topologies that minimize and/or maximize each possible efficiency (E_{diff} and E_{rout}). There are four possibilities: i) minimize E_{diff} and maximize E_{rout}, ii) maximize both, iii) maximize E_{diff} and minimize E_{rout}, and iv) minimize both. Each option results in a Pareto front (figure 1.8**b**) that, when mapped into target space, allows us to locate and relate singular topologies to each other.

We will come back to this example in chapter 3 to see what else we can learn from these fronts under the light of the theoretical framework that we will introduce in chapter 2.

1.3.4. A recent example in biology

Uri Alon et al. recently analyzed the consequences of a very elegant approach to evolutionary stability based on Pareto optimality [293, 292, 320, 141, 323, 321].

Darwinian evolution predicts that successful biological designs (body plans, topology of gene regulatory networks, sets of biochemical parameters, etc) correlate with increased fitness. This concept, *fitness*, is elusive and often impossible to calculate even in simple models. Assume that the fitness of a species depends on K tasks $F \equiv F(t_1, \ldots, t_K)$ in such a way that an improvement in the performance of any task always results in an increase of F. This shift from fitness to performance is very convenient, since the later can often be measured in the lab. Besides, even if the precise functional dependence $F(t_1, \ldots, t_K)$ is unknown, this poses an MOO problem with targets $T_f \equiv \{t_k, k = 1, \ldots, K\}$ that we can still

analyze. Biological designs which are not Pareto optimal are predicted to be removed by some Pareto optimal biological solution. Hence, when mapping the actual and the possible in nature most of phenotype space should be empty, as it is often the case [255, 256, 268]. When dealing with large numbers of biological data this amounts to a marked dimensionality reduction.

Given a series of biological designs, we could measure $(t_1, \ldots, t_K)$ for each one of them. These points should reconstruct the Pareto front in the K-dimensional target space introduced in section 1.3.1. We recall that this is a *morphospace* in which designs are arranged according to some salient features (in this case, the target functions). There are other possible choices of morphospace that can be very informative too. Indeed, the chances are that given a biological data set we are not sure about what variables are under evolutionary pressure. Alon et al. [293, 292] wonder about the shape of the Pareto front when projected in alternative morphospaces – most prominently, in trait space; where traits are a series of morphological, physiological, or any other kind of biological data readily available. A relevant problem, then, is to extract the optimization targets from these arbitrary morphospaces.

Assume that we can measure a series of M traits $\tau_1, \cdots, \tau_M$. These can be anything such as body size, mass, surface, shape of wings or limbs, metabolic efficiency, biomass consumption, gene expression, etc. The final choice depends on the specific problem and the available data. Each collection of these M measures corresponds to a phenotype, a dot in *phenotype space*, i.e. the relevant morphospace that we wish to analyze now. Let us make a few assumptions: i) For each task t_k there is one and only one phenotype that maximizes it. We note this phenotype ν_k and refer to it as an *archetype*. ii) The performance of any other phenotype in the k-th task diminishes with some distance (e.g. Euclidean) to ν_k in the M-dimensional phenotype space. These conditions are relaxed in [292], but they seem to hold when analyzing a series of biological data sets [293, 292, 321, 141].

If there are only two relevant optimization targets (t_1 and t_2), these correspond to just two archetypes (ν_1 and ν_2) in phenotype space. As-

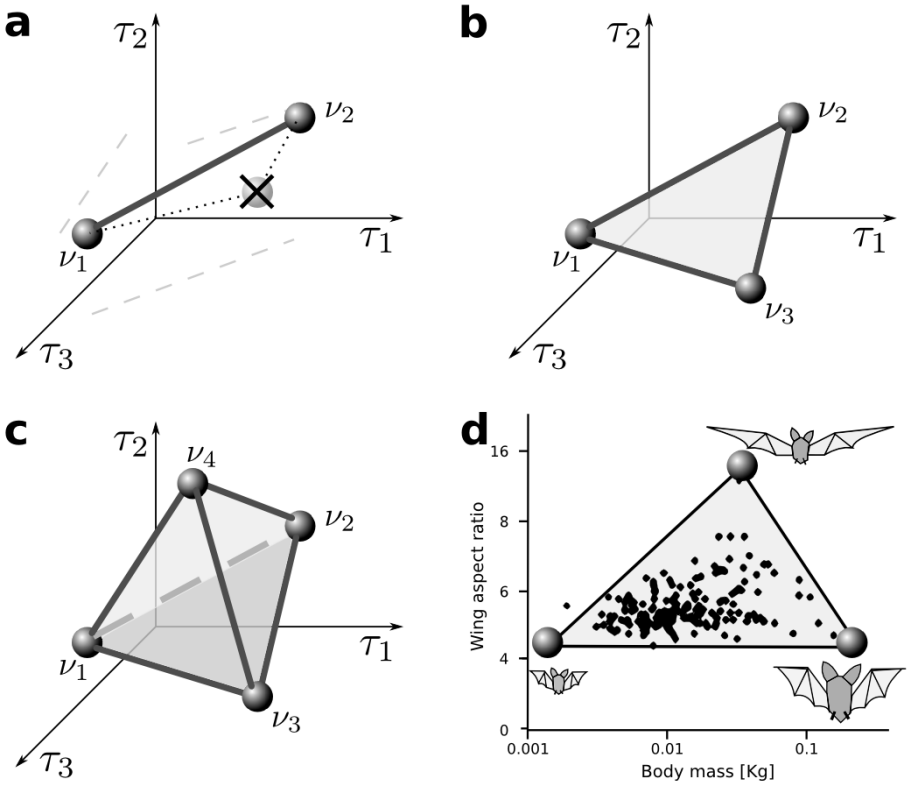


Figure 1.9: **Pareto front in phenotype space.** **a** Optimality with respect to each of the targets decreases with the distance to the archetype optimal for that target. This implies that the Pareto front looks like a straight line when plotted in phenotype space [293, 292]. Designs off this line (crossed phenotype) are not Pareto optimal and get replaced by some better solution. **b**, **c** The straight line generalizes to a polytope with K vertices, where K is the relevant number of archetypes – i.e. of underlying target functions, which remain unknown. **d** Using the software introduced in [141], a data set measuring diverse traits on several species of bats was analyzed [293]. The data consistently collapses into a triangle, indicating that three traits are simultaneously optimized. This also implies that most of phenotype space will be empty due to Pareto optimality. (Figures adapted from [293].)

suming the performance decays with Euclidean distance, any phenotype off the straight line connecting the two archetypes will be Pareto dominated by at least one phenotype along that line – hence that line is the projection of the Pareto front in phenotype space (figure 1.9**a**). Using non-Euclidean norms renders Pareto fronts with slight curvatures. The important message is that most of the M-dimensional space should be empty, with all the data collapsing into a one-dimensional manifold.

This is so when only two optimization targets are relevant. If three constraints are involved and Euclidean distances are used the Pareto front corresponds to a triangle connecting the three archetypes in phenotype space [293, 292] (figure 1.9**b**). Again, norms alternative to the Euclidean one render slightly deformed triangles. If more optimization targets become relevant, the Pareto front remains a polytope: the hypervolume enclosed by the archetypes in phenotype space (figure 1.9**c**).

It is shown for diverse, high-dimensional biological data sets how they collapse into low dimensional subsets of phenotype space (figure 1.9**d**). The examples range from morphological traits of different animals [293, 323], to genetic regulation efficiency [293, 320], to tradeoffs between mass and life span [321]. A relevant example is given by the morphology or ammonite shells [323]: they are shown to repopulate the Pareto front after major extinctions supporting that Pareto optimality is behind convergent evolutionary solutions.

Finding the polytope that better explains a dataset is a problem of statistical inference. We might vary not only the shape of the polytope, but also the number of goals (i.e. of polytope vertexes) that, we assume, might be driving the evolution of designs. Alon et al. released a valuable software that allows us to perform such inference in arbitrary biological data sets [141]. The importance of this method is that it suggests, with statistical significance, which might be the relevant constraints of the problem – i.e. it reveals the task optimized by each archetype. As an instance, analysis of bat morphology (figure 1.9**d**) results in three archetypes that strongly correlate with i) preying small insects near vegetation, ii) preying large insects while flying high, and iii) preying larger animals near vegetation [293].

Chapter 2

THE PARETO FRONT AND THE GIBBS SURFACE – PHASES, PHASE TRANSITIONS, AND CRITICALITY

In this chapter we work out the connection between the Pareto front and the Gibbs surface. This is a relevant mathematical object in thermodynamics in which phases, phase transitions, and critical points are easily revealed. We uncover that phases, phase transitions, and criticality in MOO problems share a same mathematical structure that emanates from the geometry of the Pareto front – irrespective of whether our MOO is a thermodynamic problem or not. These are, hence, the kind of universal features that Pareto optimal designs might share. As a safety test that our framework correctly reproduces well established results, we derive thermodynamics using the PO mindset. With that same MOO approach the critical point and the second order phase transition of the Ising model and the first order phase transition of the Potts model are derived.

2.1. Collapsing MOO into the simplest SOO problem

The most intuitive approach when dealing with multiple targets is to weight them linearly. (We term this operation an MOO-SOO collapse, since it turns the former problem into the later.) This introduces a series of metaparameters (Lagrange multipliers) whose values must be set externally by the person performing the optimization. Unless we have a very good reason to do this (e.g. if we know exactly how much each of the targets contributes to some *global energy* or *fitness*), this collapse implies *blindly* loosing the richness of the Pareto front. However, as we reveal in section 2.2, the shape of the front tightly determines the kind of features that we will encounter when analyzing such collapsed optimizations. Let us introduce some more notation before proceeding.

Given an MOO with K target functions $T_f \equiv \{t_k, k = 1, ..., K\}$ we define the simplest SOO problem by a linear combination of the targets through a set of external parameters $\Lambda \equiv \{\lambda_k; k = 1, ..., K\}$. This produces a global *energy function*:

$$\Omega(x, \Lambda) = \sum_k \lambda_k t_k(x). \tag{2.1}$$

Again, we assume that this function will be minimized. Because of this, global optima dwell at the minimums of a potential landscape. This will result in very intuitive visualizations of our optimal systems and justifies treating Ω as an energy, but this name is often just an analogy.

The minimization of Ω for a given Λ with fixed values $\lambda_k \in \Lambda$ yields one SOO problem. Hence, equation 2.1 defines not one, but a family of SOOs parameterized by Λ. We will study i) these SOOs, ii) the constraints that the Pareto front imposes to their solutions, and iii) the relationships between different SOOs of the same family. The validity of the results holds for any positive, real set Λ but for convenience: i) We take $K = 2$, which simplifies the graphic representations and already contains the most relevant features to be discussed. ii) We require $\sum_k \lambda_k = 1$ without loss of

generality. For $K = 2$ then $\lambda_1 = \lambda$, $\lambda_2 = 1 - \lambda$, and

$$\Omega = \lambda t_1 + (1 - \lambda) t_2. \tag{2.2}$$

iii) Finally, we impose $\lambda_k \neq 0 \ \forall k$, but allowing $\lambda_k = 0$ is parsimoniously integrated in our framework – we would just have a $K - 1$ MOO, or an SOO if $K = 2$.

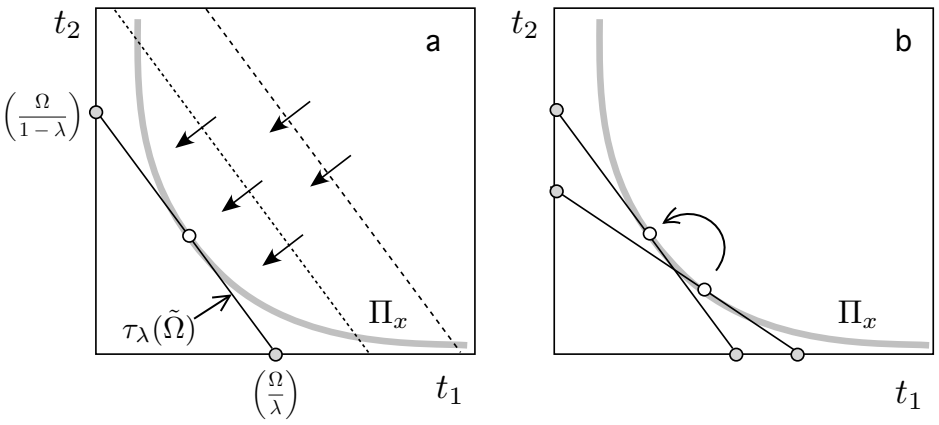


Figure 2.1: **From multi to single-objective optimization**. **a** When a linear assumption is made, the problem becomes that of minimizing a global *energy* function $\Omega(\lambda) = \lambda t_1 + (1-\lambda) t_2$ (see text). For a fixed λ one sole SOO is posed whose solution lies where $\tau_\lambda(\Omega)$ (straight lines with slope $\delta = -\lambda/(1-\lambda)$) matches the tangent of the front. **b** By changing λ, we visit other solutions of the same SOO family.

As said before, for given values of λ_k one definite SOO problem is posed. Then, equation 2.1 with fixed Ω defines *isoenergetic* surfaces noted $\tau_\Lambda(\Omega)$. Each $\tau_\Lambda(\Omega)$ constitutes a $K - 1$ dimensional hyperplane in the K-dimensional target space. For $K = 2$ (figure 2.1**a**) these surfaces

are defined as:

$$\tau_\lambda(\Omega) \equiv \left\{ (t_1, t_2) \mid t_2 = \frac{\Omega}{1-\lambda} - \frac{\lambda}{1-\lambda} t_1 \right\}. \tag{2.3}$$

This $\tau_\lambda(\Omega)$ for $K = 2$ means that, for a fixed λ, all solutions laying on the same straight line defined by equations 2.3 have the same energy Ω (i.e. they are equally optimal regarding the minimization of Ω). Solutions with lower or higher values of Ω for the same λ lay also in straight lines parallel to the original one. For general $K \geq 2$, the slope of $\tau_\Lambda(\Omega)$ along each possible direction $\hat{t}_k$ in the target space only depends on Λ so that different $\tau_\Lambda(\Omega)$ for a given SOO problem are parallel to each other (figure 2.1**a**). In particular for $K = 2$, we read the slope from equation 2.3:

$$\left.\frac{dt_2}{dt_1}\right|_{\tau_\lambda(\Omega)} = -\frac{\lambda}{1-\lambda}. \tag{2.4}$$

The crossing of $\tau_\Lambda(\Omega)$ with each axis $\hat{t}_k$ is proportional to Ω (figure 2.1**a**). With λ_k given and constant, minimizing Ω means finding $\tau_\Lambda(\tilde{\Omega})$ with $\tilde{\Omega}$ the lowest value possible such that $\tau_\Lambda(\tilde{\Omega})$ still intersects the Pareto front. Graphically, this is equivalent to *pushing* the isoenergetic surfaces against the Pareto front as much as possible (figure 2.1**a**). Hyperplanes with lower Ω exist, but the Pareto front sets the limit of feasibility: any solution with $\Omega < \tilde{\Omega}$ cannot be physically realized. The SOO optimum always lays on the Pareto front. (Take $z \notin \Pi$, then $\exists x \in \Pi,\ x \prec z$; thus at least for one $k' \in \{1, ..., K\}$ we have $t_{k'}(x) < t_{k'}(z)$, implying $\Omega(x, \Lambda) < \Omega(z, \Lambda)$ and z cannot be SOO optimal.)

The SOO optimum usually lays at the point $x_\Lambda \in \Pi$ where $\tau_\Lambda(\tilde{\Omega})$ is tangent to the Pareto front (figure 2.1**a**). Exceptions to this constitute the most interesting cases, as we will see below. The solution to different SOOs (defined by different values of λ) are found at different points along the front (figure 2.1**b**). The relationships within a family of SOO problems is thus partly encoded in the geometry of that surface.

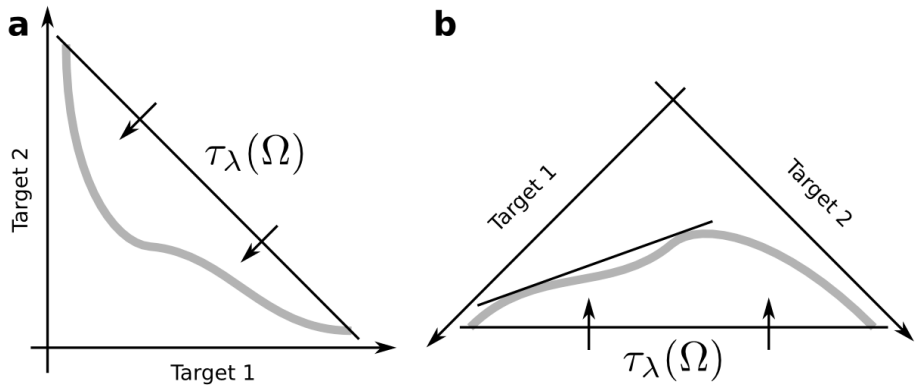


Figure 2.2: **A notion of up-down is necessary to define convexity**. Convexity and cavities are defined with respect to the direction of improvement of Ω given λ. This way, the criteria for convexity are consistent disregard of whether we deal with maximization or with problems that mix minimization and maximization.

2.1.1. Concavity/convexity and order parameters

The results that follow rely on a notion of concavity and convexity. The target surfaces $\tau_\Lambda(\Omega)$ introduce a preferred direction along which minimization proceeds. This provides a notion of *more* and *less* optimal (*lower* and *higher* energy) so that concavities and convexities are consistently defined (figure 2.2). Thanks to this definition of concavity/convexity we do not need to change the words we employ depending on whether we face maximization or minimization problems.

The next sections deal with phase transitions that are reflected in *order parameters* (that we note θ). These represent some quality of our optimal designs that varies as a response to *control parameters* – these will be the biases $\{\lambda_j\}$. When phase transitions are present, the different θ change in very characteristic ways.

As order parameters we admit any physical, geometrical, topological,

or any other features that we can measure on the elements of X. We just require i) that they make different phases distinguishable (they would be poor order parameters otherwise) and ii) that non-trivial behaviors arise out of the optimization dynamics exclusively – if not, we might encounter order parameters that become singular for some mathematical reason not relevant to our study. If the chosen indicators obey these conditions, then the singularities that we call phase transitions arise *for all* order parameters simultaneously.

For an arbitrary order parameter θ the first condition implies that if $x, y \in X$ are mapped into the same point ($T_f(x) = T_f(y)$), then $\theta(x) = \theta(y)$. The opposite ($\theta(x) \neq \theta(y) \Rightarrow T_f(x) \neq T_f(y)$) is only required whenever $x \nprec y$ and $y \nprec x$. This last condition guarantees that two points with different values of the order parameter are never mapped into the same point of the Pareto front in $\mathbb{R}^K$. The second condition is satisfied for θ such that $x, y \in X$ with $T_f(x) = T_f(y) + D\mathbb{R}^K$ implies $\theta(x) = \theta(y) + D\theta$, $D\mathbb{R}^K$ and $D\theta$ standing for arbitrary differential modifications. Then θ will not present non-analyticities other than those revealed by the theory above. Following these conditions the $t_k(x)$ themselves are valid order parameters.

2.2. Phase transitions in the Pareto front

The most simple interplay between our MOO and the corresponding family of SOOs happens when: i) the Pareto front is convex and ii) its tangent in the $\hat{t}_1 - \hat{t}_2$ plane is well defined in its interior and its slope spans the interval $(-\infty, 0)$ (figure 2.1**a**). Then, the solution to the SOO posed by a given λ is always found where the Pareto front has slope $\delta = -\lambda/(1-\lambda)$ and $\tau_\lambda(\tilde{\Omega})$ matches the tangent of the front. A differential increase $\lambda \to \lambda + D\lambda$ modifies the slope of the $\tau_\lambda(\tilde{\Omega})$:

$$\delta \quad \to \quad \delta + \frac{2\lambda - 1}{(1-\lambda)^2} D\lambda. \tag{2.5}$$

For $\lambda \in (0, 1)$, $\delta \in (0, -\infty)$. Each λ poses an SOO with a different solution. Varying λ, successive SOO solutions *roll smoothly* over the front

(figure 2.1**b**). This is similar to laying a rigid straight line ($\tau_\lambda(\tilde{\Omega})$, indeed) against the front and reading the solution for different inclinations of the $\tau_\lambda(\tilde{\Omega})$ at the contact point between that rigid line and the front. Any order parameter θ renders a continuous, differentiable function of λ.

2.2.1. Second order phase transitions

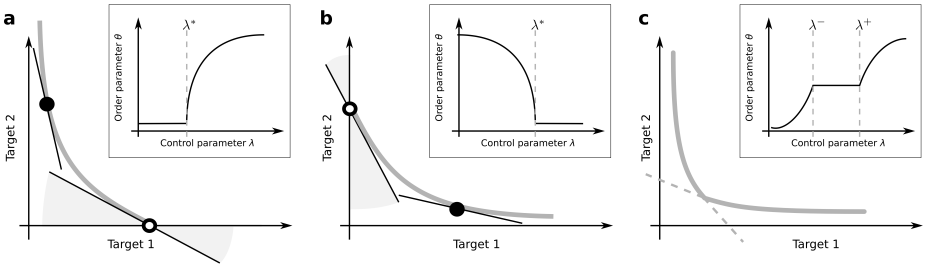


Figure 2.3: **Convex Pareto front with a tangent whose slope does not span the whole range** $(-\infty, 0)$**. a** The slope of the Pareto front spans $d \in (-\infty, \delta^*)$. The front *ends abruptly* at its bottom-right. There is a range ($\lambda < \lambda^*$, with $\lambda^* = -\delta^*/(1-\delta^*)$ indicated by the gray fan) for which well defined SOO problems exist, whose solution is persistently the same (open circle). For $\lambda > \lambda^*$ (filled circle) the front is sampled gently as in figure 2.1**b**. Any order parameter θ (inset) does not change if $\lambda < \lambda^*$ because the SOO optimum remains the same. Its derivative is not zero for $\lambda > \lambda^*$. This causes an abrupt shift in $\frac{d\theta}{d\lambda}$ at λ^* while $\theta(\lambda)$ remains continuous. **b** The exact same situation happens if the pathology is found at the top- left of the front. **c** A sharp edge is associated with two discontinuities in the derivative of any order parameter.

In figure 2.3**a, b**, and **c** we represent convex Pareto fronts whose slopes span $\delta \in (-\infty, \delta^*)$, $\delta \in (\delta^*, 0)$, and $\delta \in (-\infty, \delta^-) \cup (\delta^+, 0)$ respectively (with $-\infty < \delta^*, \delta^-, \delta^+ < 0$ and $\delta^- < \delta^+$). In all three cases we find convex stretches of the front with well defined tangents limited by points with sharp edges. In figures 2.3**a** and **b** the sharp edges happen right where the

Pareto front terminates – we say in such cases that the front *ends abruptly*. Using $\lambda = -\delta/(1-\delta)$ we reveal the intervals $\lambda \in (\lambda^*, 1)$, $\lambda \in (0, \lambda^*)$, and $\lambda \in (0, \lambda^+) \cup (\lambda^-, 1)$ respectively (note that $\lambda^- > \lambda^+$). For these well-behaved intervals a series of SOO problems exist whose solutions are always found where $\tau_\lambda(\tilde{\Omega})$ matches the tangent of the front. These can be smoothly visited as λ changes infinitesimally slow, just like before.

Consider now figure 2.3**a** for $\lambda \in (0, \lambda^*)$ – i.e. λ outside of the well-behaved range. SOO problems are perfectly defined for these values of λ, but this Pareto front ends abruptly so that nowhere in the front can we find a slope $\delta = -\lambda/(1-\lambda)$ for $\lambda \in (0, \lambda^*)$. We can graphically follow the minimization of Ω for these values of λ, by moving the corresponding $\tau_\lambda(\Omega)$ towards the Pareto front as much as possible. We realize then that the solution to all these SOOs is the same. This is indicated by the gray fan in figure 2.3**a**: several isoenergetic surfaces ($\tau_\lambda(\Omega)$) with different inclinations have been pushed all the way against the front arriving to one same optimum. This happens also in figures 2.3**b-c**: several isoenergetic surfaces (those with $\lambda \in (\lambda^*, 1)$ and $\lambda \in (\lambda^+, \lambda^-)$ respectively) reach the same solution when pushed against the front. As we vary λ within these *anomalous* intervals any order parameter remains unchanged ($d\theta/d\lambda = 0$) because we report persistently the same solution. But these same order parameters change at non-zero rates as we approach $\lambda^{*,\pm}$ from the well-behaved ranges. Because every point of the front can be reached for some λ, plotting θ (figure 2.3, insets) produces a continuous curve with a discontinuity in the derivatives. This is the fingerprint of a second order phase transition.

2.2.2. First order phase transitions

In a fully concave front (figure 2.4**a**), the straight line that joins both ends of the front has slope δ^* and defines a singular value $\lambda^* \equiv \frac{-\delta^*}{1-\delta^*}$. The solution to any SOO with $\lambda < \lambda^*$ sits at the bottom-right end of the front. For $\lambda > \lambda^*$ the solution lays at the top-left end. Both extremes coexist for $\lambda = \lambda^*$. We cannot roll smoothly over such front by varying λ. A sudden shift between radically different optima happens at λ^*. Any

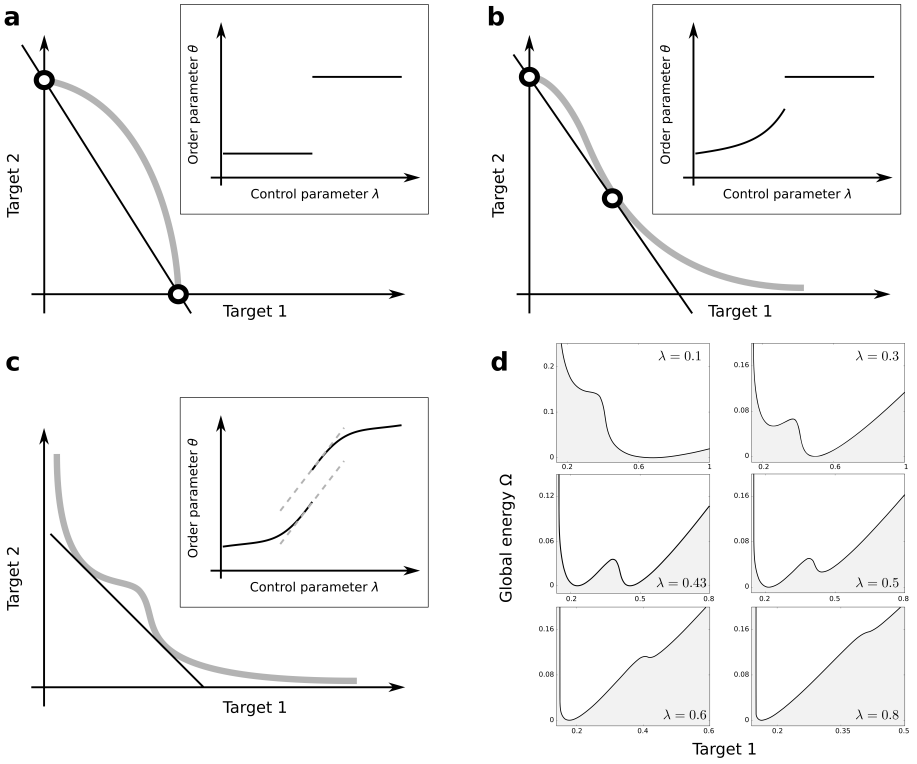


Figure 2.4: **Concave Pareto front, or fronts with concavities. a** Only two solutions are ever SOO global optima in a concave front: one if $\lambda > \lambda^*$ and another if $\lambda < \lambda^*$. For $\lambda = \lambda^*$ both solutions coexist. Any order parameter presents a sharp discontinuity at $\lambda = \lambda^*$. A similar situation happens in **b** and **c**. In the later case, while $\theta(\lambda)$ is not continuous, its derivative is. **d** An *energetic landscape potential* is built through equation 2.1. Plotting this function for all the points of the front in **c** reveals an energetic boundary below which no solutions exist. Pareto suboptimal solutions lay above the boundary and SOO optima at fixed λ sit at the bottom of energy wells. Metastable solutions are associated to local minimums and lead to hysteresis if we change λ back and forth.

order parameter remains constant below and above λ^* but a gap exists between both constant values (figure 2.4**a**, inset), as in first order phase transitions. Similar gaps are revealed when concavities are embedded within convex stretches of the front (figure 2.4**b-c**). SOO optima always lay on the convex hull of the Pareto front. Pareto optimal designs inside the cavity might be metastable – i.e. local optima – but are never global SOO optima. Metastability leads to the existence of hysteresis loops.

A convenient illustration of first order phase transitions comes through the energetic landscape enforced by $\Omega(x, \lambda)$ (figure 2.4**d**). For different values of λ we computed $\Omega(x, \lambda)$ for every Pareto optimal solution of the front in figure 2.4**c**. This renders a lower bound (thick black curves) below which feasible solutions do not exist (gray areas cannot be accessed). Heavy marbles rolling down the potential wells minimize their energy. Metastability and hysteresis dynamics are due to changes in the potential landscape (i.e. in the underlying SOO problem): As λ varies a new pocket becomes locally stable (figure 2.4**d**, $\lambda = 0.3$) and grows until it becomes the global minimum ($\lambda = 0.43$ through $\lambda = 0.5$). We might get stuck in that local minimum until it is destabilized ($\lambda = 0.8$).

2.3. Criticality in the Pareto front

Critical systems are characterized by physical quantities – i.e. order parameters – that diverge as $\theta \sim 1/|\lambda - \lambda^c|^\delta$ when $\lambda \to \lambda^c$. Criticality often requires a careful handling of control parameters, e.g. percolation probability must be set to $P = P^c$ or water becomes opalescent only at (p^c, T^c). Despite this fine tuning problem, evidence for criticality is common in complex systems, including written texts [351], populations in cities [11] or cascading events [192, 27, 189, 139]. While the hypothesis that some complex systems might be poised to criticality [175, 215] has been controversial, it is supported by several existing mechanisms known to induce power-laws in a robust manner [14, 15, 17, 45, 46, 214, 222]. This idea has been associated with potential evolutionary paths leading to optimality, but a deep understanding of criticality in evolved structures is

still largely missing. Thanks to the Pareto formalism, this section attempts to shed some light about the basis of such a connection.

We review now what it means to be critical within the Pareto formalism. This follows naturally from the geometry of the Pareto front, just as phase transitions did. In chapter 3 we will use this theoretical basis to argue that certain Pareto optimal designs must look critical *always*, and that a series of algorithms or natural forces may evolve some kind of systems robustly towards their critical states.

The change of an order parameter θ as control parameters vary is captured by a susceptibility:

$$\chi_{\theta\lambda_k} = \partial\theta/\partial\lambda_k. \tag{2.6}$$

The divergence of order parameters when a critical point is approached is mirrored by a divergence of the corresponding susceptibility. Usually we have several $\lambda_k \in \Lambda$, so that many susceptibilities are relevant in describing our system. For $K = 2$ we have just one order parameter hence only one susceptibility ($\chi_{\theta\lambda} = d\theta/d\lambda$) matters. With some algebra, we can compute it and relate it to the geometry of the Pareto front:

$$\chi_{\theta\lambda} = \frac{d\theta}{d\lambda} = \left(\frac{\partial\theta}{\partial t_1} + \frac{\partial\theta}{\partial t_2}\frac{dt_2}{dt_1}\right)\frac{dt_1}{d\lambda}. \tag{2.7}$$

The key is solving $dt_1/d\lambda$, for which we parameterize the Pareto front as a function of its slope δ. We also use $dt_2/dt_1|_{\tau_\lambda(\Omega)} = -\lambda/(1-\lambda)$ to get:

$$\frac{d\theta}{d\lambda} = \left(\frac{\partial\theta}{\partial t_1} - \frac{\lambda}{1-\lambda}\frac{\partial\theta}{\partial t_2}\right)\frac{dt_1}{d\delta}\frac{d\delta}{d\lambda}. \tag{2.8}$$

Assuming that there are not any cavities, the slope of the front matches that of $\tau_\lambda(\Omega)$, hence $\delta = -\lambda/(1-\lambda)$ and $d\delta/d\lambda = -1/(1-\lambda)^2$:

$$\frac{d\theta}{d\lambda} = \left(\frac{\partial\theta}{\partial t_1} - \frac{\lambda}{1-\lambda}\frac{\partial\theta}{\partial t_2}\right)\frac{-1}{(1-\lambda)^2} \times \frac{1}{\frac{d\delta}{dt_1}}. \tag{2.9}$$

At the same time, as long as the Pareto front behaves well from an analytic perspective, $\delta \equiv dt_2/dt_1$ so that:

$$\frac{d\theta}{d\lambda} = \frac{-1}{(1-\lambda)^2}\left(\frac{\partial\theta}{\partial t_1} - \frac{\lambda}{1-\lambda}\frac{\partial\theta}{\partial t_2}\right) \times \frac{1}{\frac{d^2t_2}{dt_1^2}}. \tag{2.10}$$

There are three reasons why this susceptibility could diverge. Two of them are not interesting for us. Namely, if $\lambda \to 1$ or if either $\frac{\partial\theta}{\partial t_1}$ or $\frac{\partial\theta}{\partial t_2}$ diverge (note that these last possibility depends on the nature of the chosen order parameter, not on the intrinsic nature of the system; and that it should be prevented by the definition of order parameters introduced in section 2.1.1). Instead, we are interested in divergences that happen whenever:

$$\left(\frac{1}{\frac{d^2t_2}{dt_1^2}} \to \infty\right) \iff \left(\frac{d^2t_2}{dt_1^2} \to 0\right). \tag{2.11}$$

By writing the Pareto front as a function $t_2 \equiv t_2(t_1)$, equation 2.11 translates into a differential geometric condition. In the most extreme case a Pareto front represents a critical system whenever it is a straight line in target space $t_2(t_1) = A + Bt_1$. This has been used in the literature to argue that certain systems trading energy for entropy in a linear way are strongly critical [215, 326, 328]. This fits our framework, especially if we acknowledge the Pareto optimality problem in thermodynamics (see section 2.2 right below). In chapters 3 and 4 we will come across other straight Pareto fronts.

We can see how more classic critical behaviors arise from this extreme situation. Consider $d^2t_2(t_1)/dt_1^2 = 0$ not for every value of t_1, but for a range $t_1 \in [t_1^-, t_1^+]$ (figures 2.5**a1** and 2.6**a1**). When rolling a rigid line over such a Pareto front, the critical range $[t_1^-, t_1^+]$ also implies a drastic rearrangement of the system happening at the critical value $\lambda^c = -\delta^c/(1-\delta^c)$, with δ^c the slope of the straight segment. Hence, besides presenting criticality (as measured by infinite divergence of any order parameter), this situation results in a gap in any order parameter (figures 2.5**b1** and 2.6**b1**).

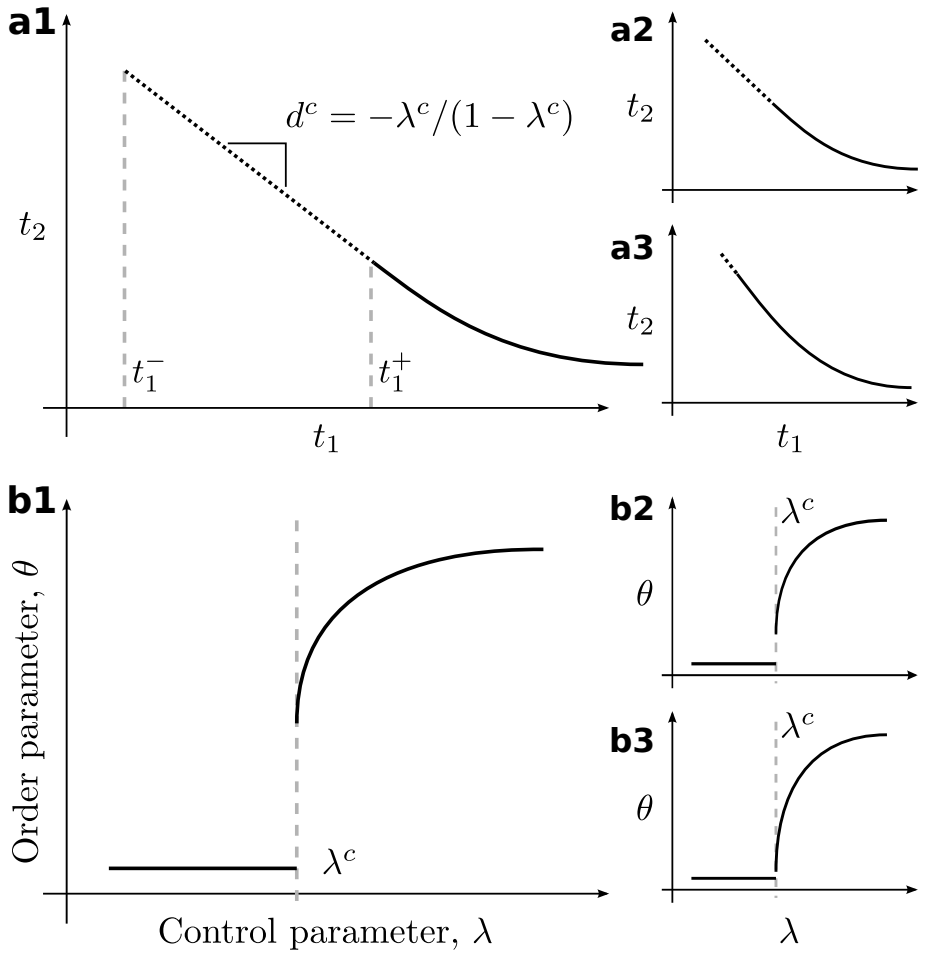


Figure 2.5: **Criticality in second order phase transitions. a** Systems critical over a range of values of the targets (here $[t_1^-, t_1^+]$) are such that t_2 is a linear function of t_1 in that range. **b** Rolling a rigid line over this front yields a first order phase transition. Besides, at the critical point any order parameter diverges. These situations containing both first order and critical elements are often referred to as *hybrid phase transitions* [96, 26]. **a1-3**, **b1-3** In this case, the limit as $t_1^- \to t_1^+$ yields a well known second order critical transition.

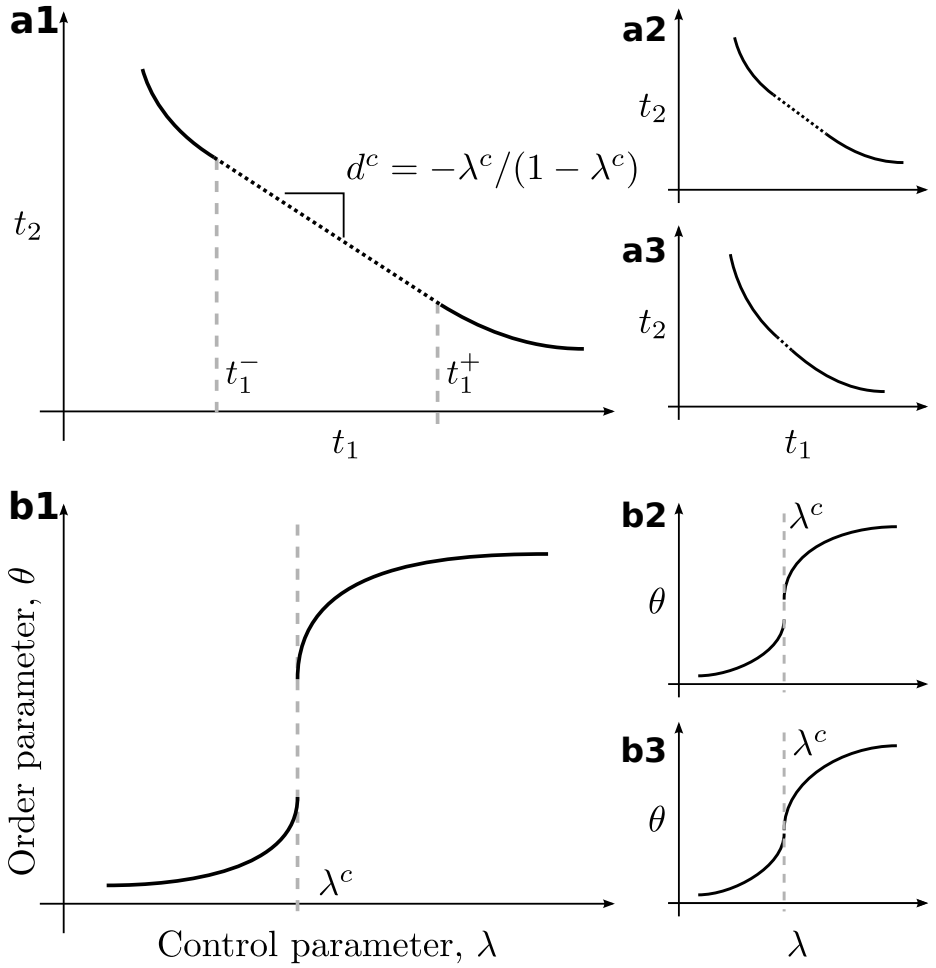


Figure 2.6: **Approaching a critical point, I. a**, **b** As before, a linear range $[t_1^-, t_1^+]$ implies a first order phase transition with a critical point at which any order parameter diverges. This again is often called a hybrid phase transition. **a1-3**, **b1-3** In this case, the limit $t_1^- \to t_1^+$ implies that the phase transition is reduced to a critical point with no phase transition associated.

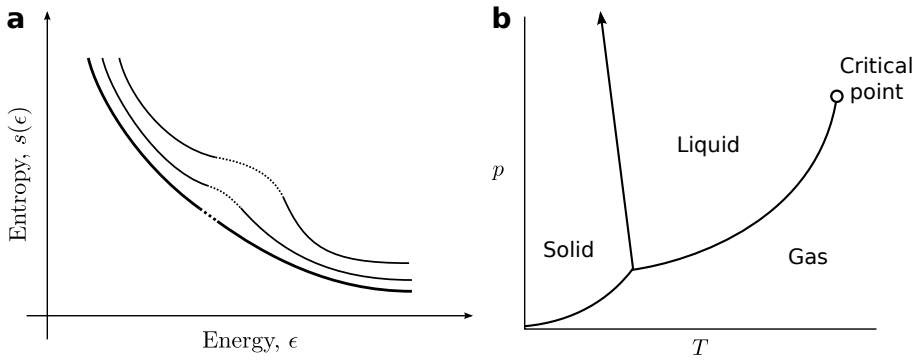


Figure 2.7: **Approaching a critical point, II. a** There is an alternative to reach the critical point in figure 2.6**a3**, **b3** without a hybrid phase transition. If three targets are involved, this situation happens whenever a cavity ceases to exist. **b** This situation reminds us of the critical point in liquid-vapor transitions.

Such behavior has been described in the literature as containing elements of both critical and first order transitions, and is often dubbed *hybrid* phase transition [96, 26]. In [309, 215], it is argued that this geometric disposition explains the criticality seemingly found in maximum entropy models of natural images.

Note that in first order transitions with a cavity we have that $\frac{d^2 t_2}{dt_1^2} \neq 0$ is well defined at either side of the cavity (i.e. at either phase), so there is no criticality in such case despite the gap in order parameters.

We can now take a limit in which the relevant range shrinks ($t_1^- \to t_1^+$), and this results either in a critical second order phase transition (Fig. 2.5**a3**, **b3**) or in a critical point with no transition associated (Figs. 2.6**a3** and **a3**). The latter case also happens if a cavity vanishes – i.e. if a first order phase transition ceases to exist. This case can be appreciated at the critical point of the liquid-gas transition (figure 2.7).

Criticality has been defined without resorting to power laws, just based on a geometric condition of the front. Of course, we would expect that

critical behavior encompassed by PO would relate to such distributions and other mathematical constructions such as fixed points of adequate renormalization schemes. We cannot offer definite answers to all these questions yet, but we will come across some power laws related to Pareto critical systems in later chapters. As an instance, if the Pareto front represents a tradeoff between energy and entropy it can be proved that the frequency of states in a probabilistic description of the system follows a generalized Zipf law [309]. We also appreciate how this representation of critical point has clear connections with neutral theory: a degenerate portion of the Pareto front becomes neutrally optimal at the same time if the condition imposed by equation 2.11 holds.

2.4. Thermodynamics as a multiobjective optimization problem

Thermodynamics emerges out of a conflict that, in the simplest case, involves energy minimization and entropy maximization – a compromise between order and disorder balanced by a temperature. It is well known that the minimization of a free energy $F = U - TS$ results in a family of solutions parameterized by T that might present common accidents such as phase transitions and critical points as T varies [166, 167]. These solutions consist of all *canonical ensembles* over which the free energy is convex, even through first order phase transitions exist. The fact that the canonical ensemble is always convex makes us wonder about the cavities that we have come across in the Pareto front.

J. W. Gibbs provided a solution by showing that canonical ensembles conform the convex hull of a surface (known as the Gibbs surface now) that might itself present cavities [128, 129, 205] – i.e. canonical ensembles are part of a wider story. Gibbs derived first order phase transitions from this surface, but not second order transitions nor critical points. His approach to thermodynamics was not widely followed, possibly because the graphic methods that he introduced posed a challenge to the technology of the time while alternative analytic calculations were already

present and allowed powerful calculations.

We proceed now to show that the Gibbs surface corresponds to the Pareto front of an MOO problem that is implicit in statistical mechanics. We are well aware that thermodynamics is a solved problem, and that many of the insights that follow are well known [128, 129, 205, 166, 167]. Although it can be (and has been) argued that the content of this section is trivial, we find many reasons to include a careful derivation of thermodynamics in MOO language anyway, even if just as an academic exercise.

First of all, it is always nice to see a same problem solved from different perspectives, check that everything fits within previous frameworks, and find out whether we can learn something new. Sometimes it happens that earlier, more complicated reasonings become simplified when analyzed under a new light; some other times, steps that were not fully understood become better contextualized.

Another reason for this exercise it that the focus of this thesis in on Pareto optimality, not on thermodynamics. By showing that statistical mechanics phase transitions can be grounded on the same framework as those of Pareto optimal systems, we wish to back the idea that the logic behind phase transitions are similar for those new systems too, and that they can be reduced to some universal mathematical properties of certain geometric objects. In our own experience, this has proven not to be easy as numerous scientists have raised concerns that our theoretical framework cannot be right, that phase transitions cannot just appear because a Pareto front happens to have a cavity or to end abruptly, or that it cannot explain phase transitions in thermodynamics. The connection between the Pareto front and the Gibbs surface has not been explicitly made in the literature as far as we know. Phase transitions and criticality are not discussed in MOO literature, to the very best of our knowledge, despite their great conceptual importance in physics and other disciplines. We feel that including this section conveys a sense of completeness about the connection between fields.

Finally, many authors get inspiration from statistical mechanics and proceed to minimize the equivalent of a free energy. These, not surpris-

ingly, are found to be convex with respect to the solutions of the optimization problem [316, 317, 314]. These works usually do not make the connection with the underlying Gibbs surface nor with the Pareto front, which we believe aid greatly in illustrating the structure of the solutions to the problem.

2.4.1. Equivalence between the Gibbs surface and the Pareto front

Take an arbitrary physical system that can occupy any state σ_j of an arbitrary space $\sigma_j \in \Sigma$. Each σ_j is a physical configuration with energy E_j. Consider an arbitrary ensemble for this system P_i, in which σ_j shows up with probability $P_i(\sigma_j)$. Consider, indeed, all possible ensembles P ($P_i \in P$), each of them an arbitrary, mathematically consistent probability distribution over the space Σ. By mathematically consistent distribution be mean that:

$$\sum_j P_i(\sigma_j) = 1, \tag{2.12}$$

We define the functions:

$$\begin{aligned} U(P_i, \Sigma) &= \sum_j P_i(\sigma_j) E_j, \\ S(P_i, \Sigma) &= -\sum_j P_i(\sigma_j) log(P_i(\sigma_j)); \end{aligned} \tag{2.13}$$

i.e. the internal energy and entropy of each ensemble P_i. These functions are rigorously defined irrespective of whether they bear any physical meaning. Since the P_i are arbitrary probability distributions there is not any guarantee (neither necessity, so far) that $U(P_i, \Sigma)$ or $S(P_i, \Sigma)$ obey any notable relationship.

These functions map $P_i \in P$ into the $U - S$ plane (i.e. into $\mathbb{R}^2$, the corresponding target space of the problem), where dominance and Pareto optimality are well defined. We can find the subset $\Pi \subset P$ of probability distributions $(P_\pi \in \Pi) \subset P$ that minimize $U(P_i, \Sigma)$ and maximize

$S(P_i, \Sigma)$ simultaneously. This is a legitimate MOO problem, again irrespective of whether it has got any physical relevance. The only difference with earlier MOOs is that one of the targets is maximized, which does not alter any of our conclusions. The solutions to this problem ($P_\pi \in \Pi$) constitute the optimal tradeoff between the targets in equation 2.13. This reduces the number of relevant ensembles for us, but still there is no guarantee nor any need that these $P_\pi \in \Pi$ present notable physical properties. They are just probability distributions solving an ad-hoc MOO.

Consider now the family of SOOs defined by:

$$\begin{aligned} &\min_P \left\{ \Omega(U, S; \lambda_U, \lambda_S) \equiv \lambda_U U + \lambda_S S \right\} \Rightarrow \\ &\Rightarrow \min_P \left\{ \hat{\Omega} \equiv \tfrac{\Omega}{\lambda_U} = U + \tfrac{\lambda_S}{\lambda_U} S \right\}. \end{aligned} \tag{2.14}$$

This collapses the original MOO into a series of SOOs whose solutions lay upon the convex hull of the Pareto front in the $U - S$ plane, as shown above, so that phase transitions arise for singular values $(\lambda_U/\lambda_S)^{*,\pm}$ due to cavities and sharp edges of the corresponding front. The quantities Ω and $\hat{\Omega}$ from equation 2.14 are naturally convex over these global minimums. Notwithstanding all these parallelisms with statistical mechanics, these are still phase transitions of a fabricated problem.

Let us connect all of this to physics by taking thermodynamics at face value. Its first law defines microcanonical ensembles as those maximizing S for a fixed value of internal energy (figure 2.8). Through equations 2.13 these ensembles can be mapped into the $U - S$ plane where they trace a curve $S \equiv S(U)$. This projection must be a function, otherwise one value of U would be assigned to two or more values of S but only one can be maximum.

The laws of thermodynamics also imply that

$$F = U - TS = U - S/\beta \tag{2.15}$$

is minimized in equilibrium at fixed temperature [205]. Thermodynamic canonical ensembles embody this minimization. Identifying $\hat{\Omega} \equiv F$ and $\lambda_S/\lambda_U \equiv -T = -1/\beta$ from equations 2.14 and 2.15, the relevant $(\lambda_U/\lambda_S)^{*,\pm}$ correspond to those temperature values at which phase transitions occur.

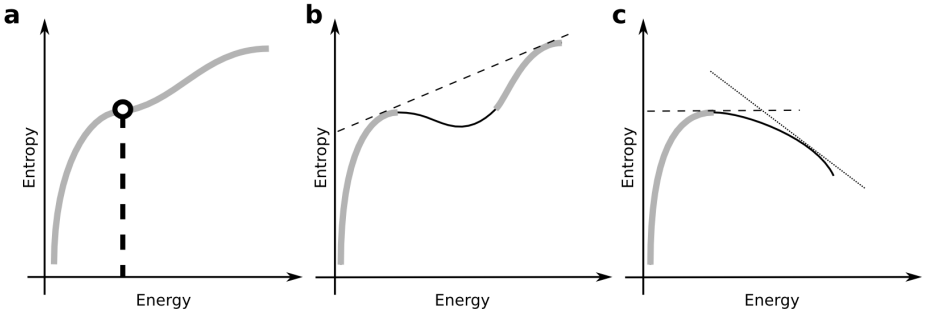


Figure 2.8: **Laws of thermodynamics and the Pareto front. a** According to the second law of thermodynamics, at constant internal energy (vertical dashed line) the microcanonical ensemble is the one that maximizes the entropy (and is hence mapped into the open circle in this example). Implementing this maximization for varying energy yields a function on the $U - S$ space that for thermodynamic systems is usually monotonously increasing with U – more energetic systems usually have more entropy. This guarantees that any two points on this curve are mutually non- dominated. There cannot be any point above this curve, thus the obtained curve must be exactly the Pareto front of the corresponding MOO problem. **b** This curve would not match the front only if the microcanonical entropy were not monotonously increasing with U. This is an odd situation in thermodynamics. These non-increasing stretches would necessarily lay inside a cavity (solid black curve) and would never show up in thermodynamic equilibrium. **c** Such situation can also happen beyond the global maximum of the entropy, which is only reached for $T = 0$ (dashed line). Points of the microcanonical entropy beyond this maximum would require $\partial S/\partial U = 1/T < 0$ (i.e. negative temperatures, dotted line). In both **b** and **c** the entropy of microcanonical ensembles still contains the whole Pareto front and, of course, its convex hull.

Irrespective of whether thermodynamics consists of an MOO-SOO collapse, canonical ensembles are constrained by the rules that reveal phase transitions in such problems. First order phase transitions are associated to cavities at which the Pareto front and its convex hull (i.e. microcanonical and canonical ensembles) must differ.

Consider the situation in which $S = S(U)$ increases monotonically. Then a greater internal energy is always associated to a greater entropy, which intuitively makes sense. This bijectivity guarantees that $S = S(U)$ matches the corresponding Pareto front. For each $U_j > U_i$ it follows $S(U_j) > S(U_i)$, thus in this curve there are not any two $U_i < U_j$ such that $S(U_i) > S(U_j)$, which would imply $U_i < U_j$. Furthermore, any ensemble mapped into a point (U_i, S_i) outside this curve is necessarily dominated by some microcanonical ensemble mapped into $(U_i, S(U_i))$ since by definition microcanonical ensembles are such that $S(U_i) > S_i$ for that given U_i. Summing up: i) points along the curve $S = S(U)$ are mutually non-dominated and ii) for any physically plausible point (U, S) outside this curve there is at least one point that belongs to the curve and dominates (U, S). This is the definition of the Pareto front.

Non-monotonically increasing entropies might show up in some mathematical idealizations of physical systems. These lead to exotic parameterizations such as negative temperatures. Furthermore they do not affect the current theory. Consider figures 2.8**b** and **c** regardless of the physical reality of such descriptions. The equilibrium thermodynamics of such hypothetical systems are still well represented by the convex hull of the Pareto front, hence our theoretical framework remains true. Given the definition of dominance, we note that the Pareto front of the relevant MOO problem is still fully reconstructed by the curve $S = S(U)$ (thick stretches in figures 2.8**b-c**). Non-increasing stretches of $S = S(U)$ lay either inside a cavity (figure 2.8**b**) or after the global maximum of the function (figure 2.8**c**). If they are inside a cavity, such situations never show up in thermodynamic equilibrium, whose canonical ensembles are strictly mapped into the convex hull of the front. These points are bypassed by a first order phase transitions. In the other situation, the slope of the Pareto front at the global maximum (whose limit from below is perfectly recon-

structed by the microcanonical ensemble) is necessarily 0, meaning that such situation is reached only at $\beta \to 0 \Rightarrow T \to +\infty$. Solutions beyond the global maximum require $\beta = 1/T < 0$, which is not realistic. Thus non monotonously increasing functions $S = S(U)$ do not affect the general framework. For MOO-SOO systems others than thermodynamics we do not rely on microcanonical ensembles, but on the Pareto front straightaway, in which non-dominated regions (including curves that would break the monotonic trend of the front) never show up.

Our work follows closely some ideas from Gibbs [128, 129, 205] that did not become so mainstream in the study of statistical mechanics. Gibbs's *graphic method* relied on the existence of a surface upon which all possible states of a thermodynamic species in equilibrium dwell. This surface is defined by the Gibbs potential $G(p,T) = U + pV - TS$, which plays the role of global energy. We identify the target functions U, S, and V and the control parameters T and p. The tangent plane at a given point of the surface is defined by a normal vector whose components are precisely related to the pressure and temperature of that equilibrium state [205].

It was argued that an MOO approach might be adequate when the different targets cannot be compared (e.g. if they have different units). In thermodynamics, the control parameters T and p transform different potentials into the same units – hence making them comparable. At fixed temperature entropy is heat and at fixed pressure volume is work – so that a low entropy and an unoccupied volume are available free energy. All free energy must be utilized to reach the thermodynamic equilibrium [205] so that $G(p,T)$ is minimized and only the convex hull of the Gibbs surface shows up. When changing T or p cavities are bypassed thus revealing first order phase transitions. Sharp edges are consistently related to second order phase transitions again.

This ingenious picture received renewed attention through the concept of *ensemble inequivalence* [334, 3, 104, 329]. The Gibbs surface is fabricated thanks to the microcanonical ensemble and it can be convex or concave, while a thermodynamic canonical ensemble can only be

convex. Whenever G becomes concave both ensembles must diverge geometrically in the $U - V - S$ space. This makes the canonical ensemble non- analytic at the inequivalence points which is reflected as a first order phase transition in the corresponding physical system. This observation is trivially true when thermodynamics is resolved from the perspective of the Pareto front. Notwithstanding, the ensemble inequivalence needed to be carefully worked out to defend its insights as late as the beginning of the XXI century, perhaps because, as noted at the very beginning of a relatively recent paper [329]:

> Most textbooks of statistical mechanics (see, e.g., [258, 156, 19, 186, 266]) have sections devoted to demonstrating that the microcanonical and canonical ensembles – the two sets of equations used to calculate the equilibrium properties of many-body systems – always give the same predictions. The arguments given are most often not actual proofs, but variations of an argument originally put forward by Gibbs in his seminal treatise [6] claiming that the canonical ensemble should be equivalent to the microcanonical ensemble in the thermodynamic limit.

This is a historically relevant case in which confusion about the intrinsic nature of the problem could be clarified by attacking it from a new angle, be it through the works in ensemble inequivalence [104, 329] or through Pareto optimality.

2.4.2. Solving the Ising and Potts models from an MOO perspective

The Ising and Potts models illustrate second and first order transitions respectively. General versions of these models have been solved using ensemble inequivalence [29, 71], which is comprehensively explained from the Pareto optimality perspective. We illustrate thermodynamic phase transitions with the Pareto front using these models because of their his-

torical importance and because they allow a complete analytical treatment.

The mean-field Ising model – a second order phase transition

We use a standard mean-field Hamiltonian for the Ising model $H_j = -\frac{J}{2}\sum_{\langle j,k\rangle} s_j s_k$ with J the coupling constant and the sum running over z neighboring spins. We parameterize the system with the probability p that we find the mean-field spin in the *up* state. It becomes easy to write down the entropy and the internal energy of the system in terms of p:

$$\begin{aligned} S &= -p\,log(p) - (1-p)\,log(1-p), \\ U &= -\frac{Jz}{2}(2p-1)^2. \end{aligned} \tag{2.16}$$

Solving this last expression for p, we can also write the entropy as a function of the internal energy alone:

$$\begin{aligned} S(U) &= -\frac{1}{2}\{1+f(U)\}\,log\left(\frac{1+f(U)}{1-f(U)}\right) \\ &\quad -log\left(\frac{1-f(U)}{2}\right), \end{aligned} \tag{2.17}$$

with:

$$f(U) \equiv +\sqrt{\frac{-2U}{Jz}}. \tag{2.18}$$

Equation 2.17 (represented in figure 2.9**a**) gives us S as a function of U ($S = S(U)$) for all possible states that the model can be found into, disregard of whether or not these states correspond to thermodynamic equilibrium situations or to microcanonical ensembles. Because for this model equation 2.17 is a function; for each U $S(U)$ is also maximal – i.e. every state of the system corresponds to a microcanonical ensemble itself. This is luckily valid for this one model, but not necessarily true in a general case.

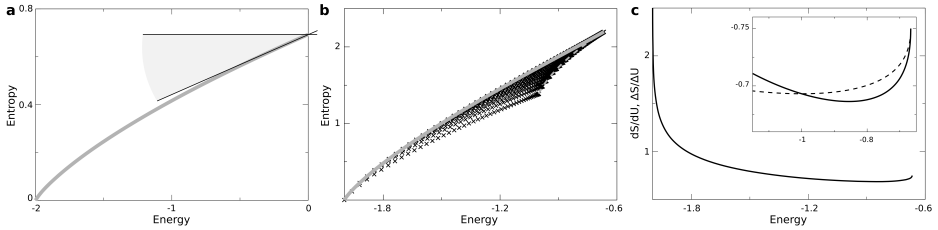


Figure 2.9: **Pareto fronts for the Ising and Potts models. a** The front of the mean-field Ising model (thick gray line) is convex but ends abruptly revealing a range $\beta \in (0, 1/Jz]$ (gray fan) for which SOOs arrive to the most entropic solution always. **b** A sample of arbitrary distributions $P = \{p_1, p_2, p_3\}$ (black crosses) for the $q = 3$ Potts model is dominated by its Pareto front whose top-right part is concave. This indicates a first order phase transition. **c** That cavity becomes noticeable when analyzing the slope of the front, which is not monotonously decreasing.

This curve also constitutes the Pareto front of the corresponding MOO problem (minimizing U and maximizing S from equations 2.16), as is expected given the correspondence between the microcanonical ensembles and Pareto optimal solutions. Let us analyze the Pareto optimality of equation 2.17. First we note that $S(U)$ is only real and well defined for $U \in [-Jz/2, 0]$, the range of available energies for the model. In this range:

$$\frac{dS}{dU} = -\frac{1}{2} f'(U) log\left(\frac{1+f(U)}{1-f(U)}\right) > 0, \tag{2.19}$$

thus $S(U)$ is monotonously increasing, which guarantees that its points in the $U-S$ plane do not dominate each other regarding energy minimization and entropy maximization. Because this curve comprises everything that is possible in the system under research and because its constituting points are mutually non-dominated, it must be the Pareto front.

Besides, $\frac{dS}{dU}$ is positive and monotonously decreasing within the range $U \in (-Jz/2, 0)$; thus there are not cavities in the Pareto front: we rule

out first order phase transitions. We can also rule out second order phase transitions in the interior of $U \in (-Jz/2, 0)$ because the derivative is well defined everywhere. Second order phase transitions are thus restricted to $U = -Jz/2$ or $U = 0$. We inspect $\frac{dS}{dU}$ as U tends to these points. After some algebra we arrive to:

$$\lim_{U \to 0^-} \frac{dS}{dU} = \frac{1}{Jz}. \tag{2.20}$$

and:

$$\lim_{U \to (-Jz/2)^+} \frac{dS}{dU} = +\infty. \tag{2.21}$$

As we saw in the results section, for the SOO posed at temperature $T = 1/\beta$ we typically reach points of the Pareto front where their tangent matches the slope of the corresponding $\tau_\beta(F) = \{(U, S)|S = \beta U - \beta F\}$. The derivative $\frac{dS}{dU}$ as we approach $U = -Jz/2$ is infinite, meaning that we will reach this end of the Pareto front only at $\beta \to \infty$ (zero temperature). There is not any remarkable behavior here. At the other end of the Pareto front the derivative is not 0, but a finite positive number. This means that already for $\beta = \frac{1}{Jz}$ the free energy optimum – i.e. the SOO solution – is located at the upper right end of the front which corresponds to the state with more entropy and energy. If we further decrease β we will not be reaching any novel solutions: the SOO optima remain the most entropic state of the system. Going back to the well behaved range of β, as we increase it above $\frac{1}{Jz}$ the SOO solutions begin to roll over the Pareto front continuously. The transition between a persistent solution for $\beta \in (0, 1/Jz]$ and a varying solution in the regime $\beta \in (1/Jz, +\infty)$ implies a discontinuity in the derivative of any order parameter. This is analogous to the cases illustrated in figure 2.3. It is associated to a second order phase transitions similar to the one known to happen in the mean-field Ising model at precisely $\beta = \frac{1}{Jz}$. Furthermore, if we calculate the second derivative over the front, we get:

$$\lim_{\beta \to (1/Jz)^+} \frac{d^2 S(U)}{dU^2} = 0, \tag{2.22}$$

meaning that the corresponding susceptibility (the heat capacity) diverges and that this phase transition presents a critical point.

The mean-field Potts model – a first order phase transition.

We repeat the same operations with the Bragg-Williams approximation to the Potts model, which has been solved elsewhere [178] using alternative methods. This choice of implementing the mean-field presents first order phase transitions for any $q \geq 3$, where q is the number of available states for each spin. For a discussion of the Braggs-Williams against other mean-field approaches to the Potts model see [349].

Following [178], we write down the entropy and energy of the system:

$$\begin{aligned} S &= -\sum_{j=1}^{q} p_j log(p_j), \\ U &= -\frac{zJ}{2}\sum_{j=1}^{q} p_j^2; \end{aligned} \tag{2.23}$$

with J and z still the coupling and the number of neighbors. Now U and S are parameterized by the probabilities p_j ($j = 1, ..., q$) of finding a spin in each of the $q \geq 3$ states. The normalization $\sum_j p_j = 1$ means that there are $q-1$ parameters and we cannot write $S = S(U)$ as before unless we make some assumption. Let us prefer one arbitrary state (say $j = 1$) over the others. Let us call p to the probability of finding a spin in that preferred state, and let us further assume that any other state is equally likely $p_{j'} = (1-p)/(q-1)$, now with $j' = 2, ..., q$. This is analytically justified in the literature [178] and later by our arguments about Pareto dominance. We note that, unlike for the Ising model, states compatible with the premises of the system will not usually be constrained to a curve because we have too many degrees of freedom. In figure 2.9**b** we represent a sample of valid points for $q = 3$: all of them can happen in theory (they are mathematically valid descriptions of the system). A few of them Pareto dominate some others thus not all of these configurations will be reached in thermodynamic equilibrium.

Thanks to the previous symmetry breaking to favor one state over the others we can write down:

$$\begin{aligned} S(U) &= -\frac{1+f_q(U)}{q} log\left(\frac{(q-1)(1+f_q(U))}{q-1-f_q(U)}\right) \\ &\quad -log\left(\frac{q-1-f_q(U)}{q(q-1)}\right). \end{aligned} \tag{2.24}$$

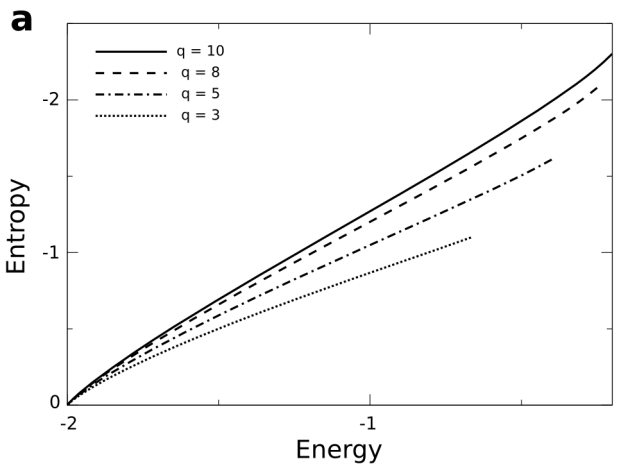
This is the counterpart of equation 2.17 only now:

$$f_q(U) \equiv \sqrt{(1-q)\left[1+\frac{2qU}{zJ}\right]}. \tag{2.25}$$

Equation 2.24 is represented in figure 2.9**b** for $q = 3$. We can appreciate that it is monotonously growing as a function of U: its points are mutually non-dominated and constitute the Pareto front. The fact that there is not any point in the previous sample that Pareto dominates any point in this curve suggests that our symmetry breaking hypothesis (favoring one spin state over the others) is correct. (It can be analytically proved that this is the correct ansatz [178].) Although it is visually difficult to appreciate, a cavity exists in the upper right part of the Pareto front. This becomes more obvious when analyzing dS/dU (figure 2.9**c**), which is not monotonously decreasing.

Most of the Pareto front is continually visited: as we vary β, the SOO solutions roll over its convex lower-left part. It can be shown that, again, the less energetic extreme of the front is reached only for $\beta \to +\infty$ ($T = 0$), so that there is not any remarkable feature there. At the other extreme of the front it exists a value β^* below which the global optimum becomes persistently the most entropic one. At exactly $\beta = \beta^*$ that solution coexists with another one in the convex part of $S(U)$.

To locate β^* we plot $\frac{dS}{dU}$ and we compare it to the slope $\frac{\Delta S}{\Delta U}$ of the straight line that connects the top-right extreme with other points along the Pareto front (figure 2.9**c**, inset). Where both functions intersect we have identified the coexisting phases. The straight line that connects these phases has slope β^* precisely. We collect β^* for the Potts models with



b

q	β_c
3	0.6932
4	0.8240
5	0.9242
6	1.0059
7	1.0751
8	1.1351
9	1.1883
10	1.2360

Figure 2.10: **Pareto front for the mean-field Potts model with different q. a** Pareto fronts of the Potts model for $q = 3, 5, 8, 10$. Although hardly noticeable, all these fronts have got concave stretches towards their upper-right ends. This indicates that all of them undergo first order phase transitions from the most disordered state to a more ordered phase where symmetry has been broken to favor just one of the states. A Pareto front for q dominates the Pareto front for every $q' < q$ indicating that this system will never leave empty one of its available states spontaneously, except in the most ordered state in which all spins are aligned. **b** Inverse temperature at which the mean-field Potts model presents its first order phase transition for $q = 3, ..., 10$. The results match perfectly those from the literature [178].

different parameter values q in figure 2.10**b**. These results match those known from the literature [178].

Because the two coexisting solutions are far away in the Pareto front, at $\beta = \beta^*$ these systems undergo a drastic change – as opposed to the continuous transition in the Ising model. This is similar to the first order phase transition situation illustrated in figure 2.4**b**.

Chapter 3

PHASE TRANSITIONS AND CRITICALITY IN PARETO OPTIMAL SYSTEMS

The formalism introduced in the previous chapter connects Pareto optimality to phase transitions and criticality. We hypothesize that these features should be prominent in most complex systems precisely because they present optimization and conflicting tradeoffs as a hallmark. This chapter and the next one illustrate the relevance of phase transitions and criticality in Pareto optimal designs. Here, complex networks are studied as a testbed for our theoretical framework. We propose that they are flexible and appropriate to illustrate phase transitions within the Pareto formalism. Additionally, optimal graphs will exemplify how Pareto selective forces must always drive certain systems towards a critical state. Because of this, we put forward Pareto optimization as a mechanism for critical self organization. The underlying theoretical framework might soften the confrontation between historically opposed positions in the literature of self-organized criticality.

Networks pervade our mathematical modeling of the reality – be it Gene Regulatory Networks, phylogeny trees, railroads, communication systems, social networks, etc – hence the resolution of Pareto optimiza-

tion problems on complex graphs presents an engineering interest of its own. Other relevant (engineering, biological, etc) PO problems have been solved in the literature and we propose that their connection to phase transitions enriches their interpretation. At the end of this chapter we review a few selected works.

3.1. Phase transitions in Pareto optimal complex networks

3.1.1. Overview of the problem

The key role of optimization outlined in previous chapters, common to both engineering problems and the intrinsic dynamics of natural evolving systems [85], can be succinctly captured by network models. The engineering perspective, associated to man-made objects and structures, is specially obvious when dealing with large-scale, interconnected units, as it occurs in very large integrated circuit design [236, 52, 25] or spatially-extended infrastructures such as power grids [342, 7, 23, 263] and transportation or distribution networks [7, 23, 173, 249, 277, 126, 47]. In these cases, abstract nodes and their connections show up as a very natural representation of the problem; and also in all these cases, engineers must cope with interfering constraints related to materials, space, packing, wiring, or dissipation costs. The staggering complexity of these designed systems can be addressed by algorithms that deal with multidimensional problems.

In biological systems, important network topologies have been shown to result from optimality [58, 211]. These include transportation networks in living organisms [218, 344, 345, 21, 125] where optimization is reached by means of fractal trees that guarantee a low cost and efficient relocation of resources. Similarly, neural circuits display optimal features over a wide range of scales [25, 77, 241, 143, 76, 343]. The packing and interconnectivity in some cortical areas seem compatible with design principles shared by high-density electronic designs [25].

In all the previous examples tradeoffs between efficiency and cost are present. Packing many components in a given spatial domain is desirable because of cost minimization of connections, but dissipation of energy or wiring constraints will also be at work. What kinds of topologies result when considering conflicting conditions? This problem has been addressed by explicitly introducing efficiency measures E (such as average path length) along with cost constraints C (such as number of connections of a given graph) [111, 63, 223]. The tension between opposite demands often leads to phase transition phenomena. We can wonder, in a more general note, when and how will phase transitions arise disregarding of the details of the problem in hand.

Within the context of network optimization, we consider the set Γ of all connected networks $\gamma \in \Gamma$ involving N nodes and any number of links. Latter on we will propose definite formulas for efficiency ($E(\gamma)$) and a cost ($C(\gamma)$) based on the structure of each network γ, thus fabricating a series of PO problems. For each such γ, we can also introduce a global *energy function* $\Omega(\gamma)$ that takes into account our optimization goals. The most straightforward way to do this is through a linear combination:

$$\Omega(\gamma, \lambda) = \lambda E(\gamma) + (1 - \lambda)C(\gamma), \tag{3.1}$$

with $\lambda \in [0, 1]$ a tunable parameter that weights the impact of each contribution, as we did in chapter 2. This is precisely the strategy in previous studies of similar problems [111, 63, 223, 112, 251, 265, 204]. If $\lambda = 1$ only efficiency constraints will be at work, whereas $\lambda = 0$ would not consider this component at all.

Such a global energy results in very illustrative visualizations of the optimization process through the notion of a *potential landscape*. Assuming minimization, more optimal network architectures lay deeper in a potential well when we plot $\Omega(\gamma, \lambda)$ for every γ superimposed on an arbitrary morphospace (figure 3.1**a**). Note, however, that this is a limited picture: a fixed value of λ is necessary to generate one potential landscape. Changing the parameter modifies the landscape rendered by $\Omega(\gamma, \lambda)$ and, accordingly, the underlaying optimization problem. To achieve a more general comprehension we should not only allow scenarios with differ-

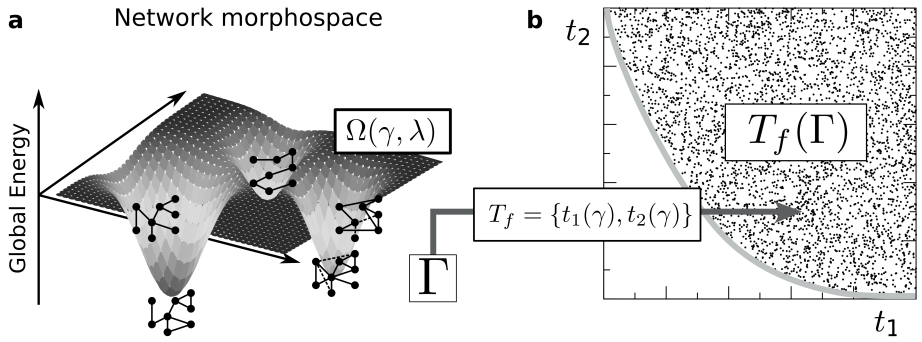


Figure 3.1: **A two dimensional example of Pareto optimality.** **a** $\gamma \in \Gamma$ are all possible connected networks with a given number of nodes. They populate some network morphospace where we seek those graphs minimizing some measurable feature. If we deal with just one fitness function, an energy landscape can be defined and the optima are easily found at the bottom of energy wells. **b** If more than one optimization target are at play, this landscape picture falls apart and we need to adopt a Pareto optimization approach. Then our task is to find a set of Pareto optimal solutions ($\Pi_\Gamma \subset \Gamma$) that minimizes all targets (here t_1 and t_2) simultaneously. These functions map each network $\gamma \in \Gamma$ into $\mathbb{R}^2$. The subset of Pareto optimal solutions is mapped into the Pareto front (thick gray curve). Along this curve it is not possible to improve both t_1 and t_2 at the same time.

ent values of λ, but we must also question the hypothesis of linearity introduced by equation (3.1). Therefore, we consider Pareto (or Multi Objective) Optimization [119, 62, 272]. We will analyze how a series of optimization problems map our networks into a target space (figure 3.1) and, taking into account the framework developed in chapter 2, we will illustrate how the rich phenomenology of phase transitions unfolds owing to the shape of the different Pareto fronts.

3.1.2. Multiobjective optimization of complex networks

Complex graphs allow us to define problems of increasing difficulty where criticality and first and second order transitions arise. Seminal work on network optimization addressed the problem from an SOO perspective [126, 111, 223, 204, 191], so some of our results can be put in context. Another advantage of complex networks is that good optimizers can be produced in the computer, simplifying the empirical work.

We propose three problems based on the conflict between the average path length between nodes and the density of edges, which roughly inform us about diffusion efficiency [132] (for which low average path length is desired) and implementation costs (less significant for sparser networks). Consider first the *topological* (or standard) average path length:

$$\langle l \rangle^t (\gamma) = \frac{1}{Z^t_{\langle l \rangle}} \sum_{i,j} \frac{d^t_{ij}(\gamma)}{2}, \tag{3.2}$$

where $d^t_{ij}(\gamma)$ denotes the distance (in number of edges) between nodes $n_i, n_j \in \gamma$ along the shortest path that connects them; and the topological (or standard) link density:

$$\rho^t(\gamma) = \frac{1}{Z^t_{\rho}} \sum_{i,j} \frac{a_{ij}(\gamma)}{2}, \tag{3.3}$$

where the adjacency matrix $A(\gamma) = \{a_{ij}(\gamma)\}$ presents $a_{ij}(\gamma) = 1$ if two nodes are linked in γ and $a_{ij}(\gamma) = 0$ otherwise. $Z^t_{\langle l \rangle}$ and Z^t_{ρ} are normalization constants discussed below.

The superindices in $\langle l \rangle^t (\gamma)$, $\rho^t(\gamma)$ remind us that we deal with the topological (or standard) average path length and link density, in which edges cost 1 unit. Geometric costs can be included if nodes are distributed, e.g., over a Euclidean space. Let $d^g_{ij}(\gamma)$ be the Euclidean length of the shortest path connecting n_i and n_j in network γ (i.e. the sum of the Euclidean lengths of the edges in the shortest path between these nodes). We introduce the geometric (or weighted) average path length:

$$\langle l \rangle^g (\gamma) = \frac{1}{Z^g_{\langle l \rangle}} \sum_{i,j} \frac{d^g_{ij}(\gamma)}{2}. \tag{3.4}$$

The shortest Euclidean distance possible between two nodes $l_{ij}(\gamma)$ only enters equation (3.4) if a direct link between n_i and n_j is present in γ (in that case $d^g_{ij}(\gamma) = l_{ij}(\gamma)$). This $l_{ij}(\gamma)$ allows us to introduce the geometric (or weighted) link density:

$$\rho^g(\gamma) \;=\; \frac{1}{Z^g_\rho}\sum_{i,j}\frac{a_{ij}(\gamma)l_{ij}(\gamma)}{2}. \tag{3.5}$$

Just as before, $\langle l\rangle^g(\gamma)$ and $\rho^g(\gamma)$ (note the superindexes indicating their geometric dependence) are normalized by $Z^g_{\langle l\rangle}$ and Z^g_ρ. A clique, or fully connected network (γ_C), has the shortest average path length possible always. As we will see later, this means that γ_C is Pareto optimal *always*, so we base our normalization on it: $Z^{t/g}_{\langle l\rangle} = \sum_{i,j} d^{t/g}_{ij}(\gamma_C)/2$, $Z^t_\rho = \sum_{i,j} a_{ij}(\gamma_C)/2$, and $Z^g_\rho = \sum_{i,j} a_{ij}(\gamma_C)l_{ij}(\gamma_C)/2$.

We combine $\langle l\rangle^{t/g}(\gamma)$ and $\rho^{t/g}(\gamma)$ as targets in different ways to generate three MOO problems:

(A) *Fully topological problem*, with $t_1 = \langle l\rangle^t(\gamma)$ and $t_2 = \rho^t(\gamma)$. Note that the geometry does not play any role in this case.

This version was originally studied in [111, 223] from an SOO perspective. From that approach, only the clique and star graphs appear relevant (as discussed in [223]) as the representatives of two phases at either side of a discontinuous phase transition. We show how this fits parsimoniously within the framework presented in previous chapters. But besides, we discuss now the whole Pareto front, its relevant shape, and some of its constituents. This front includes non-trivial complexities well differentiated from the star and clique, and it presents important connections to critical systems to which we come back in section 3.2.

(B) Partly geometrical problem, with $t_1 = \langle l\rangle^t(\gamma)$ (the same as in (A)) and $t_2 = \rho^g(\gamma)$. Geometry, through t_2, plays a relevant role now. Since the disposition of the nodes in space matters, we study this MOO in two different cases: i) nodes scattered randomly over the

$[0,1]\times[0,1]$ square in $\mathbb{R}^2$ and ii) nodes spaced evenly over a circle of radius 1.

In this problem we still use the topological average path length, meaning that we seek to minimize the number of hops or the number of relay stations between arbitrary pairs of nodes. To think about this problem we can picture an infrastructure such as a subway network whose contractor wishes to minimize the Euclidean length of line built while the users want to avoid transfers between lines.

(C) Fully geometrical problem, with $t_1 = \langle l \rangle^g (\gamma)$ and $t_2 = \rho^g(\gamma)$. In this case the geometrical cost is important for all targets involved. Again, the disposition of the nodes matters and again we study: i) nodes scattered randomly over the $[0,1]\times[0,1]$ square in $\mathbb{R}^2$ and ii) nodes spaced evenly over a circle of radius 1.

3.1.3. Outcome of network optimization

Three relevant topologies indicate major feats of all our Pareto fronts. The most prominent one is the clique: a fully connected network that presents the largest number of links possible, thus maximizes edge density (which is opposed to our desires), but it aways marks the minimum average path length. At least one global minimum of each target is always Pareto optimal and $t_1 = \langle l \rangle^{t/g}$ has the clique as its only global minimum, so it must be Pareto optimal always. It marks the top-left boundary of the Pareto front, as illustrated in Fig. 3.2**a**. This is true for all problems considered in this section 3.1.

The star presents a hub to which all other nodes are connected, while non-hubs are not connected to each other. There are N possible star graphs. If geometry is not considered, all of them are equivalent. When geometry intervenes and nodes are spaced over a circle all N stars are equivalent as well. All possible trees consist of as many edges as the star but, if geometry matters, only the minimum spanning tree (MST) minimizes *always* the edge density (t_2) – hence the MST is Pareto optimal

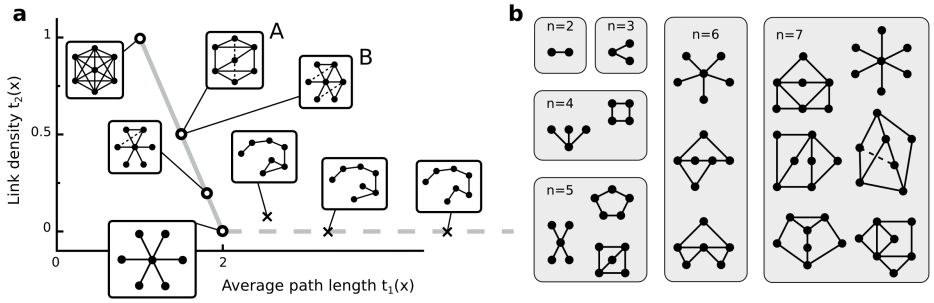


Figure 3.2: **Pareto front of the fully topological problem.** **a** The front (solid gray curve) is a straight line connecting two phases: a star and a clique. The slope of the line $d^c = -1$ determines that at $\lambda^c = 1/2$ a first order phase transition takes place. All networks laying on the front are global SOO optima at that critical value. Among them we find networks produced by attaching links to a star and others radically different from the star and from the clique (note the two graphs marked A and B: only one of them can be produced by attaching edges to the star). **b** All *core graphs* for have been listed for $N \leq 5$. Beyond that, it becomes increasingly difficult to count how many there are or even to tell apart two different ones.

whenever geometry is relevant and it always indicates the end of the Pareto front opposite to the clique (at its bottom-right).

Fully topological problem

This case has been studied as an SOO through equation (3.1) [111, 223]. That solution is incomplete from an MOO perspective which was not the chosen paradigm in those works anyway. This problem has the advantage that its front (Fig. 3.2**a**) can be found analytically and the phase transitions derived from it are independent of the number of nodes. We cannot guarantee the same for the variations studied later.

Because we normalized both targets using the clique as a reference, this network is mapped into $(1, 1)$ in the $t_1 - t_2$ plane. Any graph will

have less edges than a clique, thus the set Γ of all connected networks lays below $t_2 = 1$ in the target plane. The lower boundary of t_2 is achieved by connected networks with the minimum amount of edges possible $(N-1)$. There are a collection of such graphs, from the star to a linear chain – in between lay all possible trees. All of them have $t_2 = \rho^t = 1/N$, which tends to 0 as N goes to infinity. The average path length of these networks varies between that of the star $(2(N-1)/N \to 2)$ and the linear chain $((N^2-1)/3(N-1) \to +\infty)$. These minimally connected graphs lay on a horizontal stretch of the $t_1 - t_2$ plane (dashed line in Fig. 3.2**a**).

Among these trees (all with the same $t_2 = 1/N$), the star is the one with the lowest average path length, hence it is Pareto optimal. Any other network with a lower t_1 must have more links than the star, while the clique still sets the lower t_2 and upper t_1 bound. Thus the Pareto front must lay on a curve connecting the clique and the star – i.e. connecting $(1,1)$ and $(2,0)$ in the $t_1 - t_2$ plane.

We appreciate the following facts: i) The edge density is a function of the number of links alone and it does not depend on the topology of the network. ii) Given a network that is Pareto optimal, we generate new Pareto optimal networks by simply adding new connections. As an instance, the star is Pareto optimal and all its nodes are 1 edge apart from the hub and 2 edges apart from each other. Then, new edges can only be added that connect directly two non-hub nodes, turning a distance $d^t_{ij} = 2$ into $d^t_{ij} = 1$; but not affecting the network in any other respect. Put otherwise, once a network is Pareto optimal any addition of links has got only *local* effects in its average path length.

Adding new links to the star results in more Pareto optimal networks, the number of which grows combinatorially (that scaling saturates as we approach the clique). Take apart the $N-1$ non-hub nodes of a star: any network that we implement on this subset of nodes (connected or not), and which is subsequently embedded on the original star graph through the hub, is Pareto optimal. It is a *sufficient* (but not necessary) condition for a network to be Pareto optimal to contain a hub (Fig. 3.2**a**).

The *necessary* condition for Pareto optimality is that every node is at maximum 2 edges apart from each other. From any Pareto optimal

network (with or without a hub), adding new edges always generates new Pareto optimal graphs. Repeating this operation we always reach a clique, but this process is not reversible: Take the clique and delete connections randomly with the condition that your network remains Pareto optimal after every deletion – i.e. that every node is maximally 2 edges apart from each other. No rearrangement of the links is allowed. Let this process continue until we cannot remove any link without violating the Pareto-optimality condition. This algorithm might yield a star or any other graph from a collection of *irreducible* Pareto optimal networks, which we call *Pareto core graphs*. The star is the core graph with less edges possible. We can only construct these networks as described since defining regularities are not apparent – beyond the optimality condition that every node is at most 2 hops away from each other. Some of these graphs are represented in Fig. 3.2**b** for $N = 2, \ldots, 7$. The complexity scales from 1 core graph for $N = 2, 3$; to two core graphs for $N = 4$; to three for $N = 5$; to an unknown number for $N \geq 6$. For larger N it also becomes increasingly difficult to determine whether two core graphs are the same, given their invariance under the labeling of the nodes. Note that core graphs *are* Pareto optimal. They are representative of the staggering complexity contained in the front (which grows combinatorially) and they cannot be trivially composed as a mixture of stars and cliques. Because of this they constitute Pareto and global optima that have not been previously reported.

Even if we cannot list down all Pareto optimal networks, we can find where they live on the $t_1 - t_2$ plane. Adding one edge always modifies $\langle l \rangle^t (\gamma)$ by an amount $\Delta \langle l \rangle^t = -1/N(N-1)$, thus t_1 is decreased. The same operation increases t_2 by $\Delta\rho^t = 1/N(N-1)$. Because $\Delta\rho^t / \Delta \langle l \rangle = -1$ does not depend on the number of edges, Pareto optimal graphs thus generated lay on a straight line with slope $d^c \equiv \Delta\rho^t / \Delta \langle l \rangle$ (Fig. 3.2**a**). According to sections 2.2.2 and 2.3 of chapter 2, this front implies a first order phase transition with a critical point at $\lambda^c = -d^c/(1 - d^c) = 1/2$. The clique and the star are found at either phase (correspondingly for $\lambda > \lambda^c$ and $\lambda < \lambda^c$). Right at the critical value λ^c any Pareto optimal network is a global optimum. The plot of any order parameter as a function of λ just

presents a chasm between two constant values (not shown). At the critical point, the value of any order parameter is degenerated.

Partly geometrical problem

Figure 3.2**a** provides an archetype for the Pareto front that will be repeated (with variations) in the more elaborated MOO problems. Our fronts will present a first, stepped stretch that trades off between the clique (top-left) and some intermediate networks (usually the star); and a second, flat stretch with little variation in the vertical dimension (t_2) and a broad variation in the horizontal axis (t_1). In the previous case, this second stretch (dashed line in Fig. 3.2**a**) does not belong to the front, but it will in the following problems.

Nodes scattered over a plane Figure 3.3**a** shows the first example of this archetype. A very stepped stretch of the front trades off between the clique and the star just as before. However, this is a convex curve now, which we discuss below. The second archetypal stretch of the front trades off between the star and the MST, and is mapped onto an almost horizontal curve in the $t_1 - t_2$ plane with a slight convexity. This Pareto front ends up smoothly in its bottom-right extreme, so we dismiss any phenomenon associated to it. Because the whole front is convex, first order phase transitions are ruled out.

The first, stepped stretch of the front (Fig. 3.3**a**, inset) presents a feature that appears in most subsequent cases. This stretch is a convex curve that ends abruptly (with the notion of *abruptness* introduced in chapter 2, section 2.2.1). This indicates that a second order phase transition takes place. This transition trades off between the clique (persistent global optimum for $\lambda > \lambda^* \simeq 0.61$) and dense but incomplete graphs reached as we move below λ^*. Because the clique is optimal for $\lambda \geq \lambda^*$, anything that we measure on this global optimum stays constant as a function of λ until $\lambda < \lambda^*$, for which our wandering over the Pareto front yields a changing global optimum as λ decreases. Then, any measurement performed on the SOO optimum will vary steadily with a derivative (with respect

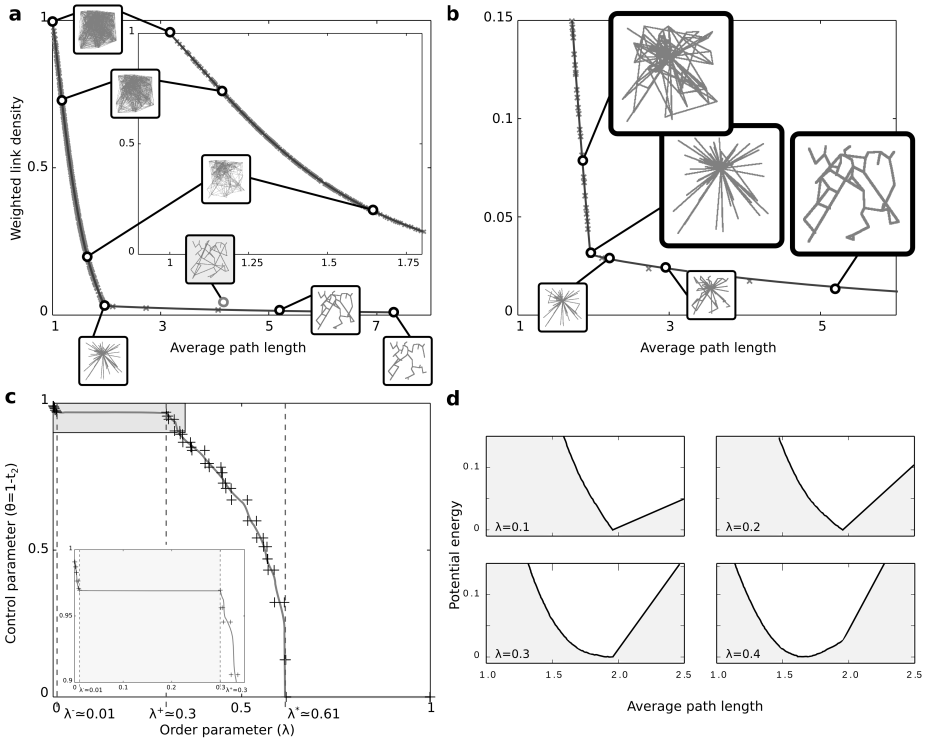


Figure 3.3: **Partly geometrical problem on nodes scattered over a plane.** **a** The front follows the archetype of the topological problem with two roughly perpendicular stretches that trade off between the clique (top-left), the star, and the MST (bottom-right). Incomplete cliques are reached after a second order phase transition because the Pareto front ends abruptly in its top-left (inset). The other extreme of the front ends smoothly. **b** A sharp edge indicates a second order phase transition with the star graph being optimal for a range $\lambda \in (\lambda^-, \lambda^+)$. **c** Plotting an order parameter as a function of λ reveals both transitions at $\lambda \simeq 0.61$ and at $\lambda^- \simeq 0.01$ and $\lambda^+ \simeq 0.3$ (inset). The star is optimal in the range $\lambda \in (\lambda^-, \lambda^+)$, thus any order parameter is constant in that range. **d** The non-analyticity of the Pareto front is inherited by the energetic landscape also as a sharp edge. The SOO is vividly illustrated thanks to this potential landscape, whose minimum is occupied by one same network for several values of λ.

Figure 3.3: As lambda changes, the potential well is deformed until the minimum drifts away from the sharp edge.

to λ) different from 0. This discontinuity in the derivative indicates that a second order phase transition takes place (Fig. 3.3**c**). As discussed in the previous chapter, this kind of transition might have associated critical points. To asses that, we would need to evaluate a diverging susceptibility. We only have a numerical approximation to that susceptibility, so every comment about the criticality of this phase transition would be more speculative than desired. We think that this is question worth researching: do nearly completed cliques present some kind of universal critical behavior? Which one is it?

A sharp edge in the front (Fig. 3.3**b**) indicates yet another second order phase transition similar to that described in Fig. 2.3**c**: The slope of the front is well defined as we tend towards the sharp edge from the left (yielding a slope $d^- \simeq -0.43$ such that $\lambda^- = -d^-/(1-d^-) \simeq 0.3$), and as we tend to the sharp edge from the right (now with $d^+ \simeq 0.01$ such that $\lambda^+ = 0.01$). (Following the notation introduced in chapter 2, $d^- < d^+$ with both $d^-, d^+ < 0$ and $\lambda^- > \lambda^+$.) For any $\lambda \in (\lambda^+, \lambda^-)$ SOOs are well defined through equation (3.1), but there are not any points of the front with a slope $-\lambda/(1-\lambda) = d \in (d^-, d^+)$ where to locate the optimum. Instead, the same one network laying precisely at the sharp edge is consistently optimal for this range of λ. Anything that we measure about this optimum will remain constant as a function of λ in (λ^+, λ^-), but samplings of the front run smoothly below λ^+ and above λ^-, with well defined derivatives for any order parameter as a function of λ. Hence, two discontinuities are evident in this derivative (Fig. 3.3**c**, inset).

Qualitatively, browsing the front through λ is a continuous transition from the clique (which is a global optimum for a wide range of λ), through the star graph (also a persistent optimum for a continuum of λ), to the MST. The weighted density of edges ($t_1 = \rho^g$) penalizes large connections first, which are dropped as we leave the clique. But enough of them survive among Pareto optimal graphs so that the star can be reached continuously, without needing a drastic rewiring that would leave an imprint

in the order parameters. Note that these few long edges survive because they enable a low average path length ($t_2 = \langle l \rangle^t$), which is still measured as the number of hops between nodes. Finally, to rearrange the star into the MST, the surviving long edges are replaced by winding branches that extend visiting many nodes on their way. Alternative strategies, like hybrid MSTs that incorporate non-essential shortcuts between far-apart nodes, fall off the Pareto front (note the sub-optimal graph shaded in gray in figure 3.3**a**).

It is very useful to consider the potential landscape introduced by equation (3.1) to stretch our intuition. For a fixed value of λ we compute the energy $\Omega(\gamma, \lambda)$ for every Pareto optimal network. The result is a lower energy boundary (Fig. 3.3**d**). Non Pareto-optimal networks must present yet higher energies. The SOO solution becomes now very intuitive since the global optimum lays at the minimum of $\Omega(\gamma, \lambda)$, which has got a vivid graphic representation. But this potential landscape changes as a function of λ, and consequently its minimum. The sharp edge of the Pareto front is inherited by the potential landscape, which presents a persistent minimum for a range of the control parameter.

Nodes spaced over a circle The front of this problem (Fig. 3.4**a**-**b**) again follows the archetype: i) a stepped stretch to the left that trades-off between the clique and dense (yet incomplete) graphs and ii) a flatter stretch that encompasses graphs with roughly the same edge density but a wide variation along the average path length dimension. This second stretch extends to large values of t_1: a region populated by minimal, circlelike networks with little long-range connections. It seems a convex stretch that ends smoothly, so nothing remarkable happens there. Again, the stepped stretch of the front is convex (Fig. 3.4**a**, inset) and ends abruptly revealing a second order phase transition. It is similar to the one encountered before by the clique and happens at a similar value $\lambda \simeq 0.59$ (Fig. 3.4**c**).

The notable feature that this front introduces is a concavity at the junction between the two archetypal stretches of the front (Fig. 3.4**b**). The cavity lays at the confluence between the three relevant network topolo-

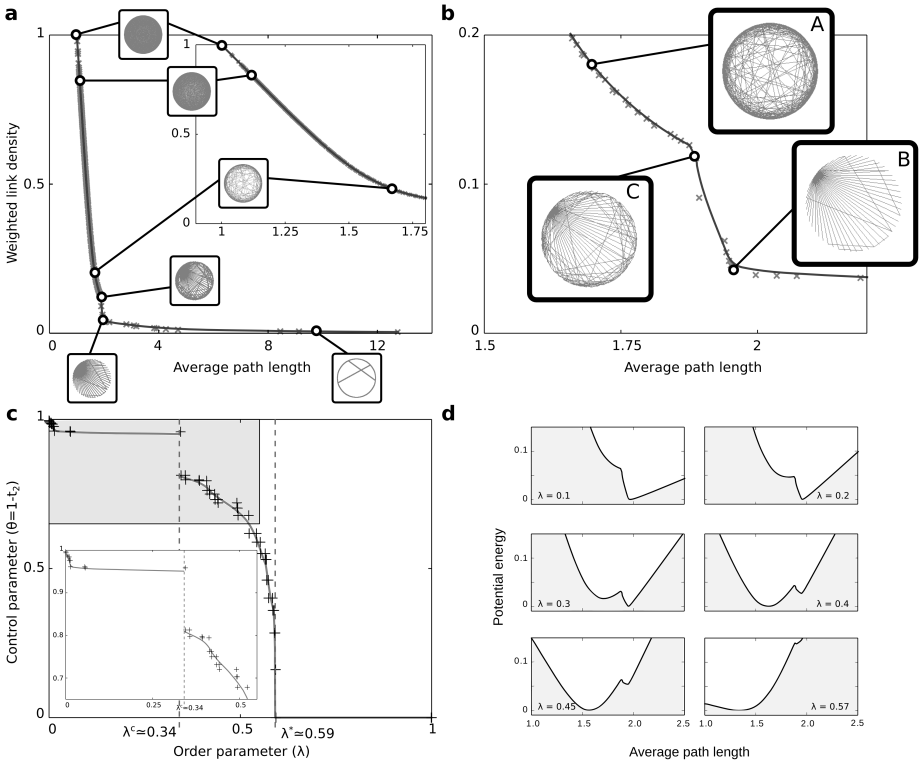


Figure 3.4: **Partly geometrical problem on a circle.** **a** Again, the front follows the archetype of the topological problem with two roughly perpendicular stretches that trade off between the clique (top-left), star networks, and the MST. Around the clique it is observed the same phase transition as before. In the cavity it is solved a complex rearrangement. Networks that drop their larger connections first (A) must morph into a star (B), which requires some of these far-reaching edges. Therefore, some Pareto optimal networks are produced that never get to be SOO optima (C), yielding a first order transition. **c** Order parameters as a function of λ reveal the first ($\lambda^c \simeq 0.34$, magnified in the inset) and second ($\lambda^* \simeq 0.59$) order transitions. **d** The landscape potential unveils the mechanisms for local equilibrium and hysteresis associated to first order transitions. At low levels of the control parameter ($\lambda \sim 0.1$) only one minimum exists in the global energy Ω.

Figure 3.4: A pocket becomes locally stable for $\lambda \sim 0.2$. This grows for larger λ, until it becomes the global extreme of the energetic landscape ($\lambda \sim 0.4$). Optimizing our networks through numerical algorithms can get us stuck in local minimums, so to transit from one potential well to the other we need to increase our control parameter until one of the wells get destabilized ($\lambda \sim 0.57$). Repeating the operation with decreasing λ can get us stuck in the other well, thus engaging in a hysteresis loop. Some Pareto optimal networks inside the cavity are reached at these metastable states.

gies: incomplete cliques, the star, and encircled nets (the MST of the problem). As before, longer edges are dropped first. In that previous example the scattered nodes managed to retain enough long-range links, thus enabling a continuous transition through the star. This is not possible now, and the symmetry of the circle might be crucial therefore. Earlier, the distribution of lengths were varied, while now all long-range edges are the same: the moment one is dropped, the others follow. As we leave the clique, we converge quickly to encircled graphs with little long range connections; and these lay inside a cavity of the front (Fig. 3.4**b**). To reconstruct a star (which remains Pareto optimal due to its low topological average path length) a drastic rewiring is unavoidable. This prompts a first order phase transition (at $\lambda^c \simeq 0.34$) whose imprint is, indeed, that cavity. That transition is reflected in any order parameter θ that we plot as a function of λ (Fig. 3.4**c**, inset).

We can resort again to a potential landscape to visualize this. Plotting $\Omega(\gamma, \lambda)$ for all Pareto optimal networks we obtain the lower energy boundaries portrayed in Fig. 3.4**d**. This landscape changes as λ varies, producing two potential wells associated to local minimums. At $\lambda = \lambda^c$, both minimums present the same energy, thus both phases coexist. Moving away from the transition point, one of the wells is unstabilized. Note that moving λ back and forth could get us temporarily trapped in metastable states (the most energetic local minimum) and hysteresis loops would be observed.

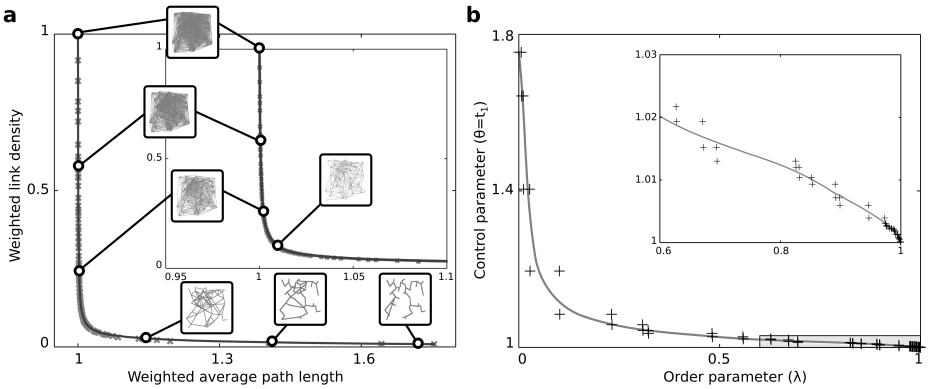


Figure 3.5: **Fully geometrical problem for nodes scattered over a plane.** **a** The front has no accidents. It is completely convex and spans all possible slopes so that each $\lambda \in (0, 1)$ poses an SOO with a different solution. As we roll over the front, the clique gently leads to less connected networks, towards the MST. **b** The absence of phase transitions renders smooth plots of any order parameters.

Fully geometric problem

Introducing geometry in both target functions has the effect of smoothing the Pareto front, removing the first order phase transition. Some relevant second order transitions disappear. Others persist, but only at the extremes of the front. The picture becomes closer to a soft trading-off between the clique and the MST.

Nodes scattered over a plane This problem presents a quite uninteresting front (Fig. 3.5**a**) without phase transitions. The front spans all possible slopes $d \in (-\infty, 0)$, so that SOOs with different solutions can be posed for each $\lambda \in (0, 1)$. Any order parameter renders a continuous plot (Fig. 3.5**b**) even when zooming in into tiny details (inset). The first derivative of the order parameters also behaves properly.

All phase transitions from previous cases have vanished. Besides the

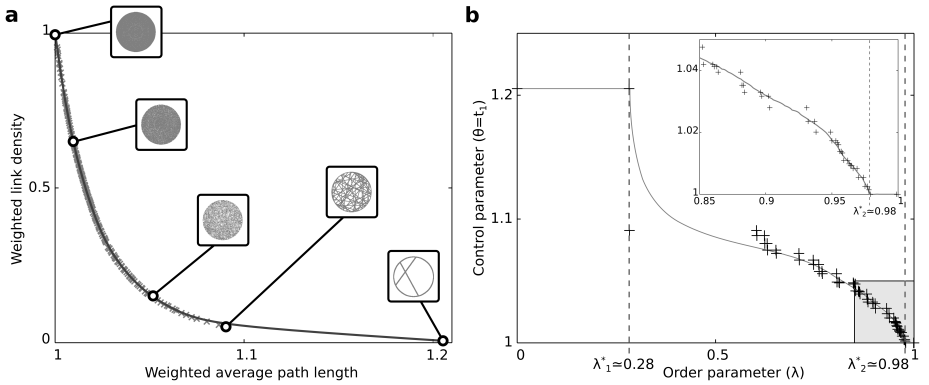


Figure 3.6: **Fully geometrical problem on a circle.** **a** The front presents a smooth transition between the clique and the open circle, with no relevant feats except in the extremes of the front. These end up abruptly, as in second order phase transitions. **b** Plotting any order parameters reveals these phase transitions at $\lambda_1^* \simeq 0.28$ and $\lambda_2^* \simeq 0.98$ (inset).

geometric disposition of the nodes (which has some obvious influence over what transitions are present), it is notable that choices of optimization targets exist for which previously existing transitions disappear. Given a same set of networks $\gamma \in \Gamma$, a choice of targets erases previous transitions. This implies that an optimal yet gentle evolution between radically different topologies (clique and MST) can happen, despite the drastic modifications that we might envision necessary *a priori*, and despite the phase transitions that do take place on the same graphs for other choices of targets. This stresses the role of target functions to frame phase transitions properly.

Nodes spaced over a circle Again, introducing geometry in both targets smooths the Pareto front (Fig. 3.6**a**). The first order transition found in the circle before has disappeared. There is no cavity now and the tradeoff between clique and circle happens gradually as we roll over the front. That previous transition took place because the gradual drift from clique to cir-

cle was interrupted by the presence of the Pareto optimal star, which kept the average path length low because it was measured as the number of hops between nodes. But now geometry also enters through $t_1 = \langle l \rangle^g (\gamma)$ and using only two links to get from one node to another is still costly if these are far reaching connections. It is more economic now to circle around even if that implies visiting many more nodes. Thus the star is retracted from the Pareto front and the transition from the clique to the MST proceeds smoothly.

The bottom stretch of the front seems convex but abruptly terminated, suggesting a second order transition at $\lambda_1^* \simeq 0.28$ that trades off between the MST (an almost complete circle, which is persistently optimal for $\lambda \leq \lambda_1^*$) and other, more connected graphs. The characteristic plot of a second order transition is noted in any order parameter (Fig. 3.6**b**). As before, the possibility that this transition presents a critical point remains open. At the other end of the front we find the usual transition associated to the clique, which did not disappear but has been moved to $\lambda_2^* \simeq 0.98$. The same characteristic order parameter plot can be appreciated (Fig. 3.6**b**, inset). For $\lambda \in (\lambda_1^*, \lambda_2^*)$ any order parameter is a smooth function of λ.

3.1.4. Discussion of Pareto optimal networks

We have solved three MOOs defined on complex networks. These problems allow us to explore interesting aspects of Pareto optimality. Following recent contributions in biology [79, 12], our work is an exploration of a morphospace. A first approach to such spaces is to list all possible morphologies for a system and locate them quantitatively with respect to some relevant aspects – here, complex networks are characterized in an average-path-length vs. edge-density two-dimensional space. We propose that a natural selection process based on Pareto optimality shall constrain further such a morphospace and we study the effects of these conflictive restrictions. In doing so, we continue recent efforts [293, 270, 292, 320] to illustrate how Pareto optimality reduces the effective dimensionality of certain complex systems.

Alternatively, a recent theoretical framework brings together statisti-

cal mechanics and MOO [282, 285], thus enriching the analysis of Pareto optimal designs. This framework reveals universal features of Pareto optimal systems that correspond to phase transitions or critical phenomena.

First order phase transitions indicate that a system must undergo important structural changes despite little variation of some control parameter. In thermodynamics, this implies great investments of energy in reshaping matter, e.g., as it transits from solid to fluid. Similar demands might be requested of complex Pareto optimal systems, specially if the parameters controlling the conditions for optimality may change over time. This underscores the importance of gathering knowledge about the Pareto front before implementing solutions to any optimization problem.

Second order phase transitions can also be very informative about the nature of a system. They indicate that some solution is stable for a large range of the control parameters, thus making it more likely if evolution or design has taken place under many different scenarios. On the other hand, if such stable solutions would not show up in an evolutionary setup, we would have strong evidence that a large set of possible circumstances do not occur naturally.

We have found a variety of first and second order transitions. These depend very much on the precise mathematical expression of the optimization targets (see, for example, how transitions disappear as we change the measure of average path length). Hence, looking at the problem from an alternative perspective, the presence or absence of expected phase transitions in real systems could be informative about the nature of the optimization pressures that these systems might be subjected to.

Finally, the fully topological problem presents a quite singular case: a first order phase transition takes place between the star graph (for $\lambda < \lambda^c$) and the clique ($\lambda > \lambda^c$), while all other Pareto optimal networks are also global optima at $\lambda = \lambda^c$. This would suggest that cliques and star graphs should happen overwhelmingly more often than any other topology when geometry is not relevant, unless every optimal network had evolved under the unlikely condition $\lambda = \lambda^c$. This is notably at odds with the reality and a possible solution (with connections to critical phenomena, [284]) will be explored elsewhere in the next section.

3.2. Systems poised to criticality through Pareto selective forces

In section 2.3 of the previous chapter we introduced criticality through the Pareto formalism. We also mentioned how power laws (often associated to critical phenomena) are quite abundant among complex systems [351, 11, 192, 27, 189, 139, 35], despite the need to fine tune parameters to reach criticality in physical systems. This led to hypotheses of how systems might poise themselves to critical points and to a rich literature on how power-laws might emerge, with or without criticality being involved [14, 15, 175, 17, 45, 46, 92, 214, 222, 215, 146]. These ideas have been connected to optimally evolved structures, but a general framework is largely missing. Indeed some authors have actively downplayed the importance of this connection between optimal designs and criticality [45, 46].

The Pareto front unites optimal designs to a framework in which criticality is defined in a robust and natural way, just as an expansion of the thermodynamic concept. Besides, we proceed to show now that certain Pareto optimal systems *must* look critical when described through statistical mechanics ensembles – hence implying that *Pareto selective forces* are yet another mechanism capable of driving some systems towards a critical state. We term Pareto selective forces to any procedure (natural, artificial, algorithmic, or otherwise) that, applied upon a set of designs (species, candidate solutions to an optimization problem, distributions of goods in a market), makes them evolve towards some underlying Pareto front. Genetic algorithms designed to locate the Pareto front (like those from chapter 1, section 1.3.2) are examples of Pareto selective forces that were *artificially designed*. In section 3.2.2 we discuss the existence of these forces in more natural contexts.

3.2.1. Descriptions of Pareto optimal sets based on statistical ensembles

If we come across a naturally occurring set of objects it is often useful to study them through *Maximum Entropy models* (MaxEnt) models. Take some measurements ($\{f_k\},\ k = 1,\ldots,K$) performed on the set that we wish to describe, and that might be found in any of $\sigma_j \in \Sigma$ *microstates* – potentially including configurations not observed in the empirical set. Among all possible probability distributions $\{P_i(\boldsymbol{\sigma}_j)\}$, the one that most faithfully describes our observations $\{f_k\}$ is the one with largest entropy and takes the form [166, 167]:

$$P\left(\sigma_j,\{\lambda_k\}\right) \;=\; \frac{1}{Z}\exp\left(-\sum_k \lambda_k f_k\left(\sigma_j\right)\right), \tag{3.6}$$

where λ_k are inferred from the data and $Z = \sum_j \exp(-\sum_k \lambda_k f_k(\sigma_j))$. MaxEnt has been successfully applied to diverse complex systems: distributions of species across ecological niches [142], letters within words [310], antibody coding regions [216], pixels in natural images [309], or spiking neurons [326, 325, 328]; and we will make used of them in future chapters too. From equation 3.6 it is possible to evaluate susceptibilities $\chi_{kl} = \partial\left\langle f_k\right\rangle_P/\partial\lambda_l$, where $\left\langle f_k\right\rangle_P$ is the average value of the observable f_k given the model $P\left(\sigma_j,\{\lambda_k\}\right)$. Note that these susceptibilities are similar to the ones introduced in chapter 2, section 2.3. Previously we had optimal systems whose order parameters θ changed as a function of λ (and this change was given by the corresponding $\chi_{\theta\lambda}$), now we have an optimal model (the best one possible to describe our data given by equation 3.6 and parameterized by λ_k) and observables f_k of that model would change according to χ_{kl} if we would vary the model (by changing the λ_k) slightly. The MaxEnt model from equation 3.6 would also minimize the free energy $F = \Omega - TS$, which entails the minimization of Ω – hence such models are connected in multiple ways to the framework introduced in chapter 2.

Once again, we identify critical systems through diverging susceptibilities. From chapter 2 we know how these relate to the geometry of the

Pareto front. Among others, we know that the case in which the whole front is a straight line corresponds to a critical point. Because Pareto selective forces evolve ensembles of designs towards their front, any system with a straight Pareto front that has been subjected to such forces long enough must seem poised to a critical configuration when described from a statistical mechanics perspective. The criticality of such a description should be robust for these systems, while Pareto optimal ensembles with other kinds of fronts should generally not look poised to any kind of relevant model.

We illustrate this point with three of the systems studied in section 3.1.3, whose Pareto fronts are schematically reproduced in figure 3.7 for convenience. The most relevant case that will be analyzed, and that we label **A** now and on, entails the minimization of both topological path length and edge density (the fully topological problem from section 3.1.3, in which geometric space is irrelevant). In this case the Pareto front (Π_A) is a straight line with slope $\delta = -1$ (Fig. 3.7**a**), which implies a first order phase transition with a critical point at $\lambda = \lambda_A^c \equiv -\delta/(1-\delta) = 1/2$. The other two cases (labeled **B** and **C**), are those in which each link still contributes the same to the average path length, but edge density is weighted by the Euclidean distance (i.e. the partly geometrical problem from section 3.1.3). In **B** nodes are scattered randomly over $[0,1]\times[0,1] \in \mathbb{R}^2$ and in **C** nodes are regularly spaced over a circle. The fronts (Π_B and Π_C) associated to these problems are not full straight lines, hence the Pareto optimal sets should not seem related to any singular point when studied though MaxEnt models.

The networks used to reconstruct the fronts $\Pi_{A,B,C}$ have been subjected to Pareto selective forces: a genetic algorithm was designed precisely for that. Π_A can be computed analytically, but we used the same genetic algorithm anyway so that we can compare empirical data produced in similar ways. For each one of the Pareto optimal ensembles we use the global energy defined by equation 3.1 to derive MaxEnt models:

$$P^{\lambda}(\gamma;\alpha) = \frac{1}{Z}\exp(-\alpha\Omega(\gamma,\lambda)), \tag{3.7}$$

with λ and α the only free parameters of the model. α represents an arbi-

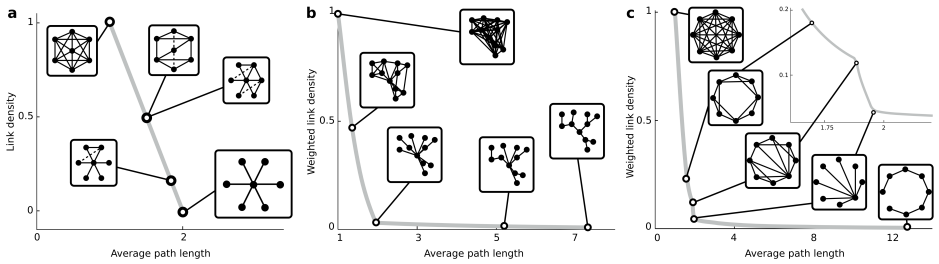


Figure 3.7: **Simultaneous minimization of average path length and edge density. a** Purely topological graphs lie on a linear Π_Γ. In **b**, **c** the Euclidean distance weights the cost of each link when computing the density of edges. Graph drawings are qualitative, units refer to the systems in [285]. **b** Nodes distributed over a plane. Optima trade between a clique, a star graph, and the Minimum Spaning Tree through two second order transitions. **c** Nodes placed over a circle display a first and a second order transition.

trary scaling factor that will be handy to test the robustness of the criticality of our systems. (Note that α would also correspond to the Lagrange multiplier of the entropy – i.e. a temperature – if we would include it in the model.) Equation 3.7 is a model of how often each network should be present in the Pareto optimal set if networks were to minimize $\Omega(\gamma, \lambda)$.

We can use this model to compute p-values

$$p^{\alpha}_{A,B,C}(\lambda) \quad = \quad \prod_{\gamma \in \Pi_{A,B,C}} \exp(-\alpha \Omega(\gamma, \lambda)) \tag{3.8}$$

that should be larger the better the considered model. With these p-values we can infer the free parameter λ that better matches the predictions of the model to the statistics from the simulations:

$$\tilde{\lambda}_{A,B,C}(\alpha) \quad = \quad \max_{\lambda} \left\{ p^{\alpha}_{A,B,C}(\lambda) \right\}. \tag{3.9}$$

Put otherwise, this tells us the most likely model given the data. Figure 3.8**a-c** shows p-values $p^{\alpha}_{A,B,C}(\lambda)$ for different α. p-values for constant

α have been normalized to ease their comparison. Only $\tilde{\lambda}_A$ remains unchanged at $\tilde{\lambda}_A = \lambda_A^c$, the critical value of the system, suggesting that it is poised to that singular statistical description. $\tilde{\lambda}_{B,C}$ change depending on α and do not correspond to relevant parameters in the phase space.

Alternatively we use the Kullback-Leibler divergence to measure the information that we lose when we use equation 3.7 to describe the Pareto optimal sets from the numerical experiments:

$$D^{KL}(P^e_{A,B,C}||P^\lambda(\alpha)) = \sum_{\gamma \in \Pi_{A,B,C}} P^e(\gamma) log\left(\frac{P^e(\gamma)}{P^\lambda(\gamma;\alpha)}\right), \quad (3.10)$$

where $P^e_{A,B,C}(\gamma) = 1/||\Pi_{A,B,C}||$ is the empirical frequency of each network in the front. D^{KL} (figure 3.8**d**) is minimal but not zero at $\tilde{\lambda}_{B,C}(\alpha)$, while it vanishes at $\tilde{\lambda}_A(\alpha) = \lambda_A^c$. Again $\min_\lambda\{D^{KL}_{B,C}\}$ (but not $\min_\lambda\{D^{KL}_A\}$) depends on α (not shown).

It turns out that a MaxEnt model tuned to the critical point is a good description of empirically obtained Pareto optimal sets in case **A**, but not in cases **B** nor **C**. Put otherwise, from the point of view of the favorite statistical mechanics model, the Pareto selective forces that have driven the network ensemble in case **A**, brought them towards a state that we know is critical. As stated above, whenever a Pareto front is a straight line, the Pareto optimal set will appear tuned to a critical point. We have found this critical behavior for Pareto optimal graphs and for a model of language evolution [286, 111, 251, 265, 301] that will be studied with more detail in the next chapter.

3.2.2. Pareto selective forces in real life

Do Pareto selective forces (like the ones simulated in [285, 286] and for this thesis) exist in nature?

Note first that *constrained optimization* can reconstruct a Pareto front: Find the networks $\gamma^1 \in O_1$, $\gamma^2 \in O_2$, $\gamma^3 \in O_3$, ... (with $O_i \subset \Gamma$) with a fixed number of nodes (N), constrained to have precisely $L_1, L_2, L_3, \ldots$ edges respectively (being $L_i \in [N-1, N(N-1)/2]$ randomly distributed

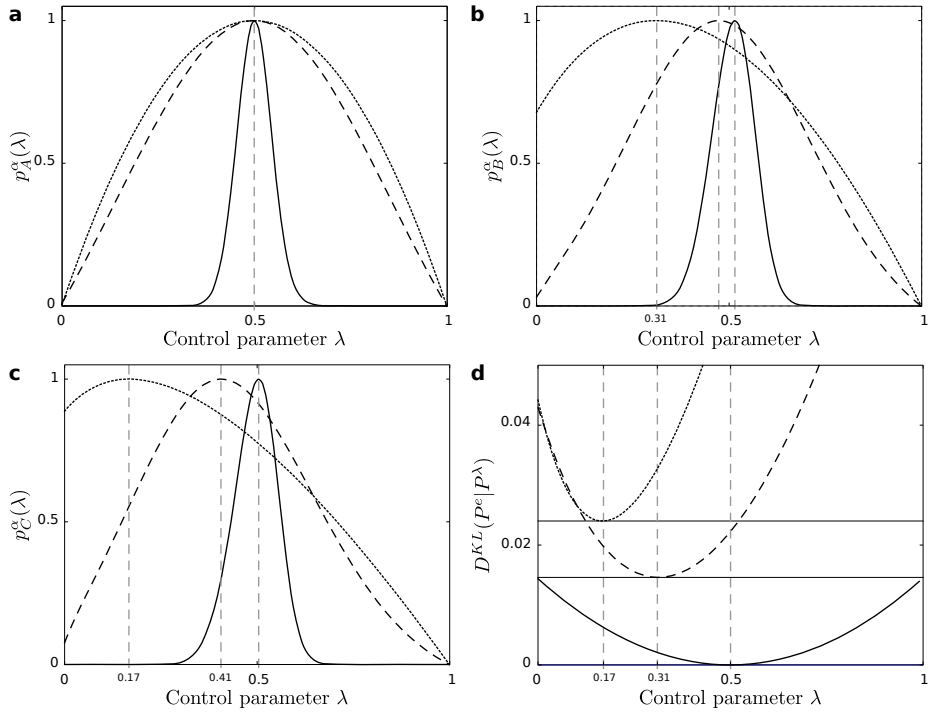


Figure 3.8: **Testing MaxEnt models.** **a-c** p-values for MaxEnt models of Pareto optimal sets $\Pi_{A,B,C}$ as a function of λ and α ($\alpha = 0.0125$, solid; $\alpha = 0.1$, dashed; and $\alpha = 1$ dotted lines; curves have been normalized for comparison). **a** α does not affect the critical ($\tilde{\lambda}_A = \lambda_A^c$) description of Π_A. **b**, **c** Changing α changes the best model, so that an α-invariant, consistent description does not arise for these Pareto optima. $\tilde{\lambda}_{B,C}(\alpha)$ usually do not correspond to relevant parameters in phase space. **d** The best model misses the least information about each data set (Π_A, solid; Π_B, dashed; and Π_C dotted lines; $\alpha = 1$). This loss is vanishingly small in the critical case.

integers), and with the lowest average path length possible given each L_i. The set $O_1 \cup O_2 \cup O_3 \cup \ldots$ samples from the Pareto front in Fig. 3.7**a** and will appear critical if described through equation 3.7. *Highly Optimized Tolerance* (*HOT*) states [45, 46] have been proposed as an alternative to *Self- Organized Criticality* [14, 15, 17] (SOC) or *edge of chaos* dynamics [175] to explain power-laws in complex systems without resorting to critical states. HOT generates power-laws through thoughtful design that optimizes several constraints simultaneously. This ability to generate power-laws through design was used to argue against SOC and to dismiss the importance of critical configurations. One of the strategies used in the design of HOT states is, precisely, constrained optimization. Disregarding the way to achieve HOT designs, since they solve a PO problem, if the corresponding front is flat these designs will appear critical when studied through statistical mechanics – which is the appropriate, theoretically sound way to asses criticality. We propose that PO might help close the gap between SOC and HOT and put under the same light those seemingly confronted approaches.

Recent works [227, 293, 270, 148, 320, 235] (some of them reviewed in section 1.3.4) account for a series of dimension reduction and allometric scalings in real biological data using PO. These data appear distributed over a front, suggesting that Pareto selective forces might be operating. Numerical evidence also indicates that ecological populations will evolve towards a Pareto front through prey-predator dynamics when different predators select preys with different criteria [188, 138]. If some of these biological systems would belong in a straight Pareto front, that system would automatically look critical from a statistical mechanics perspective.

Finally, PO plays a relevant role in economy, as seen in chapter 1. The first fundamental theorem of economic welfare guarantees that *any competitive market is Pareto efficient at equilibrium* [8, 9]. If an equilibrium competitive market belongs in a straight Pareto front, it must hence appear critical when studied through MaxEnt methods. The conditions for a *competitive market* are stringent and relate to (often incomplete) available information. However, such markets are an interesting reference of

academic importance [34, 202].

3.3. Reviewing some selected literature with Pareto optimal designs

To the very best of our knowledge, references to phase transitions and critical points are absent in the literature about Pareto optimality. We believe that the contributions introduced so far in this PhD thesis offer a novel, interesting perspective on PO. Behind every Pareto front there are potential phase transitions and critical points. The later might be more difficult to identify unless the front presents clear straight stretches. Pareto optimization has been used to analyze diverse systems including complex biological processes or models of phenomena out of thermodynamic equilibrium. A tool to locate relevant phase transitions or critical points in these cases should be taken seriously.

We analyze now three recent contributions from the literature in which our perspective might add interesting knowledge. The first one deals precisely with models of dynamical processes, specifically: the relay of information through complex networks. The second example studies optimal networks whose Pareto front suggest that a critical point must exist. The last example illustrates how, in a strictly physical setup, previously existing methods in the MOO literature suggest designs that cannot be physically stable – and our contributions reveal the reasons of this instability in a clear manner.

3.3.1. Networks for efficient communication

In chapter 1, section 1.3.1 we introduced a tradeoff between the efficiency of diffusing and routing messages across network topologies, a problem studied by Goñi et al. [132, 12]. They located Pareto optimal networks in target space depending on whether they maximized and/or minimized these efficiencies. Additionally, there are some further interesting features (phase transitions) that we can extract from those Pareto

fronts.

Let us look at those transitions that we will find from an information theoretical mindset. Therefore, consider for a moment what diffusion and routing entails. The former is a process that happens spontaneously in every physical system. If diffusion is efficient in performing some task, we might get some work done for free. The downside is that it is a maximally random process. In the case of message passing, diffusion is alright as long as you do not care that the message reaches everybody on average. Because your message is randomly broadcast, you might need several copies of it to be sure that it reaches your target. This is, in a way, a leak of energy that should be taken into account. On the other hand, routing requires knowledge of the global topology of the network. This information is reflected in costly bits. Routing also requires that nodes invest energy *against* diffusive processes: think of the i-th node directing a bit along one out of k_i possible alternatives given by its out-degree. While diffusion maximizes entropy $h_i \sim log(k_i)$, the node must perform a work to ensure that only efficient paths are taken.

The work described above entails the modeling of important processes out of equilibrium. Take structural information being articulated through message passing between units – e.g. as in ontogeny or in inference carried out by ants [133, 118] or by fluid neural neurons [300, 297, 86]. Then, the phase transitions inherent to the Pareto fronts found in [132] may be manifested in sound, physical substrates and they might, perhaps, be accessible for experimental testing under laboratory conditions.

Each of the fronts in [132] (figure 1.8**b**) results in a series of transitions. We focus on the cases that we find more relevant. In front 1 nodes attempt to make work *against* the diffusive inertia. For them, it is wished that the routing efficiency is maximized while that of diffusion is minimized. In figure 3.9**a** we reconstructed front 1 manually, smoothed it (following the same procedure as for our networks in section 3.1, see appendix A.1), and normalized it. This renders two first order transitions, as revealed by two cavities (one of them barely noticeable at the bottom left) and the gaps in order parameters (figure 3.9**b**). An energy landscape tells us how accessible different topologies are under a given circumstance

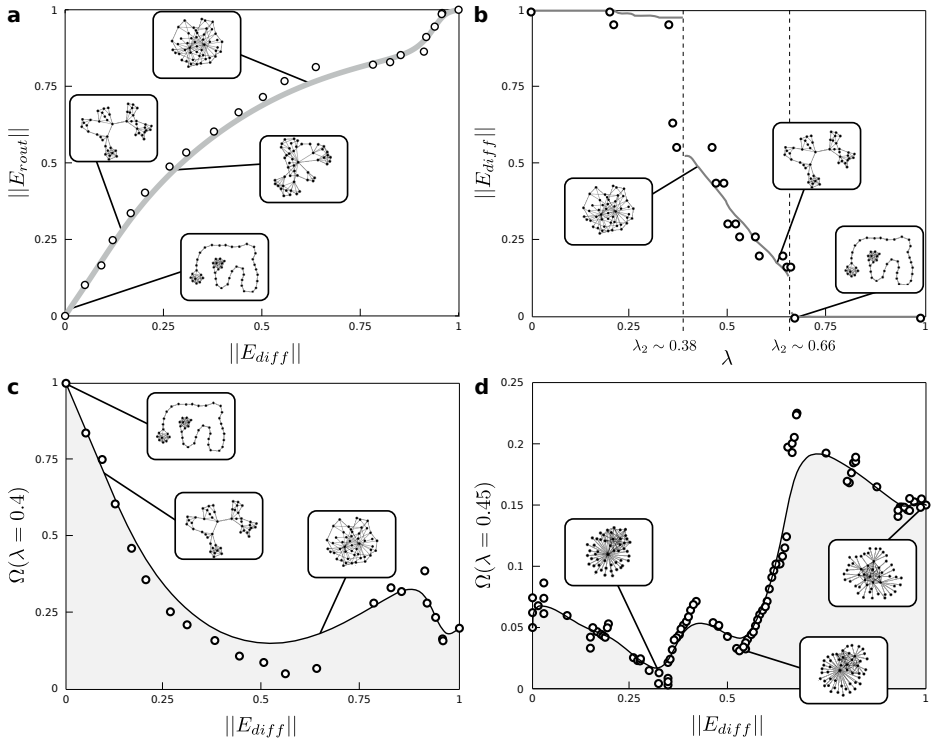


Figure 3.9: **Pareto optimal networks for communication.** **a** Pareto front extracted from [132] (reconstructed manually from front 1 in figure 1.8**b**) that entails the optimization of graphs with respect to: i) minimizing diffusion efficiency and ii) maximizing routing efficiency. A large cavity in the front and a smaller one (slightly noticeable at the bottom left) reveals two first order phase transitions that are noted in any order parameter **b**. These transitions result in important rewirings of the graphs to switch between distinct topologies: from almost linear chains to tree-like structures containing clusters in its branches, to networks seemingly compacted around a unique cluster. (Network drawings corresponds to actual topologies found in [132]. Their location along the front and at the order parameter plot are qualitatively correct, but not precise.) **c** An energy landscape illustrates how networks can become local minimums and how phases can coexist at a given λ (λ = 0.4). **d** Similar analysis are possible for the other fronts in figure 1.8**b**.

Figure 3.9: Here, the energy landscape generated for front 2 with $\lambda = 0.45$ reveals a structure with more local minimums than that of front 1. (Note that the energy landscapes in **c** and **d** should delimit the least energetic networks possible, but we find several dots below this lower bound. These should correspond to unfeasible solutions. This is a numerical artifact that stems from the smoothing used to build the energy landscape.)

(codified by a value of λ).

Thinking again about structural information that uses messages to articulate itself, consider a situation in which network arrangements that successfully convey messages get reinforced. This reminds us of some connections between Darwinian theory, Hebbian learning, and cortical organization [100]. Such a setup shall tend to build up structures that take advantage of diffusion and routing efficiency simultaneously. The most efficient tradeoff resulting of maximizing both quantities is captured by front 2 in figure 1.8**b**. This front appears more jagged, resulting in several first order transitions (not shown) and local minimums in the corresponding energy landscapes (figure 3.9**d**) which shall difficult the ability of these systems to adapt in comparison to those of figure.

3.3.2. Robust topological networks

As we argued above, very different processes in biology, sociology, economy, etc can be modeled by interactions between units or agents, which can be captured by graphs. The robustness of such networks is relevant to the wide variety of processes that these fields describe. Electric grids, as an example of economic relevance, should be designed to resist different kinds of failures [263, 150]. This justifies a large activity in the study of robustness with theoretical network models during the last decades [44, 43, 269]. The right functioning of the dynamical processes going on between nodes rely on the network structure persisting despite arbitrary circumstances that might damage the connections or the nodes themselves. Consider, e.g., manufacturing defects, which are usually spread randomly across the components. On the other hand, if a node

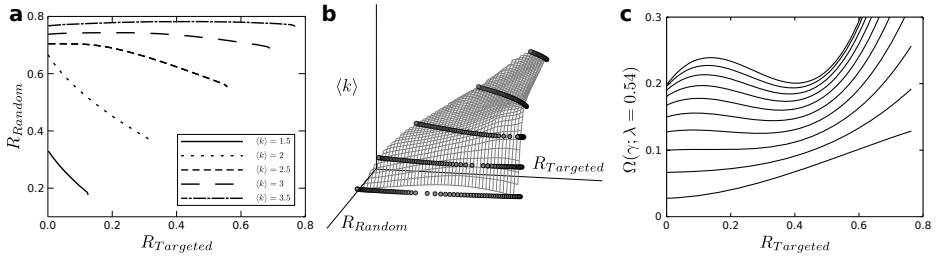


Figure 3.10: **Pareto optimal networks against targeted and random attacks.** **a** Pareto fronts reconstructed from [250]. **b** Pareto front projected onto a 3-D space where a new constraint (average degree of nodes) adds up a new dimension. The corresponding Pareto front presents a critical point. **c** The characteristic double well of first order phase transition appears as we move close and beyond the critical point.

or an edge is heavily used the chances increase that it might fail. This is similar to a malevolent attack that seeks to shut down key elements in a graph. These two different sources or network failure (random malfunction vs. directed attack) invite us to consider Pareto optimality, as Priester, Schmitt, and Peixoto have done [250].

They employ *stochastic block models* [174] in which, instead of networks, a generative model is given that samples from an ensemble of graphs. Each model specifies: a number of blocks, the number of nodes and their average degree within each block, and the number of connections between pairs of nodes. From these elements it is easy to compute a series of traits (averaged over the ensemble). Importantly: the size $S(q)$ of the largest component after a fraction q of nodes has been removed (either randomly or by directed attack). The integral $\int_0^1 S(q)dq$ over every q is used as a measure of robustness (R_{Random} or $R_{Targeted}$ depending on the kind of attack performed). The parameters defining these generative models are explored through an MOO algorithm that converges towards the optimal R_{Random} – $R_{Targeted}$ tradeoff, shown in figure 3.10**a** for different average degrees, which adds up an extra external constraint.

The different fronts trade between stochastic block models that generate core-periphery graphs [240, 250] and those that produce Poisson-like randomly connected networks. The former network kind contains an abundantly linked core with large average degree and a sparse periphery. This topology is less robust to targeted attacks. Fully Poissonian graphs at the other extreme of every front nimbly resist directed attacks since they do not contain any remarkable targets.

The most striking feature of the Pareto fronts in figure 3.10**a** (reproduced from [250]) is that they change their convexity as a function of the average degree. This is also illustrated by the surface in figure 3.10**b**. That surface is a Pareto front itself – just add the external constraint of the average degree. Based on the framework introduced in chapter 2, this surface with a vanishing cavity reflects a first order phase transition that ceases to exist as a control parameter is varied. This also implies the existence of a critical point similar to the one associated to the liquid-vapor phase transition. Successive energy landscapes show with clarity how the typical double well pattern emerges out of a single well as a control parameter is varied (figure 3.10**c**).

The parallelism with thermodynamics becomes even deeper thanks to the stochastic block model: points over the fronts in figure 3.10 do not correspond to individual networks, but to generative models describing full ensembles just as the canonical and microcanonical ensembles do in statistical mechanics (chapter 2, section 2.2) – hence figure 3.10**b** is quite close to a Gibbs surface in every sense. The existence of a critical point invites us to further explore its properties, potentially related to a great flexibility in dealing with different constraints (not only with the ones contemplated here).

3.3.3. Protein folding

Finally, let us look at a very relevant problem in biochemistry and bioinformatics: that of Protein Structure Prediction (PSP)[1]. Proteins are

[1]We follow [78] in noting that PSP focuses in the static aspect of the more general problem of Protein Folding (PF), which would include the dynamics as well.

the building blocks of cellular structure and function, and are hence involved in every task including signaling, enzymatic reactions, cell integrity, molecular transport, intercellular communication, structural stability of cells, and cell replication among others in a very long list. The correct performance of their tasks relies on their folding into the right shape. Failing to do so often leads to malfunction and disease. A relevant example are those maladies, such as Creutzfeldt-Jakob's, caused by prions. These are proteins folded in configurations that not only impede their proper functioning, but also enables them to induce their improper folding onto other proteins, thus propagating through tissues and organs. Understanding protein structure allows us to understand the dynamics of these processes and, perhaps, how to revert them. The invasion of cells by virus or bacteria relies on key proteins too. Knowing their shape can lead us to design molecules that mechanically block their activity – i.e. drugs to prevent or cure an infection.

Protein structures are determined through X-ray crystallography or nuclear magnetic resonance but, as pointed out in [78], these costly methods are often slow, and there is a huge number of proteins that we would want to investigate. On the other hand, DNA sequencing is extremely cheap and standard templates of the most important proteins are known across many species. This gives us a list with the linear chain of amino acids making up a protein, but it does not tell us how they fold. Much effort has been put in PSP. The problem consists in guessing the stable folding of a protein from those amino acid sequences. Each amino acid is itself a molecule whose atoms and electrons interact with those of neighboring peptides in different ways. Through these interactions a free energy landscape is defined. This can exhibit a number of energy minimums on which stable protein shapes should dwell owing to the laws of physics. This energy can be calculated from accurate theories, such as quantum mechanics; but this can be computationally expensive and approximations shall be used instead – e.g. statics, classic electrostatics, van der

Waals forces, etc. A typical energy function would read:

$$E(\boldsymbol{R}) = \sum_{bonds} B(\boldsymbol{R}) + \sum_{angles} A(\boldsymbol{R}) + \\ + \sum_{torsions} T(\boldsymbol{R}) + \sum_{non-bonded} N(\boldsymbol{R}), \tag{3.11}$$

where $\boldsymbol{R}$ represents the position of each atom in the protein and the different terms ($B(\boldsymbol{R})$, $A(\boldsymbol{R})$, $T(\boldsymbol{R})$, and $N(\boldsymbol{R})$) stand for bond energies, mechanical energies due to angles or torsions of the peptide chain, or so-called non-bond contributions (due, e.g., to van der Waals forces) respectively.

Many more details can be included, but it is enough for our purposes to know that an equation similar to 3.11 must be minimized, for which SOO is the usual approach. This is sometimes well justified because energy minimization might not admit control parameters – think, e.g., of the *kinetic* plus *potential* energy of a classic pendulum. But occasionally the different forces do depend on thermodynamic variables such as temperature or pressure. More challenging scenarios show up if we consider solvation and the role of water in filling in pockets within a molecule, which might induce local modifications to the energy function. Finally, energies such as equation 3.11 include numerous terms, often in conflict with each other. In such cases, SOO is prone to finding local solutions that depend largely on arbitrary initial conditions. PO often aids in avoiding local minimums even if SOO is used [341, 276]. Considering all these circumstances, we wonder whether MOO can give us important insights not only about optimal protein structures but also about the relationships that (through the Pareto front) Pareto optimal protein shapes may share with each other.

An MOO approach to PSP is pursued by Cutello et al. [78], who consider bound vs. non-bound energies. The former consist of forces due to the stretching of chemical bonds, Urey-Bradley energies, others associated to angles, torsions, and impropers; while the non-bound terms include the electrostatic and van der Waals forces between pairs of atoms. We do not need to understand all these standardized terms, which are

handled by computational packages such as CHARMM [50, 51], the tool used in [78].

Because all these contributions are grouped as bound vs. non-bound, we have a two-dimensional MOO problem whose Pareto front is a curve on the $t_1 - t_2$ plane (figure 3.11**a**). Cutello et al. derive the fronts for a series of proteins. As argued above, it would be interesting to analyze the Pareto optimal ensemble of protein shapes. From that ensemble we extracted phases and phase transitions (figure 3.11**b**). Given the ultimate physicality of this problem, these transitions shall show up in nature and be connected to released heat, work exerted against pressure, etc. Such relevant quantities may be obtained from the energy potentials (figure 3.11**c**). Note that those potentials (that we derived from the geometry of the fronts in [78]) must be correct up to a rescaling of the units.

We argue, hence, that considering Pareto optimal protein shapes as an ensemble reveals interesting insights about PSP and we hope that our findings can be useful to future explorations of this PO problem. However, instead of analyzing this ensemble perspective, Cutello et al. focused in one of the most interesting aspects: that of finding the *right* protein structure among all the candidates. Therefore they needed to chose one Pareto optimal solution. They relied in the intuitive concept of “knee”, which is well established in the MOO literature [83, 40, 140]. A knee is found in those regions of the front where a small displacement in either direction leads to a comparatively large loss in either of the targets. (Graphically this translates in a prominent convex bump in the Pareto front. The best illustration encountered so far might be figure 1.5**b** for Pareto optimal models of spiking neurons.) Knees often correspond to phases in the framework presented in chapter 2, but some methods (e.g. those in [140]) might detect knees that do not correspond to a phase.

Following [140], Cutello et al. locate relevant protein structures among those that are Pareto optimal [78]. They also compare those shapes to a series of known conformations (e.g. the native, functional shape), finding good agreement. At least in two of the examples the selected structures lay within a cavity of the Pareto front, meaning that they have a higher free energy than some other shape for every configuration of the control pa-

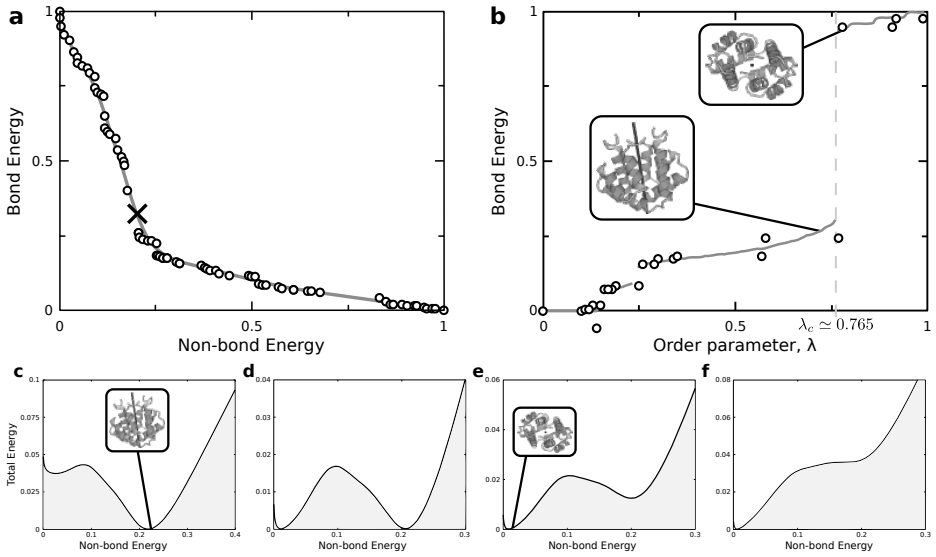


Figure 3.11: **Pareto optimal Protein folding.** **a** Pareto front for the simultaneous minimization of bonding and non-bonding energy in the 1UTG protein. A prominent cavity is revealed that implies a first order phase transition as the control parameters trading both kinds of energy change. Following established methods from the MOO literature [140] it is found that the best protein structure is around the region marked with a large cross. Such structures lay within the cavity of the front. **b** Protein shapes laying in the convex hull of the Pareto front are visited for different values of the control parameter. Foldings within the cavity (including the one selected in [78]) are bypassed by a first order phase transition. (Protein drawings correspond to 1UTG, but not to the actual foldings at either phase. They are intended as an illustration.) **c-f** This phase transition is physical in nature so that these Energy landscapes have a direct interpretation in terms of necessary work or heat to fold/unfold the protein, or of energy barriers that must be overcome. The geometry of the Pareto front may offer key information about the dynamics as control parameters are varied.

rameters of the free energy. Put otherwise: these shapes cannot be stable protein foldings. Because of the ultimate physical nature of this problem, the fronts in [78] *must* be considered under the light of the framework in chapter 2, suggesting that alternative protein shapes need to be selected. Alternatively, we might leave the ideal of having just one protein shape and study all the Pareto optimal solutions as an ensemble. In either case, both Pareto optimality and the methodology that we advocate for become relevant for PSP.

Chapter 4

EXPLORING THE CRITICALITY OF HUMAN LANGUAGE

Communication is based on the interchange of signals between agents. Because these signals happen through a physical channel, they entail a cost that is to be traded by the amount of transmitted information. The search for efficient communication constitutes a fundamental compromise that we will find in disguise here and in chapter 6. Because communication is carried out by agents, these shall try to push the different costs onto their counterpart using different strategies and depending on their role in communication. In a nutshell: lazy speakers shall be ambiguous and hope that the receptors of their message will worry about cracking its content; while hearers would wish to find as little ambiguity as possible, which should be solved by a speaker's accurate phrasing. This establishes yet another communicative tradeoff that will be the matter of this chapter. We hypothesize that different communication systems can be localized in a Pareto front, and that phase transitions or critical phenomena will in turn be manifested as different aspects of communication. Human language should itself occupy a place within this tradeoff, and we speculate that it must be a special place indeed, so that the notable flexibility and

complexity of natural language can be accounted for. We begin by reviewing the existing literature on so-called *least-effort language*, which already offers important hints to guide the validation of our hypotheses. Then we move on to review the problem in the context of Pareto efficiency and to analyze the important shape of the Pareto front and its content.

4.1. Ambiguity in language networks

4.1.1. Introduction

One of the latest and yet more profound evolutionary transitions involved the appearance of a new form of communication. Human language represented the triumph of non-genetic information, in a scale and quality that allowed a virtually infinite repertoire of meaningful constructs out of a collection of basic lexical units. Cultural evolution became a major player in shaping the character of human societies [207, 144].

It is fair to say that language, and human language in particular, has received the most dedicated multidisciplinary efforts. These include a vast range of fields, from genetics and anthropology to cognitive sciences, artificial intelligence or game theory. And yet, despite its undeniable importance, the origins of language remain largely unknown. Moreover, a graded transition to this complex form of communication does not exist. It is a sharp, drastic change what mediates between human languages and other animal communication systems. This enormous gap makes difficult to retrieve information by comparing our tongues to any midway stages.

We deal with a complex system that involves multiple scales and intricate interactions between levels and component units [6]. As such, a proper approach to its complexity needs a framework that explicitly considers systemic properties. Born by this complexity, language displays all kinds of apparently odd features, from the sometimes quirky appearance of syntactic rules to the ubiquitous presence of ambiguity. Ambiguity is specially puzzling: it seems to make little sense when we consider language from an engineering perspective or even under a standard optimization view based on communicative pressures [248, 56]. Under this

view, selection for comprehensible symbols would act removing unreliable components, thus reducing ambiguous features to the minimum.

Following the optimization line of thought, the ultimate basis of our discourse will be that a *least effort* principle is a driving force of languages. Always focused on this argument, in this chapter we present recent theoretical advances that share a common systems-level perspective of language structure and function. We adopt a non-reductionist approach towards human language [175, 299, 176] that largely relies on a network view of its structure – closer to a structuralist view of evolution. Within this scheme, constraints and genuine, endogenous features manifest themselves promoting (and/or being masked behind) universal statistical regularities. The discussed theoretical arguments are preceded by the description and discussion of experimental facts – always following the same systemic approach – that clearly show the kind of universal traits that we refer to and that happen to pervade every known language.

4.1.2. Scaling in Language

Language structure has been very often contemplated under the perspective of word inventories. The properties of isolated words and how these properties can be used to classify them within given general groups provide a first way of studying language architecture. The abundance of words, how they become adopted over language acquisition, or how different levels of language structure shape their relative importance define major research areas within linguistics. When exploring word inventories, one is faced with a dual character of languages that confronts the heterogeneity of tongues with the deep universality of a variety of their traits.

So, on the one hand languages are diverse. This is reflected in several features displayed by its constituents. Word inventories obviously differ from one dialect to another. Many characteristics, such as the number of letters in a word, show a statistical pattern with a distinctive single-hump distribution, but the average number of letters is rather different across

languages. In Mongolian or German this is close to 12 letters per word, whereas for Croatian or Serbian this drops down to around seven. The diversity in this trait might originate in historic contingencies idiosyncratic to each language and is not – a priori – the kind of universalities that we wish to study.

On the other hand, all languages seem to share some remarkable universal patterns, best exemplified by the so called Zipf's law [351]. Earlier noted by other authors, but popularized by G. K. Zipf, this law states that the frequency of words in a given word inventory – such as the one we can obtain from a book – follows a universal power law. Specifically, if we rank all the occurrences of words in a given text from the most to the less common one, Zipf's law states that the probability $p(s_k)$ that in a random trial we find the k-th most common word s_k (with $k = 1, ..., n$) falls off as:

$$p(s_k) \quad = \quad \frac{1}{Z} k^{-\gamma}, \tag{4.1}$$

with $\gamma \approx 1$ and Z the normalization constant – i.e., the sum $Z = \sum_{k \leq 1} k^{-\gamma}$. We can observe this regularity in any modern human language when analyzing any adequate corpus. This is the kind of traits that we are interested in, and of which we demand an explanation with the hope of gaining a deeper understanding about the origins of language or the constraints that shape it.

Roughly speaking, Zipf's law tells us that the most frequent word will appear twice as often as the second most frequent word, three times as often as the third one, and so on. Instead of using a word's rank, an alternative form considers the use of the standard probability $p(m)$ that we come across a word that is repeated m times throughout a text. Then the corresponding Zipf's law scales as:

$$p(m) \quad = \quad \frac{1}{Z} m^{-\alpha} \tag{4.2}$$

where now the normalization constant is $Z = \sum_{m \leq M} m^{-\alpha}$ with M the maximum observed frequency. Now the scaling exponent is $\alpha = 2$. In

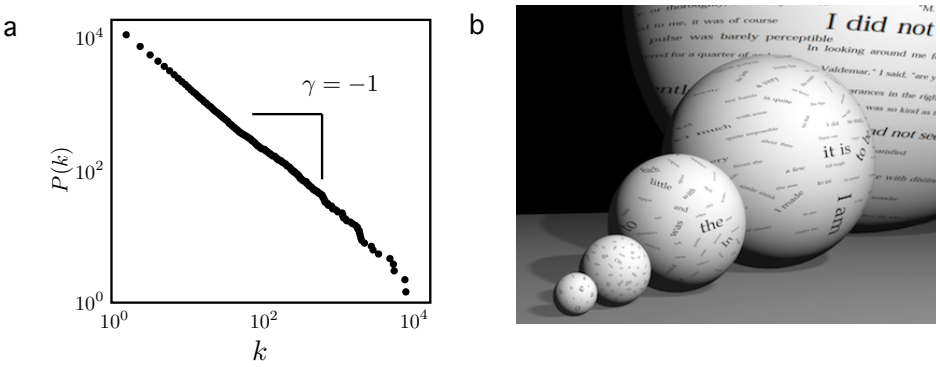


Figure 4.1: A seemingly universal feature of all known human languages is Zipf’s law, illustrated in Figure **a** from the rank-abundance statistics obtained using Melville’s Moby Dick (see text). Moreover **b** language contains multiple levels of nested complexity, illustrated here by means of an idealized collection of spheres whose size rapidly grows as the objects being considered at one level are combined to obtain those in the next level. Letters and syllabus are the first levels, followed by words and pairs of words and eventually sentences. The diagram actually underestimates the real proportions of combinatorics.

Figure 4.1**a** the frequency-rank distribution of words collected from Herman Melville's Moby Dick is shown in logarithmic scale. The scaling law $P(k) = k^{-\gamma}/Z$ is plotted against the rank k. The logarithmic plot provides a direct way of testing the presence of a scaling law, since it gives a linear relationship:

$$\begin{aligned} \log p(k) &= \log\left[\frac{1}{Z}k^{-\gamma}\right] \\ &= -\log[Z] - \gamma \log k, \end{aligned} \tag{4.3}$$

the slope of which is the scaling exponent γ.

The widespread, virtually universal presence of Zipf's law in all known languages, and perhaps even in the context of DNA and the genetic code [198, 273, 231] suggests two potential interpretations. It might be the case that the observed scaling is so widespread that it is essentially a meaningless signal. (Note the discussion about Zipf's law in random texts [110].) The other possibility is that its universal presence has to do with some relevant feature shared by all languages, perhaps associated to some deep functional role. Given the disparate trajectories followed by human languages over their evolution, it seems unlikely that such a unique scaling law would be so robust unless it involves a relevant constraint.

An additional component related to the logical organization of language deals with its enormous combinatorial potential. Language defines a non-genetic form of heredity and as such allows rapid cultural exchanges, the formation of a collective memory, and an enormous plasticity while facing environmental challenges. Its success is tied to the brain's capacity for storing a large number of communication elements. However, an inventory of words can only be part of the whole story. Another important aspect must be the associations that these units can build between them. This will be treated in more detail in the next section. Let us explore first the scaling facet of such associativity to have a scope of the relevance of the generative power of language.

Words are combined and related to each other in multiple ways. Such combinatorial potential pervades all linguistic levels from phonemes to

whole texts. As we move towards higher levels, the potential universe of objects expands super-exponentially (Figure 4.1**b**). We can appreciate this inflationary behavior explicitly when moving from words to sentences to texts. Let us assume a set of words $\mathcal{L}'$ is sampled from the whole repertoire of words defining a language $\mathcal{L}$ (i. e. $\mathcal{L}' \subset \mathcal{L}$). Our set $\mathcal{L}'$ is finite and involves $|\mathcal{L}'| = N_w$ words. Of course the combinatorial nature of word arrangements easily explodes with N_w. Now consider a finite (but long) written text, to be indicated as $\mathcal{T}$. It is composed by a set of M sentences S_μ, each one formed by an ordered, sequential collection of words extracted from $\mathcal{L}'$:

$$S_\mu = \{w_{1,\mu}, w_{2,\mu}, ..., w_{n_\mu,\mu}\} \tag{4.4}$$

with $\mu = 1, 2....M$ and thus we have our text defined as the union:

$$\mathcal{T} = \bigcup_{\mu=1}^{M} S_\mu \tag{4.5}$$

If we indicate by $|S_\mu|$ the length of a given sentence, the average sentence size in $\mathcal{T}$ will be

$$\langle S \rangle = \frac{1}{M} \sum_\mu |S_\mu| \tag{4.6}$$

A very rough first approximation assuming that all components can be combined in similar ways – i.e. leaving syntactic constraints aside – provides a total number of (possible) sentences as given by the power law:

$$|\mathcal{T}| \sim N_w^{\langle S \rangle} \tag{4.7}$$

which gives, for $N_w \approx 80000$ and $\langle S \rangle \approx 7$ (two reasonable estimates) a hyperastronomic number: 2.097×10^{34}. In natural language, many of these combinations will never appear, most of words will be extremely rare and a few of them extremely frequent (as we saw above) since there exist nontrivial rules for a string of symbols to make sense as a word of $\mathcal{L}$. The plausibility of a sentence existence and its frequency will be

constrained as well because there are further nontrivial (syntactic) rules for the use of words from $\mathcal{L}$ in a real context. Nevertheless, this quick calculation allows us to grasp the scope of the expressive power of this system.

The enormous potential for combination that is present in human language embodies the uniqueness of such complex form of communication. No other species in our planet shares such a spectacular capacity and a chasm seems to exist between us and all the other species inhabiting our planet. This uniqueness is also interesting for another reason. Major innovations that have occurred through evolution have been found independently a number of times. Multicellularity, sight, or sex have emerged in many different groups through different paths [207, 136, 180] thus indicating that the same basic innovations can be obtained following different paths. By contrast, the complex communication system that we use as a species is unique [207]. No other parallel experiments in evolution leading to such achievement have taken place.

However, storing words is one thing; combining them, another; and being able to relate each other in a flexible, efficient manner is yet another one. Our potential for storing a large inventory of words together with an astonishing potential of relating them in complex ways through intricate paths (sentences being just one of them) is at the core of the evolutionary success of humans. In this chapter we consider language organization in terms of a statistical physics picture, where networks instead of word inventories play a central role. By using them, we will argue that ambiguity is an expected feature of human language, and a specially relevant and perhaps inevitable one. It is ambiguity what hides behind Zipf's law and an essential element that makes our use of language so efficient and flexible.

4.1.3. Small World Language Networks

In our previous illustration of the combinatorial potential of language, we used sentences as higher-order structures obtained as linear chains that combine words in syntactically meaningful ways. Sentences provide us

with a first example for the kind of recursive linguistic structures that we are capable of forming. They will also serve us to introduce networks and how language can be interpreted in terms of these complex webs.

The simplest case of language network that can be introduced is defined in terms of co-occurrence [108]. Two words in a sentence that appear one after the other are said to co-occur. We will build a graph (a network) using these words and their co-occurrence as follows: Words w_i $(i = 1, ..., N_w)$ are the fundamental units, defining a set W. The relationships between words are encoded in a matrix $\Gamma = \{a_{ij}\}$ called the *adjacency matrix*. An undirected link $a_{ij} = 1 = a_{ji}$ will be defined between two words $w_i, w_j \in W$ if they follow one another within at least one sentence (otherwise the matrix element is set to $a_{ij} = 0 = a_{ji}$). The resulting *language production network* (LPN) Ω_L is thus defined as a pair $\Omega_L = (W, \Gamma)$, where $\Gamma = \{a_{ij}\}$ constitutes the set of unweighted links of the graph. It should be noticed that the mapping $\Gamma : W \to W$ is expected to capture some of the underlying rules of word ordering. This web provides in fact a glimpse to the production capacity of the underlying grammar structure and shares, as we will see below, a large number of common traits with syntactic webs [113].

In Figure 4.2**a** we display an example of LPN network. This particular one has been obtained from the words that appear in Paul Auster's short story *Augie Wren Xmas tale*. Here spheres correspond to specific words and connections among them indicate that the pair of word co-occurred at least within one sentence throughout the tale. The size of the spheres has been increased in some cases to indicate their high frequency of appearance in the text. Several interesting features need to be noticed. One is that the network is highly heterogeneous: a vast majority of words have only one or two links with others, whereas a small number of them (the hubs) have a very large number of connections. These super connectors can be seen in Figure 4.2**c** and correspond to words that are very common and highly ambiguous. Figure 4.2**b** gives us a glance of the “local” organization stemming from the sentence structure. We can actually read well defined chains that make sense in a given direction. These readable chains become less and less common as the size of the word inventory

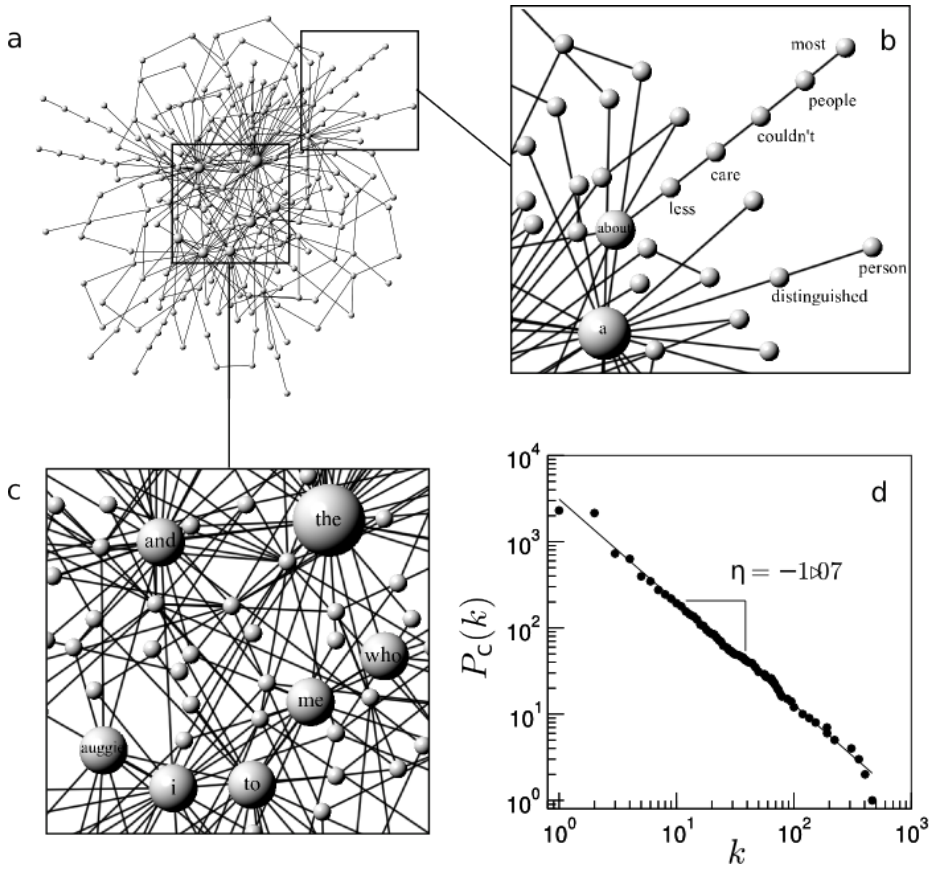


Figure 4.2: A language network can be build in different ways. The simplest one is considering co-occurrence between words within sentences from written corpuses. Here **a** we have used the first chapter of Paul Auster's "Augie Wren Xmas Tale", from which we draw our network. Each ball is a different word, whereas an undirected link between two balls indicates that those words appeared one after the other within a sentence in the text. Two parts of the web are zoomed in **b** and **c**. In **b** we observe multiple linear structures and chains associated to particular sentences. Meanwhile, in **c** we can see that some words have a very large number of links with others and are referred to as "hubs", whereas most words have just one or two connections. LPNs follow scale-free degree distributions, as exemplified in **d**.

Figure 4.2: This is the same statistical feature of word frequency illustrated by the Moby Dick data set (see section 4.1.2 and figure 4.1**a**).

grows and more and more crossings occur.

A distribution of connections, or *degree distribution* $P(k)$, can be defined by measuring the number of links k of each node (also known as its degree) and calculating the relative frequencies for each k. In a randomly wired graph of N nodes (where we simply connect every two elements with some probability p) the number of links associated to arbitrary words would follow a Gaussian distribution centered around the average degree value $\langle k \rangle = p(N-1)/2$. We call such a graph *homogeneous* because the average value represents fairly well everything that can be awaited of the graph. But many real networks – including language graphs – follow a functional form that displays a scaling law, namely

$$P(k) \quad = \quad \frac{1}{Z} k^{-\alpha}. \tag{4.8}$$

Once again, we have $Z = \sum_k k^{-\alpha}$ and, for all LPN networks, $\alpha \approx 2$. Let us note once more the remarkable universality of this observation: for any language, from any adequate collection of sentences, despite the disparity that both elements (languages and sentences – and collections of sentences, indeed) can present we will derive such a degree distribution with roughly the same exponent α; just as if some inner mechanisms of the human language were eventually responsible of such scaling. As opposed to the Gaussian, these kind of power law distributions feature an extreme variability that the average alone cannot capture. This is a consequence of the existence of a miscellany of structures within the network. The real world example from Auster's short story is shown in Figure 4.2**d**, where we have used (to smooth out the statistics) the cumulative distribution, defined as

$$P_>(k) \quad = \quad \sum_k^M P(k) \sim \int_k^M P(k)dk \sim k^{-\alpha+1} = k^{-\gamma}. \tag{4.9}$$

We find an exponent $\alpha \sim 2$, which is actually the same that we observed in Zipf's law (in its frequency form). This is not surprising, since there

is an almost perfect correlation between the frequency of a given word and the number of co-occurrences it can establish within W. Therefore, it could be argued that the only thing that matters is the frequency distribution of words: this would eventually determine the degree distribution. However, there is more to the structure of the network than this power law distribution of its degree k. To appreciate it we must look at some other traits.

A randomly connected graph following the previous $P(k)$ scaling would not recover many observable properties exhibited by the original graph based on co-occurrence. As an example, there is a widespread feature that is present in the LPN and not in a randomized version of it: hubs are usually not connected in the former but they can be so in the later. This particular result tells us that, despite not being a true syntactic network, LPNs do preserve some essential constraints associated to syntactic rules.

There is another interesting property. The LPN graph is sparse: the average number of connections per word is small. Despite this sparseness and the local organization suggested by the previous features, the network is extremely well connected. In complex networks theory, this is known as a *small world* graph [342, 5]. Small world networks were first analyzed by Stanley Milgram in the context of social ties within a country [212]. It was found that only a small number of links separates, within the network of social acquaintances, two randomly chosen individuals. Since a given country involves millions of humans, the basic result – that only about six jumps are needed (on average) to connect any two random persons – was highly surprising. This qualitative property can be quantified by means of the *average path length* (D) defined as $D = \langle D_{min}(i,j) \rangle$ over all pairs $w_i, w_j \in W$, where $D_{min}(i,j)$ indicates the length of the shortest path between two nodes. Within the context of a LPN, a short path length means that it is easy to reach a given word $w_i \in W$ starting from another arbitrary word $w_j \in W$. The path cannot be interpreted here in terms of meaningful trajectories (such as sentences) but instead as a measure of accessibility.

An additional measure of network organization that characterizes small world graphs is the so called *clustering coefficient* (C). It is defined as the

probability that two vertices (words, in our context) that are neighbors of a given vertex are neighbors of each other as well. In order to compute the clustering, we associate to each word w_i a neighborhood Γ_i, defined as the set of words linked to w_i, i. e.

$$\Gamma_i = \{w_k \in W \mid a_{ik} = 1\} \tag{4.10}$$

Each word $w_j \in \Gamma_i$ has co-occurred at least once with w_i in some sentence. The words in Γ_i can also be linked among them. The clustering $C(\Gamma_i)$ of this set is defined as the fraction of triangles found, compared to the maximal number expected from an all-connected scenario. Formally, it is given by:

$$C(\Gamma_i) = \frac{1}{k_i(k_i-1)} \sum_j \sum_{k \in \Gamma_i} a_{jk} \tag{4.11}$$

and the average clustering is simply $C = \langle C(\Gamma_i) \rangle$. Many triangles in a sparse graph indicate an excess in local richness of connections. Such an excess needs to be compared with a null model of random connections among words – i.e. with a randomized version of the LPN as we did to compare the likelihood that the hubs are connected.

Concerning the average path length, for random graphs with Poissonian structure – i.e. with nodes simply connected with a probability p and thus their degree distribution following the more standard Gaussian distribution – it is possible to show that we have a logarithmic growth in the number of degrees of separation with N [342, 5]:

$$D_{random} \approx \frac{\log N}{\log \langle k \rangle}; \tag{4.12}$$

whereas the clustering is expected to decay inversely with system size – i. e.

$$C_{random} \approx \frac{1}{N}. \tag{4.13}$$

On a first approximation, it is said that a network is a *small-world* when $D \approx D_{random}$ whereas the clustering coefficient is much larger $C \gg C_{random}$ [342, 5]. LPNs happen to be small worlds, as remarked above. This nature of LPNs and other language networks tells us that despite their locally ordered, correlated structure (far from that of a random graph) association and routing between words can be highly efficient.

Network theory does not offer a full explanation for the cognitive substrate responsible for word association and optimal search – this last property being related to the easy navigation that small worlds enable. This theory does provide, though, a valid formal framework within which relevant questions can be consistently stated. Hopefully, the answers attained also constitute compelling knowledge about human language.

4.1.4. Ambiguity in Semantic Networks

The relational nature of language can be analyzed from different scopes including semantics, syntax, morphology, and phonology [253, 254, 55, 267, 330]. They define the different relationships between units and the structures made by such units. We saw a syntactic example in the previous section. Moreover, at the community level social interactions also describe a web within which languages are enforced. This social structure can play a determinant role, for example, in the success or failure of a contingent linguistic trait and even in the emergence of further universal regularities [298]. We see that network theory is not only useful but perhaps inescapable to understand our communication system. All these networks must somehow contain information concerning the way in which components – generally, but not necessarily, words – are organized within sentences or how they are related in terms of their semantic content. The links can thus have a very different nature in each graph and the overall patterns of organization of such graphs do not need to be the same.

A prominent subfield of linguistics, semantics has traditionally been defined as the study of the meaning of (parts of) words, phrases, sentences, and texts. Semantic organization is a widely explored topic in psy-

cholinguistics. As a search for an adequate characterization of meaning, semantic relations have strong ties with memory and categorization. Semantic relations are also known to deteriorate in patients with Alzheimer's disease and other types of brain impairment [49]. Such a semantic decline can also be appreciated in the kind of properties (e.g. Zipf's law) that we are interested for other diseased patients [114].

Semantic networks can be built starting from individual words that lexicalize concepts and by then mapping out basic semantic relations such as isa-relations, part-whole, or binary opposition. They can potentially be built automatically from corpus data [179, 217, 294, 154, 313] and also from retrieve experiments in which subjects are asked to quickly list down words as they come to their minds [313, 131]. One of the most interesting efforts in understanding the organization of semantic relationships is the Wordnet project [213, 107]. This data set explicitly defines a graph structure where words from the English lexicon are connected through various kinds of semantic links. A possible subset of such kind of web is displayed in Figure 4.3**a**-**b**. As pointed out by Sigman and Cecchi [294] mental concepts emerge as a consequence of their interrelationships, and meanings are often related through chains of semantic relations. Linking “stripes” with “lion” requires following a mental path through a sequence of words, such as lion-feline-tiger-stripes [313]. Different paths are possible on a semantic network – as exemplified in Figure 4.3**a**-**b** – and experience shows that we find them easily despite the very large set of items potentially available.

The efficient character of the semantic network is associated to an important, universal, and yet apparently undesirable property of language: polysemy. All languages exhibit polysemy, meaning that a given word form corresponds to two or more meanings. At first sight we would think that polysemy is a rather undesirable feature, since some ideal language should be expected to avoid such ambiguity. The analysis of the large-scale architecture of semantic networks reveals a likely reason for polysemy to exist and be so widespread. The answer lies on the global organization of these graphs which are both highly heterogeneous and exhibit

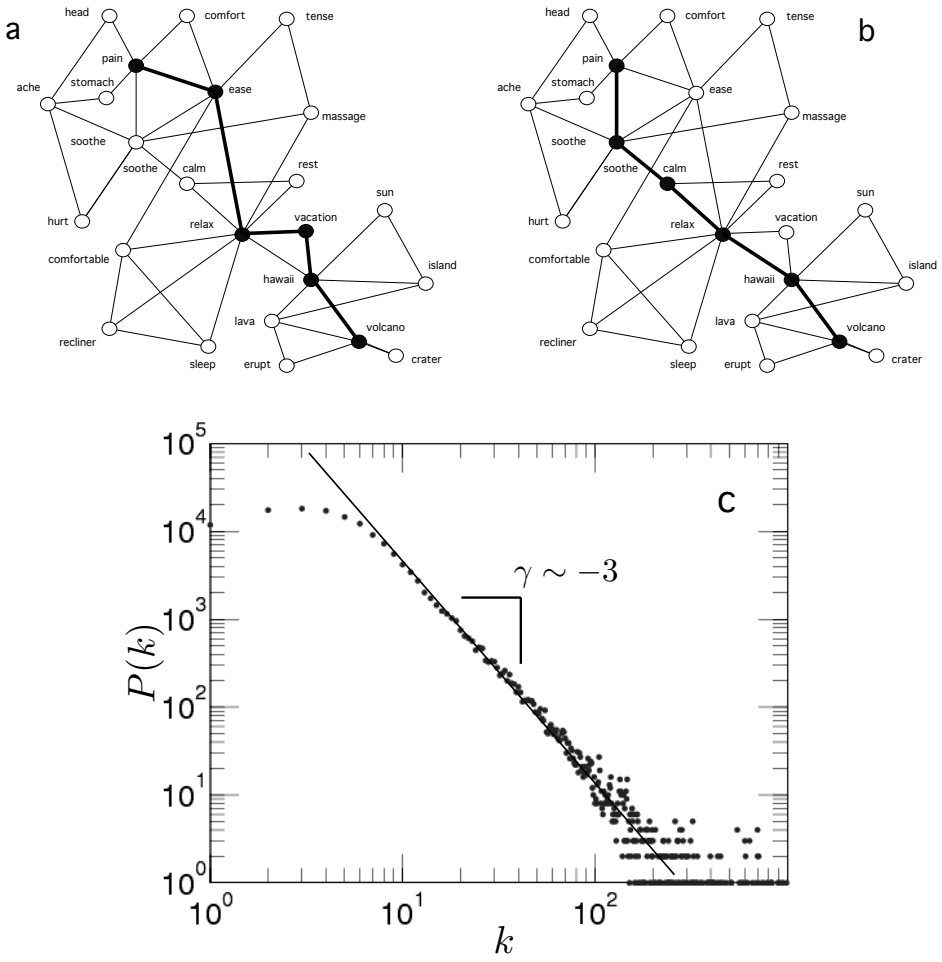


Figure 4.3: **A simple network of semantic relations among lexicalized concepts.** Nodes are concepts and links, semantic relations between concepts. This would correspond to a very small subset of a vast set of words and semantic relationships. Associations between words allow us to navigate the network. Locally, the number of triangles is very large, allowing multiple ties among semantically related words – and contributing to a high clustering, as seen in the text. Moreover, given two words, such as “volcano” and “pain” can be linked through different paths, two of which are illustrated here using thick lines.

Figure 4.3: The degree distributions associated to these semantic graphs are broad, with fat tails. In **c** we display the distribution of links for WordNet, with a scaling exponent close to three.

the small world phenomenon. The network analysis of Wordnet shows a scale-free structure (Figure 4.3**c**) where most elements would be more specialized, and thus semantically linked to just a few others. By contrast, a few of them would have a large number of semantic links. As before, we have a degree distribution $P(k) \sim k^{-\gamma}$, now with $\gamma \sim 3$ and thus a higher scaling exponent that indicates a much faster decay in the frequency of high-degree elements. This network is a small world *provided that polysemy is included*. The high clustering found in these webs favors search by association, while the short paths separating two arbitrary items makes search very fast [217] even if distant fields need to be reached. Additionally, as discussed in [313], the scale-free topology of semantic webs places some constraints on how these webs (and others mentioned above) can be implemented in neural hardware. This is a remarkable example of how statistical regularities could be hiding a very relevant constraint of language evolution.

To summarize, the mapping of language into networks captures novel features of language complexity far beyond word inventories. It provides further evidence for universal traits shared by all languages and how to characterize and measure them. More interestingly, they suggest novel ways to approach old questions related to language efficiency and how it might have evolved. But they also allow us to formulate new questions that could not be expressed without using the network formalism. Among them, how these network patterns might emerge and how they might be linked to Zipf's law. In the next section, we will review a model of language evolution that also involves graphs and that is based on an early proposal by Zipf himself. That model provides a first approximation to the potential components that make human language unique. It turns out that ambiguity might actually be a key component behind some of our more remarkable singularities.

4.1.5. The Least-Effort Language Agenda

As we insisted throughout the text: statistic regularities are a narrow window that allows us to glimpse the existence of universal laws driving the emergence and evolution of human languages. Zipf's law remains the most singular of such universal observations. Opposed to partial collections of words – such as the analysis performed on *Moby Dick* in Section 4.1.2 – a careful analysis of extensive corpora clearly indicates that the whole of a language does not feature the pattern observed by Zipf [109, 243]. Just a *core vocabulary* does so, but the observation remains universal anyway. Furthermore, recent analysis indicate that diseased patients as well as lexicon not in the core might follow a version of Zipf's law with a generalized exponent $\gamma \neq 1$ [109, 114]. In sight of this evidence, the general scientific intuition has a broad consensus about the importance of Zipf's law and efforts to find model explanations to it do not diminish over time.

In its original account, Zipf proposed that a tension between minimizing user's efforts and maximizing the communication power of a language would be the main driver towards the statistic regularity that he observed empirically, thus he coined the *least effort language* principle [351]. Our main concern in this section is not necessarily Zipf's law, but the least effort optimization as a mechanistic driving force – which, anyway, has been shown to be a mechanism for the generation of scale-free distributions [332]. There are strong evolutionary reasons why a least effort principle might be acting upon human languages. To appreciate the selection for least effort in communication we can adopt any of two complementary view points – both of which are visited in [348]. On the one hand we could argue that a human group with a more efficient code could enjoy an evolutionary advantage over other groups. Those with less adequate dialects would be selected against and their tongues would perish with them. The other possibility is to look at each language as a system enduring natural selection. We can conceive different codes simultaneously spreading over a population. Those fitter to be transmitted by humans – i.e. those better coping with our biological, social, and technological con-

straints – would be naturally selected for and become dominant. Because the fitness now is the ease of tongues to humans we can see a least effort driving language evolution quite directly, not necessarily through an intermediate step of human selection.

How can we approach language evolution from a sensible facet? There are in principle multiple ways and scales of approximation that can be used. They span an enormous range of views, from game- theoretic models to computational linguistic or language evolution in embodied, robotic agents. Perhaps the answer to previous questions needs to be tied to another, more basic one: What do we want to understand? Here we are concerned with ambiguity as part of the fabric of language organization. We would like to understand if ambiguity plays any role in how the previous scaling laws emerge and why there might be sharply defined classes of languages – perhaps separated by some sort of barrier – thus directly tackling the harsh gap between human and any other form of communication. Following the steps indicated in [112], we will use Zipf's least effort hypothesis to derive a model within which we can frame these kind of questions properly. We will ultimately study communication between pairs of agents sharing a given channel, so information theory (as formulated by Claude Shannon) is the natural framework.

In [112], the tension between simplicity and communicative power proposed by Zipf rests upon the trade-off between speaker and hearer's requirements of a language. The former prefers to name every possible object with the same signal – there lays their least effort to find an object's proper name – and the latter prefers to have a one-to-one mapping between available signals and existing objects, so that no decoding effort is necessary. Note that the speaker's option is the most ambiguous language possible in which communication is not possible. Meanwhile, the hearer's proposal is not degenerated at all. The conflicting needs of different users pose an evolutionary game for languages. These are modeled by allocations of available signals $s_i \in S$ (with $|S| = n$) to name existing objects $r_j \in R$ (with $|R| = m$). The assignments that identify a given tongue are encoded in the entries of a matrix: $A = \{a_{ij}\}$ with $a_{ij} = 1$ if signal s_i refers to object r_j and $a_{ij} = 0$ otherwise.

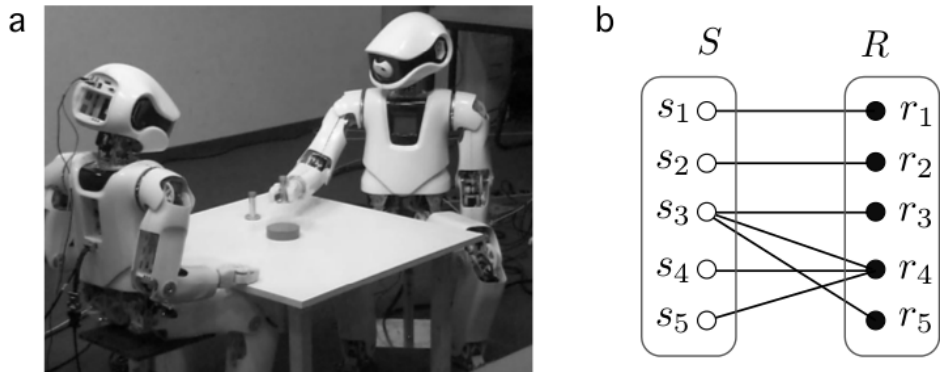


Figure 4.4: **Modeling language evolution through robots or matrices.** In order to model language evolution, one can use a number of artificial systems, including among them robotic, embodied agents **a**. Here two robots (image from the Neurocybernetics group at Osnabrück, see https://ikw.uni-osnabrueck.de/~neurokybernetik/) share a common environment seeded by a number of objects, which they can name. Robots can evolve a rudimentary grammar that goes beyond the simple word inventory that we could expect. Additionally, simple mathematical models can also be used in order to capture essential features of language organization. A model of language can be formulated in terns of a matrix **b** that relates a set of n signals (indicated as $s_1, s_2, ..., s_n$) with a set of m objects or actions of reference $(r_1, ..., r_m)$. A simple case with $n = m = 6$ is displayed. A signal is associated to an object using a link connecting them. Here for example signal s_5 is used to refer to object r_4.

Similarly to the matrices introduced in Section 4.1.3, A is known as an *adjacency matrix*; only before it linked elements from within a set to one another and now it connects the constituents of two different sets, R and S, thus accounting for their relationships and other relevant features. A very important trait is related to the presence of ambiguity. As defined, the model and its matrix representation include both polysemy (i. e. presence of multiple meanings associated to a given signal) as well as synonymy, where different signals refer to the same object. The two traits can be detected by direct inspection of the rows and columns of the adjacency matrix. If we look at the example given in Figure 4.4**b**, using $n = m = 5$ the matrix reads:

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \tag{4.14}$$

The A matrix structure apprehends both the capacity for a signal to have multiple meanings (by referring to multiple objects), and synonymy, where multiple signals refer to the same object. These two features are directly detectable here by looking at rows and columns within A. Synonyms are associated to vertical strings of ones, indicating that the same object r_k can be labelled or referred to by multiple (synonymous) words. Conversely, a polysemous word would correspond to a signal having multiple ones in a row. This contributes to the ambiguity of the language. In our example, r_4 is connected to three synonyms, whereas signal s_3 is used to label three different objects.

In [112] it is assumed that objects are recalled randomly with uniform frequency $p(r_i) = 1/m$. A speaker then chooses from among the available signals that name the required object in its language $A = \{a_{ij}\}$, yielding a frequency for each signal:

$$p(s_i|r_j) = \frac{a_{ij}}{\omega_j}, \tag{4.15}$$

with $\omega_j = \sum_j a_{ij}$. We will indicate the joint probability (of having a signal and a given object) and the corresponding probability of a given signal as:

$$\begin{aligned} p(s_i, r_j) &= p(r_j)p(s_i|r_j), \\ p(s_i) &= \sum_j p(s_i, r_j). \end{aligned} \tag{4.16}$$

We can write the entropy associated to the signal diversity, which in the proposed framework stands for the effort of the speaker:

$$H_n(S) = H(\{p(s_1), ..., p(s_n)\}) = -\sum_{i=1}^{n} p(s_i) log_n(p(s_i)). \tag{4.17}$$

Recalling our needs of information theory, Shannon's entropy $H_n(S)$ provides a measure of the underlying diversity in the system. It is also a measure of uncertainty: the higher the entropy, the more difficult it is to predict the state of the system. For this reason H is often considered a measure of randomness. Its maximum value is obtained for a homogeneous distribution. In our case, it corresponds to $p(s_i) = 1/n$ for all signals involved:

$$H\left(\left\{\frac{1}{n}, ..., \frac{1}{n}\right\}\right) = -\sum_{i=1}^{n}\left(\frac{1}{n}\right) log_n\left(\frac{1}{n}\right) = \log n. \tag{4.18}$$

Conversely, the lowest entropy is obtained for $p(s_i) = 1$ and $p(s_{k\neq i}) = 0$. For this single-signal scenario we obtain $H_s(S) = 0$.

Another key quantity involves the noise associated to the communication channel. Using the definition of conditional probability, namely $p(r_j|s_i) = p(s_i, r_j)/p(s_i)$ we define a measure of noise associated to a given signal as follows:

$$H_m(R|s_i) = -\sum_{j=1}^{m} p(r_j|s_i) log_m p(r_j|s_i). \tag{4.19}$$

This entropy weights the uncertainty associated to retrieving the right object object from R when signal s_i has been used. The average uncertainty is obtained from:

$$H_m(R|S) = \langle H_m(R|s_i)\rangle = \sum_{i=1}^{n} p(s_i) H_m(R|s_i). \tag{4.20}$$

For simplicity, let us assume $n = m$. If each signal were used to refer to a single and separated object, we could order our set of objects and signals so that $p(r_j|s_i) = \delta_{ij}$ where we define $\delta_{ij} = 1$ for $i = j$ and zero otherwise. In this case, it is easy to see that $H_m(R|s_i) = 0$ and thus no uncertainty would be present: given a signal, the right object can be immediately fetched without ambiguity. This corresponds to a perfect mapping between signals and meanings/objects. The opposite case would be a completely degenerate situation where a single signal s_μ is used to refer to all objects indistinctly. Then $p(r_j|s_\mu) = 1/n$ for all $j = 1, ..., n$. In this case, it can be shown that $H(R|S) = \log n$ – thus the uncertainty that the hearer faces is maximal.

Summing up, this conditional entropy $H(R|S)$ works as the average ambiguity perceived by the hearer, and thus stands for its effort when *decoding* language $A = \{a_{ij}\}$. Finally, both communicative costs are collapsed into the following energy function:

$$\Omega(\lambda) \quad = \quad \lambda H_m(R|S) + (1 - \lambda) H_n(S). \tag{4.21}$$

Using this as a kind of "fitness" function, an evolutionary search was performed in order to minimize $\Omega(\lambda)$. The minima obtained from this algorithm provide a picture of the expected graphs – as defined by the adjacency matrices – compatible with the least effort minimization principle.

Along with the relative efforts defined above, two key properties were also measured. The first is the information transfer (or mutual information) obtained from:

$$I(R, S) \quad = \quad H(S) - H(S|R), \tag{4.22}$$

which plays a central role within information theory and is interpreted as *how much information* do signals convey about which object needs to be retrieved. The second is the effective lexicon size $|\mathcal{L}|$, i. e. the number of signals that are used to name objects. This was defined as

$$|\mathcal{L}| = \left| \left\{ j | \mu_j = \sum_{k=1}^{N} a_{jk} > 0 \right\} \right|, \tag{4.23}$$

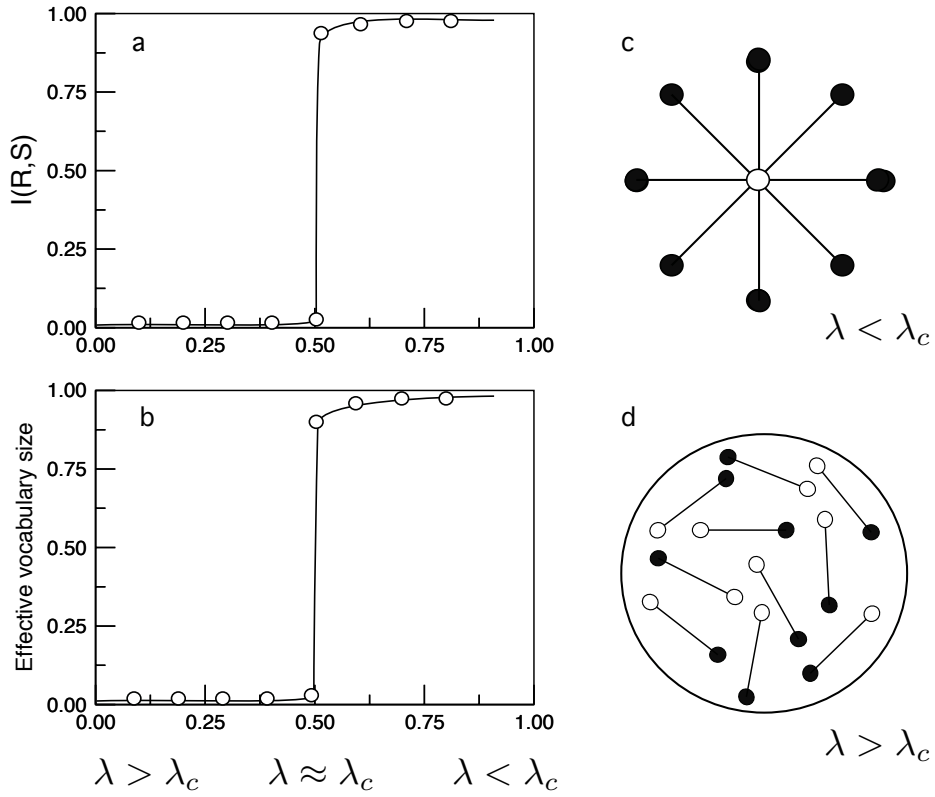


Figure 4.5: **Phase transition in least-effort language.** As we vary λ, equation 4.21 awards different importance to a speaker's or a hearer's requirements of a tongue. Accordingly, we move from a scenario that contents the former to one that pleases the later. But the change is sharp and happens at a very precise value of $\lambda = \lambda_c \equiv 0.5$, in accordance with the description of a first order phase transition. The simulations to generate these plots – a Genetic Algorithm (GA) that proceeded to minimize equation 4.21 with different values of λ – are in good agreement with this numerical critical value. Because of this sudden regime shift we can observe very abrupt changes in some *order parameters* than can be measured in a language: **a** The mutual information between signals and objects (whose average value across the top population of the GA is plotted) says how much information the signals of a language convey about the named world.

Figure 4.5: For $\lambda < \lambda_c$ one only signal serves to name every object – fully complying with the speaker's needs – and the language does not bear any information about the external world, thus communication is not feasible with such a language. For $\lambda > \lambda_c$ tongues map one-to-one between signals and objects and a maximal amount of information is conveyed. This is compared to animal codes in [112]. These require a perfect mapping, thus exploit the whole range of available signals as we can see in panel **b**, where the proportion of used signals to those available is reported. In **c** and **d** we represent the signal-object association graphs that emerge in the two extreme regimes: $\lambda < \lambda_c$ and $\lambda > \lambda_c$ respectively.

where μ_j actually indicates whether or not the signal is being used.

Clearly the meta-parameter λ weights the importance of the hearer and speaker's needs. In [112] a phase transition is uncovered at a certain value λ_c when varying λ between 0 and 1, as it is illustrated in Figure 4.5. For $\lambda < \lambda_c$ the speaker's effort is minimized and completely ambiguous languages are persistently achieved. The A matrix for the extreme case in a $n = 4 = m$ system would be

$$A = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{4.24}$$

As expected, in that scenario communication is impossible, given the complete degeneracy associated to the unique signal used to refer to every item within R. This is revealed by the vanishing mutual information between signals and objects (Figure 4.5**a**). Obviously the vocabulary requirements of this solution are minimal (Figure 4.5**b**). The word-object association graph that we would obtain is illustrated in Figure 4.5**c**.

For $\lambda > \lambda_c$ the one-to-one mapping preferred by the hearer (Figure 4.5**d**) is always optimal. In this special case, the adjacency matrix for the

signal-object association can be written in a diagonal form:

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{4.25}$$

Most models of language evolution that explore the origins of communication under natural selection end up in finding these type of diagonal matrices. This is compared to animal communicative systems in [112]. Such systems present a non-degenerated mapping between objects and signals. The exhaustive vocabulary needs of this regime is illustrated in Figure 4.5**b**. This case would be favored in a scenario where few signals suffice for communication, and it would be restrained by the memory capacities of hearer and speaker. Indeed, it has been shown how memory constraints could prompt the development of a fully articulated human language when vocabulary size overcomes a certain threshold [228] so that units might be reused, but at the expense of making them ambiguous. In the least-effort framework proposed in [112], such a language would show up only at the phase transition $\lambda \sim \lambda_c$. Then, hearer and speaker's needs are equally taken into account, language instances with a moderate level of ambiguity are found, and communication is still possible – as the sharply varying mutual information between signals and objects around λ_c points out (Figure 4.5**a**).

In the original work [112] the phase transition reported was of second order, meaning that the shift from non-communicative codes to one-to-one mappings was a smooth drift across several intermediate steps – any of them could be a relatively fit candidate of human language, not so urgently needing to tune λ to its critical value λ_c. But further investigation of the problem clearly indicates that the transition is of first order in nature and that $\lambda_c = 0.5$ (for m=n), as Figure 4.5 clearly shows. This means that the jump between the two extreme cases happens swiftly at $\lambda = 0.5$, that a graduated range of possibilities that solve the optimization problem for $\lambda \sim \lambda_c$ does not exist, and that only at $\lambda = \lambda_c$ could we find a phenomenology akin to human language.

The analysis in [112] is complemented with an investigation of the frequency with which different signals show up for a given language in the model. This can be made thanks to equations 4.15 and 4.6. Remarkably, at the phase transition it was found that the frequency of different words obey Zipf's law [112], thus closing the circle with one of the observations that opened our quest.

This work [112] has been featured here for its historical importance in promoting the least-effort language agenda. However, its results have been contested and can not be held as correct anymore without a critical revision. The first and foremost claim has been that the algorithm employed in [112] usually only achieves local minimization, thus the portrayed languages would not correspond to global least-effort codes [251]. Furthermore, when analyzing the global optima of the problem we find ourselves with a degenerated solution – i.e. multiple assignments between objects and signals optimize the trade-off between speaker and hearer needs at the phase transition [115, 117, 251].

Three observations are pertinent about these critics: i) Among the several solutions to the least-effort problem at the indicated phase transitions we find Zipf's law as well [251]. This is not the dominating solution, though – i.e. there are more solutions with some other frequency distribution of signals than solutions whose signal usage follows equations 4.1 and 4.2 [251]. Thus we would expect that when choosing randomly among all least-effort solutions for $\lambda = \lambda_c$ we would likely arrive to some other distribution but to Zipf's. However, ii) the original investigation of least- effort communicative systems and the framework that this model introduces remain valid and very appealing, even if they do not suffice to produce Zipf's law. The least-effort principle has still got robust experimental and theoretical motivations, and we should not discard further forces operating upon language evolution that would select Zipf's law against others. In such a case, the least-effort game described in this section would be just a sub-problem that language evolution has solved over time. Finally, iii) concerning the main topic of this volume; even if Zipf's law were not recovered, robust evidence exists indicating that the trade-off

posed by the least-effort procedure is a way in of ambiguity into human language.

The featured model has been furthered by successive works. The hunt for a robust mechanism that generates the Zipf distribution continues and interesting proposals are being explored. A very promising one relies on the open-ended nature of human language [67]. Previous work by the same authors showed how Zipf's law is unavoidable in a series of stochastic systems. A key feature of those systems is that they grow by sampling an infinite number of states [66]. When applied to language, not only the unboundedness of human language is necessary but also the sempiternal least effort, so that Zipf's law can be successfully obtained for communicating systems. Interestingly, the approach in [67] applies the least-effort principle upon the transition between stages of the language as it grows in size – by incorporating new signals to its repertoire. This explicit role of the contingent historical path is an interesting lead absent in the main body of literature. A slightly different research line followed by these authors uses the proposed model to quantify precisely how much information is lost due to the ambiguity of optimal languages when the trade-offs discussed above are satisfied [122, 69].

Finally, several authors elaborate upon the model described above. In the critical review mentioned earlier [251] it is noted how the original model is not sufficient to always derive Zipf's law for the optimal model languages. The authors modify equation 4.21 and propose:

$$\Omega(\lambda)^0 \quad = \quad -\lambda I(S|R) + (1-\lambda)H(S) = -\lambda H(R) + \Omega(\lambda), \quad (4.26)$$

as a target for minimization; where $I(S|R)$ is the mutual information between signals and objects in the sets S and R respectively. This new target becomes eq. 4.21 if all objects are equally probable. Equation 4.26 is more adequate to “better account for subtle communication efforts” [251], as more costs implicit in equation 4.26 but absent in equation 4.21 are considered. In a follow up paper [265] it is demonstrated how an ingredient to robustly derive Zipf's law in their model is to take into account signal costs, which makes sense considering that different signals require different time, effort, or energy to be produced, broadcast, collected, and

interpreted. This, as we will see in the following section, can also be an important element for the presence of ambiguity in human languages.

4.1.6. Ambiguity, Principles of Information Theory, and Least Effort

Several recent empirical observations illustrate an optimization force – that justifies our least effort point of view – acting upon different linguistic facets such as prosody, syntax, phonology, and many others [165, 123, 164, 245, 196]. This evidence accumulates with other, previously shown global-level language organizational features epitomized by the properties of the small worlds (see Sections 4.1.3 and 4.1.4). All this indicates that optimization principles and natural selection should play a paramount role to understanding human communication in a broad sense. As we have seen, entropies arise or need to be explicitly introduced with a twofold purpose: as a metric and as a specific optimization target. The ubiquity of this mathematical construct – that, we recall, gives a measure of degeneracy and, more specifically in our context, of degeneracy of meanings – is a first clue that the price to pay for a least effort language is ambiguity, as we will argue right below again and as suggested by the results from Section 4.1.5.

In [246] a formalization of this trade-off between least-effort and ambiguity is presented. They argue that any optimal code will be ambiguous when examined out of context, provided the context offers redundant information; and they do so presenting extremely elegant, easy, and powerful information theoretical arguments that apply beyond human communication. Specially the first argument is of general validity for *any communicative system* within a context that is informative about a message. The two alternative – but similar – paths that the authors provide towards ambiguity are the following ones (the quotes are from [246]):

– *“Where context is informative about meaning, unambiguous language is partly redundant with the context and therefore inefficient.”*

  The authors conceive a space M consisting of all possible meanings

m such that inferring a precise meaning out of a signal demands at least

$$H[M] \;=\; -\sum_{m\in M} P(m) log\{P(m)\} \tag{4.27}$$

bits of information, with $P(m)$ the probability that meaning m needs to be recalled. Similarly, they assume a space C that encompasses all possible contexts c, compute the entropy of each meaning conditioned to happen within each context, and average over contexts:

$$H[M|C] \;=\; \sum_{c\in C} P(c) \sum_{m\in M} P(m|c) log\{P(m|c)\}. \tag{4.28}$$

This accounts for the average number of information (in bits) that a code needs to provide to tell apart different meanings within discriminative enough contexts. If context is informative it is likely that $H[M] > H[M|C]$ [246, 72].

With this in hand the authors have shown how “the least amount of information that a language can convey without being ambiguous is $H[M|C]$”, which is lower than $H[M]$; thus any optimal code will seem ambiguous when examined out of context and any unambiguous code will be suboptimal in that it produces more information than strictly necessary.

Note once more the elegance of the argument and its generality: no requirements are made about the meanings or the contexts, and the later are general enough as to include any circumstance of any kind affecting communication in any way.

– *“Ambiguity allows the re-use of words and sounds which are more easily produced or understood.”*

This second argument only diminishes in generality because the authors must consider that different signals in a code vary in cost – i.e. that they are not of equal length or complexity, or that distinct signs

require different amount of effort when they are used. This becomes obvious in human speech, e.g., considering the longer time that larger words demand. Note anyway that this is a quite general scenario still affecting most conceivable communicative systems and, of course, any kind of human communication.

The argument acknowledges that it is preferable to use simpler signals. Then, ambiguity enables us to re-use the same signal in different contexts, assuming always that the context provides the needed disambiguation.

According to these ideas, that optimal codes are ambiguous if the context is informative does not imply that human languages must be ambiguous, neither that any ambiguous coding is more optimal than any unambiguous one. However, ambiguity – say polysemy, in certain contexts, but not only – is an extremely extended phenomenon in human language when tongues are analyzed out of context, and the authors propose that such simple yet forceful reasoning explains its pervasiveness. In previous sections a much stronger point was made based on empirical observations: this polysemy not only does exist, but it also shapes the structure of tongues such that a global order arises in many aspects of it (e.g. semantic networks), and such that it presents very convenient features that render human language optimal or very effective (e.g. for semantic navigation). Thus not only ambiguity is present, it seems to be of a very precise kind in order to comply with several optimization needs at a same time, such as Zipf’s least effort paradigm proposed [351].

4.1.7. Discussion and Prospects

The models and real networks presented above provide a well-defined theoretical and quantitative framework to address language structure and its evolution. The sharp transition between non-communicative and communicative phases is a remarkable finding – and the fact that intuitive models can reproduce this feature is impressive. This suggests that a fundamental property associated to the least effort minimization principle

involves an inevitable gap to be found among its solutions. From another perspective, both real language networks and the simple graphs emerging from the least effort algorithm(s) introduce ambiguity as a natural outcome of their heterogeneous nature.

While the path explored this far invites us to be optimistic, several open problems arise from the results reviewed. These will require further research until a complete picture of human language – beyond the role of ambiguity – is attained. Here is a tentative list of open issues:

1. Both the topological analysis of semantic networks and what can be proposed from simple models are typically disconnected from an explicit cognitive substrate. Some remarkable works on semantic webs have shown that the structure of semantic webs includes a modular organization where groups of semantically related words are more connected among them than with other items. Individuals mentally searching on this space seem to make fast associations between items within modules as well as seemingly random jumps between modules [131]. The pattern of search is actually related to the ways search is performed on computer networks. Moreover, there is a literature on neural network models of semantic association [200, 159, 160] that could be explored in order to see how the space of neural attractors and the underlying categorization emerging from them are linked to a semantic network. Models of damage in semantic webs (using topological methods) already suggest that relevant information might be obtained in relation with the process of cognitive decay associated to some neurodegenerative diseases [53, 37].

2. A very promising field within language evolution involves using embodied agents (robots or physical simulations of them) that are capable of learning, memory, and association [304]. A protogrammar has been shown to emerge in these embodied communicating agents [303, 306, 30]. The study of lexical and grammatical processing in these robotic agents using so called Fluid Construction Grammars (FCGs) [307, 305] reveals that language evolution

might take place by optimizing lexicon size and the construction structures in order to minimize search. More traditional approaches to computer languages – as in programming languages – explicitly reject ambiguity for the challenges it presents. It is made clear that FCGs seek more malleable structures (thus the *Fluid*), ready to evolve and be adopted and adapted by a population – in this case, of robots. The population is usually not expected to share the exact same grammatic structure as it emerges, thus clearing a path for ambiguity. Notwithstanding this, part of the problems solved by this novel approach is one of reducing ambiguity out of the messages being interchanged by the talking agents [308]. Also, the emergence of grammatical rules is a direct consequence of this ambiguity reduction [30].

3. In all studies so far developed, models of language evolution involve only one type of network level of description. However, semantic, syntactic and even phonologic levels interact and any relevant analysis should include several network levels. How are different networks connected to each other? What is the impact of their special topological and scaling properties on the global behavior of language as a whole?

4. Statistical physics is at the core of many of the approximations considered in this chapter. Despite the biological nature of language and its historical origins, we have seen that some strong regularities are inevitable and are more fundamental than we would expect. There are many other ways of approaching language structure using these methods, including the analysis of language ontogeny [64, 13] and the structure of syntactic networks. Available evidence from data and models suggests that, once Zipf’s law is at work, a number of well known regularities exhibited by syntax graphs are obtained [116]. This would be consistent with an evolutionary scenario where syntax might come for free, as a byproduct of possibly inevitable features of correlations among words following Zipf’s law [296]. The idea is appealing and worth researching and, once

again, complex networks and information theory might provide a valid framework.

5. A twin problem to that of ambiguity is revealed when we consider synonymy. This trait might be a contingency, and it is considered rare by scholars [229]. Indeed, while different models account for it [229, 112, 265], all of them predict that synonymy should not be present in optimal languages or languages in equilibrium; but yet we observe some degree of synonymy in every human code.

4.2. Exploring the morphospace of communication codes

Now we turn our attention back to the least-effort language agenda reviewed in section 4.1.5. Several contributions to this field in the last decade are inspired by the minimal model proposed by Ferrer i Cancho and Solé [112] in which a matrix A encodes whether a signal $s_i \in S$ in a language names a given object $r_j \in R$ ($a_{ij} = 1$) or not ($a_{ij} = 0$) as introduced above (see figures 4.4**b** and 4.5**c** and **d**, as well as equation 4.14).

These matrices allow us to compute a series of characteristics about the language that they represent, such as the average ambiguity perceived by a hearer ($H_m(R|S)$, equation 4.20) or the average number of bits that a speaker must browse in finding the appropriate name for a given object ($H_n(S)$, equation 4.17). The simultaneous minimization of these quantities (which would account for efficient languages that are easy to use for both hearers and speakers) poses a MOO problem that has historically been wrapped into a single global target ($\Omega(\lambda)$ or $\Omega(\lambda)^0$) through equations 4.21 or 4.26 [112, 251, 265, 301, 286]

The global optima of equation 4.21 have been found as a function of the external parameter λ. They turn out to trace a first order phase transition between the one-to-one mapping, with each signal naming one and only one object, and the most degenerate code in which one word names

every object in S (figure 4.5**a** and **c**). The matrix A defines a bipartite network with nodes in S and R, hence we may refer to these extreme languages as the one-to-one and the star graphs or codes indistinctly (figure 4.5**c** and **d**).

It is often argued that human-like communication systems would lay precisely at this transition, while the extreme phases would correspond to animal communication (one-to-one mapping) or to a code in which communication fully depends on context (star) [115]. According to what we have seen in chapter 2, first order phase transitions appear due to cavities or straight stretches in the front – which implies a critical point. In section 4.2.1 we analyze the Pareto front of this problem and see how it fits together with previous findings about efficient communication.

The model from [112] defines up to $2^{n\times m}$ different matrices where $n \equiv |S|$ is the number of signals available and $m = |R|$ is the number of objects that are named by the code. Notwithstanding the interest of the Pareto front, this defines a huge morphospace that might be worth exploring. We do so in a systematic way in sections 4.2.2 and 4.2.3. In section 4.2.3 we report a series of interesting measures about the codes found across the morphospace, including those along the Pareto front. In section 4.2.4 we perform both principal component and clustering analysis to find out what salient aspects of codes tend to happen together. Also in section 4.2.4 an attempt is made to locate real languages in the morphospace. This analysis suggests a key role for grammatical particles in opening up the space of possibilities to human language.

4.2.1. The criticality of least effort communication

Prokopenko et al. [251, 265] computed analytically the global minimizers of equation 4.21. These turn out to be all matrices A that do not contain synonyms when interpreted as a communication code – i.e. those matrices that only have a 1 in each column. For those codes, using some

algebra we come to the next expressions for the target functions:

$$H_m(R|S) = log_m(n) \sum_{i=1}^{n} \frac{\rho_i}{m} log_n(\rho_i), \tag{4.29}$$

$$H_n(S) = log_n(m) - \sum_{i=1}^{n} \frac{\rho_i}{m} log_n(\rho_i), \tag{4.30}$$

$$H_n(S) = log_n(m) - \frac{1}{log_m(n)} H_m(R|S); \tag{4.31}$$

where ρ_i is the number of objects associated to the i-th signal.

Equation 4.31 implies that all Pareto optimal codes lay along a straight line in target space – if $n = m$ this is the line $y = 1 - x$ with slope -1 (figure 4.6**a**). In other words, the Pareto front Π_{LE} of the least-effort communication problem is given by:

$$\Pi_{LE} = \left\{ A = (a_{ij}) | H_n(S) = log_n(m) - \frac{H_m(R|S)}{log_m(n)} \right\}. \tag{4.32}$$

The one-to-one and the star codes lay at the extremes of Π_{LE}, which in target space lay at $(0, 1)$ and $(1, 0)$ respectively (assuming $n = m$ again). According to chapter 2 this implies a first order phase transition and a critical point at $\lambda^c = 1/2$, at which the global optimum is hugely degenerated because all the Pareto optima become available at once – i.e. as it happened with critical graphs in chapter 3, critical codes at λ^c are neutrally optimal with respect to the global energy (equation 4.21).

The first order transition has already been noted in the literature thanks to the marked discontinuity in the order parameters. The possibility that this transition could be critical has been informally mentioned owing to the presence, only at the transition point, of codes whose rank distribution of signals follows a power law with exponent -1. As discussed above, this distribution is known as Zipf's law and is a paramount, universal trait of human languages. This evidence of criticality is only circumstantial because processes shall exist, not necessarily critical, that generate Zipf's distribution. Now, using the definitions of section 2.3 in chapter 2 or the MaxEnt methods from section 3.2, we can further argue that least-effort communication codes generate a critical ensemble.

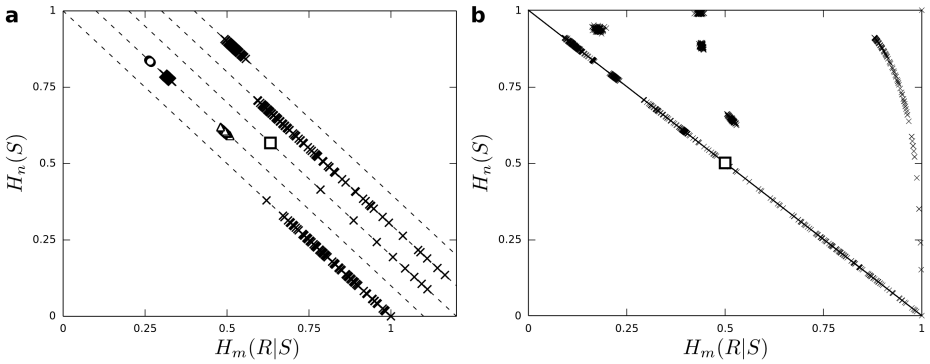


Figure 4.6: **Seeding the morphospace of communication codes.** **a** Pareto optimal codes generated randomly following different prescriptions always fall in the $y = 1 - x$ line. The different process from left to right (components along the x axis have a 0.1 offset) are: i) one dice process, ii) an iterated dice process, iii) power laws with different exponent, iv) preferential attachment, and v) uniform association between objects and signals. **b** After seeding the morphospace with a series of random processes, it remains almost empty. It was necessary to apply an evolutionary algorithm to achieve a uniform sampling of the morphospace.

Since we know the condition that Pareto optimal codes fulfill, we can use different processes to generate them and see where they lay along the Pareto front. A good reference point is that code that follows Zipf's law (square in figure 4.6**a** and **b**). Since Pareto optimal codes do not have synonyms, the frequency with which each signal shows up within our model is proportional to the number of objects that it names. Hence, in this model the most degenerate signal is associated to N_1 different objects while the i-th most degenerated one is associated to N_1/i. Fortuny and Corominas-Murtra have shown that such a language emerges for $H_m(R|S) = 1/2 = H_n(S)$, right at the middle of the Pareto front [122].

With this in mind we generated more Pareto optimal languages following a series of procedures, some of which are known to induce power-law distributions. The first one is described in [70] and it consists in

throwing a dice to allocate objects to signals. Start with $N_0 = m$ objects that must be allocated. Throw an N_0-sized dice to obtain a random number $N_1 < N_0$. The first signal in the language is assigned $N_0 - N_1 + 1$ objects. Then throw an $(N_1 - 1)$-sized dice to generate the number of objects associated to the second signal, and so on. Sooner or later a 1 shows up (either randomly or because the dice became so small). At this point, all objects have been allocated to some signal. Pareto optimal codes produced with this process (presented in the left-most diagonal of figure 4.6**a**) appear mostly along the inferior half of the Pareto front. Simple as it might seem, it has been proved that this process can generate Zipf's distribution [70], especially when the previous process is iterated many times – i.e. if after exhausting the size of the dice, a new one with size N_0 starts the process anew so that eventually we end up with $m = N_0 N_{it}$ objects, with N_{it} the number of iterations. The second diagonal in figure 4.6**a** contains codes generated by an iterated dice with different choices of N_{it} and N_0. They present a lower spread (once N_{it} and N_0 are fixed), and tend to appear at the upper half of the front.

Codes with an explicit power law distribution and different exponents are plotted on the third diagonal in figure 4.6**a**. The exponent increases from $\gamma = 1$ to $\gamma = 3$ as we approach the star graph. These codes show mostly in the lower half of the front. Preferential attachment is another process known to generate power law distributions. Codes generated in this way (fourth diagonal) concentrate towards the middle of the front and extend towards the lower half. Finally, objects were assigned to signals with a uniform distribution (fifth diagonal). These tend to locate themselves closer the one-to-one mapping.

4.2.2. Sampling the morphospace

The star-like language sets the upper boundary of $H_m(R|S)$ and the one-to-one mapping does the same for $H_n(S)$. It is easy to see that a matrix filled with ones (that we shall call a block code) is mapped onto $(1, 1)$ in the $H_m(R|S) - H_n(S)$ plane. In between these three singular solutions lays a triangle containing every possible communication code

with $n \leq m$. We treat this subset of the $H_m(R|S)$ – $H_n(S)$ plane as a morphospace.

For a modest $n = 200$ and $m = 200$ there are $2^{nm} = 2^{40000}$ possible codes. We would like to sample them fairly so that the morphospace is evenly represented. A first strategy is to wire up signals to objects randomly with a given probability p. If we do so, most codes appear clustered in small clouds in the upper part of the morphospace. The same thing happens if we induce small mutations on the one-to-one or the star mappings. All these random initializations of codes resulted in a very poor exploration of the region available (figure 4.6**b**).

To further sample the interior of the morphospace we used an evolutionary algorithm. We split the target space in a 30×30 grid and seeded it with the codes shown in figure 4.6**a**. We detected squares with low occupation and applied mutation and crossover between codes from occupied squares nearby. Codes from the most occupied areas were randomly removed. We iterated this process until every square in the upper left part of the target space approached the average per-square occupation (~ 20 codes). Then we measured a series of characteristics of each code (detailed and discussed in the next section) and we averaged them over codes within each square.

The costly algorithms involved and the memory demands of storing the code matrices limited our experiments to $n = m = 200$. For Pareto optimal languages a more compact notation is possible (we just need the number of objects associated to each signal) so we produced extra Pareto optimal codes with $n = m \sim 1000$ to further assess the morphology of codes along the front.

4.2.3. The morphospace of communication codes.

Although a large part of the literature on language evolution modeling considers bipartite, signal-object (word-meaning) associations, no systematic effort has been made to characterize the available landscape of communication codes. By considering the morphospace of matrices A, we aim at providing a quantitative description of its properties. As will

be shown below, this morphospace turns out to be a far from trivial or monotonous landscape.

Characterizing the vocabulary

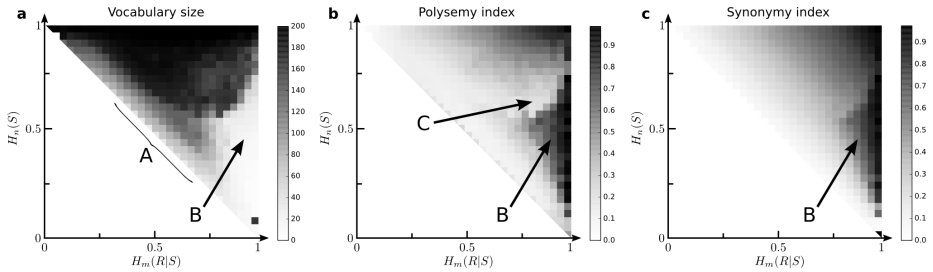


Figure 4.7: **Vocabulary size, polysemy, and synonymy over the morphospace.** **a** Vocabulary size represented the number of languages used among all those available and it is only low near the star graph (in a prominent area labeled B) and along the Pareto front. Most of the morphospace presents large vocabularies. **b** Polysemy is large in the B region and as we complete the matrix A towards the block code. **c** Synonymy increases uniformly as we move apart from the front except for codes within B. This locates them far away from the Pareto optimality condition (zero synonymy).

First of all we measure the effective vocabulary size of the codes (L, figure 4.7**a**). While we force that every object must be associated to some signal, the opposite is not true: there might be signals that are actually not used (e.g. the star code consists effectively of just one signal, $L = 1$). By plotting L across the morphospace a non-trivial structure is revealed. Codes with little number of signals occur mostly near the star and in a narrow region adjacent to the Pareto front (marked A in the figure). Far apart from the front there is yet another region (marked B) with less than 30% of all available signals being used. The transition to codes that use

more than 50% of available signals (central, darker region in figure 4.7**a**) seems to be abrupt wherever we approach those codes from.

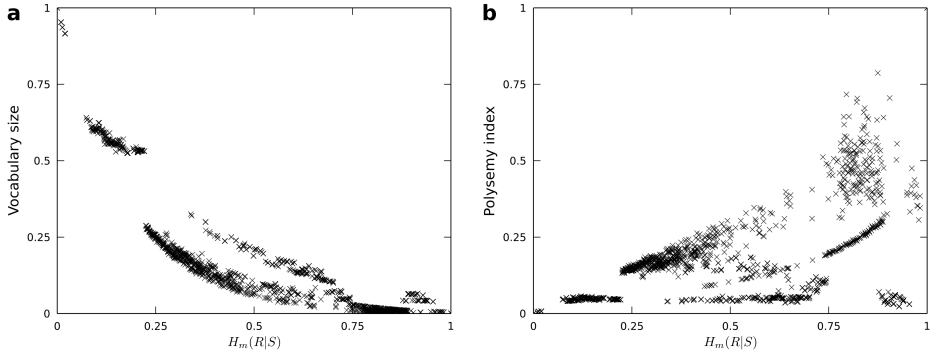


Figure 4.8: **Vocabulary size, polysemy, and synonymy along the Pareto front.** **a** Codes along the Pareto front keep a relatively low vocabulary except close to the one-to-one mapping. Also, two branches seem noticeable around the middle of the front, suggesting that similar Pareto optimal values of $H_m(R|S)$ and of $H_n(S)$ can be achieved with differently wired codes. **b** A reduced vocabulary size does not result in a strictly monotonous increase of polysemy as we approach the star code. Instead, languages with similar $H_m(R|S)$ may present different polysemy levels. The range available grows as we approach the maximally ambiguous code.

Codes along the Pareto front keep a relatively low vocabulary when compared with similar languages just a little bit into the front (figure 4.8**a**). In the more central part of the front ($0.25 < H_m(R|S) < 0.75$) two branches seem visible. This indicates that similar values of both targets ($H_m(R|S)$ and $H_n(S)$) can be reached with distinctly structured Pareto optimal codes. This will be confirmed when analyzing other characteristics of the codes.

It is important to take into account the effective vocabulary size when measuring certain properties. As an instance, we introduce the next poly-

semy index I_P and synonymy index I_P:

$$\begin{aligned} I_P &= \sum_{s_i \in S} \frac{log_m(\rho_i)}{L}, \\ I_S &= \sum_{r_j \in R} \frac{log_L(\sigma_j)}{m}; \end{aligned} \tag{4.33}$$

where σ_j is the number of signals associated to object r_j and ρ_i is the number of objects associated to signal s_i. These indexes measure the average logarithm of σ_j and ρ_i respectively – i.e. the average number of bits needed to decode an object given a signal (I_P) and the averaged degeneracy of choices to name a given object (I_S).

We appreciate that the region B, which presented a very low vocabulary size, is mostly made of polysemic signals (figure 4.7**b**). Right next to it I_P drops suddenly (marked C in figure 4.7**b**) and then increases steadily as we tend towards the block language, in which every signal names every object. The region B starts close to the star and is also associated to a large I_S (figure 4.7**c**). This implies that synonymy increases sharply as we move away from the Pareto front if departing from the star code. Moving away from any other part of the front, I_S builds up more gradually. Keep in mind that the condition for Pareto optimality is that codes do not have any synonyms. This indicates that Pareto optimality degrades almost uniformly anywhere but near the star. This might have evolutionary implications: that part of the morphospace might be difficult to reach if Pareto selective forces are at play. Note that languages in B require more contextual information to be disambiguated – that shall be a handicap for communication systems.

I_S is trivially zero along the front (not shown). We could expect that polysemy would build up as we approach the star code. Instead we see that at each value of $H_m(R|S)$ there are very different codes showing a range of polysemy. The maximum of this range does grow with $H_m(R|S)$. The fact that similar Pareto optimal codes present such diverse I_P again suggests a great diversity within critical communication codes.

Network structure

As noted at the beginning of section 4.2, each code embodies a bipartite network (figures 4.4**b** and 4.5**c** and **d**) that we refer to as the *code graph*. We can derive two more networks from each code: one named R-graph in which objects $r_j, r_{j'} \in R$ are connected if they are associated to one same signal, and another one named S-graph in which signals $s_i, s_{i'} \in S$ are connected if they are synonymous. Because Pareto optimal codes do not have synonymous signals, their code graphs consists of disconnected components in which the i-th signal binds together ρ_i objects. Accodingly, each Pareto optimal R-graph is a set of independent fully connected clusters and S-graphs are isolated nodes.

We extracted the set of connected components $C = \{C_i, i = 1, \ldots, N_C\}$ for each network. The size $||C_1||$ of the largest connected component along the Pareto front corresponds to the number of different meanings of the most degenerate signal (figure 4.9**a**), which is precisely m for the star and drops fairly monotonously along the front until the one-to-one mapping is reached. For intermediate values of $H_m(R|S)$ we observe certain degeneracy of $||C_1||$, again associated to a variety of similarly Pareto optimal codes. The largest connected component in R-graphs $||C_1^R||$ (figure 4.9**b**) is trivially connected to that of the bipartite network, while that of S-graphs is just one node (in figure figure 4.9**c** $||C_1^S||$ is plotted as a fraction of the vocabulary size L, hence it grows with $H_m(R|S)$).

If we keep track of $f(||C_i||)$ (the frequency with which connected component sizes show up in a graph) we get:

$$H_C = -\frac{1}{||C||}\sum_{i=1}^{N_C} f(||C_i||) log(f(||C_i||)), \tag{4.34}$$

the entropy of the distribution of connected component sizes. This is a first measure related to the internal complexity of each code and we find no clear trend for Pareto optimal codes (figure 4.9**d**), again suggesting a huge diversity among them. H_C^R and H_C^S (not shown) turn out to be similar to that of the bipartite network.

As we move apart from the front code graphs get a large connected component very quickly (figure 4.10**a**). Around the region B objects

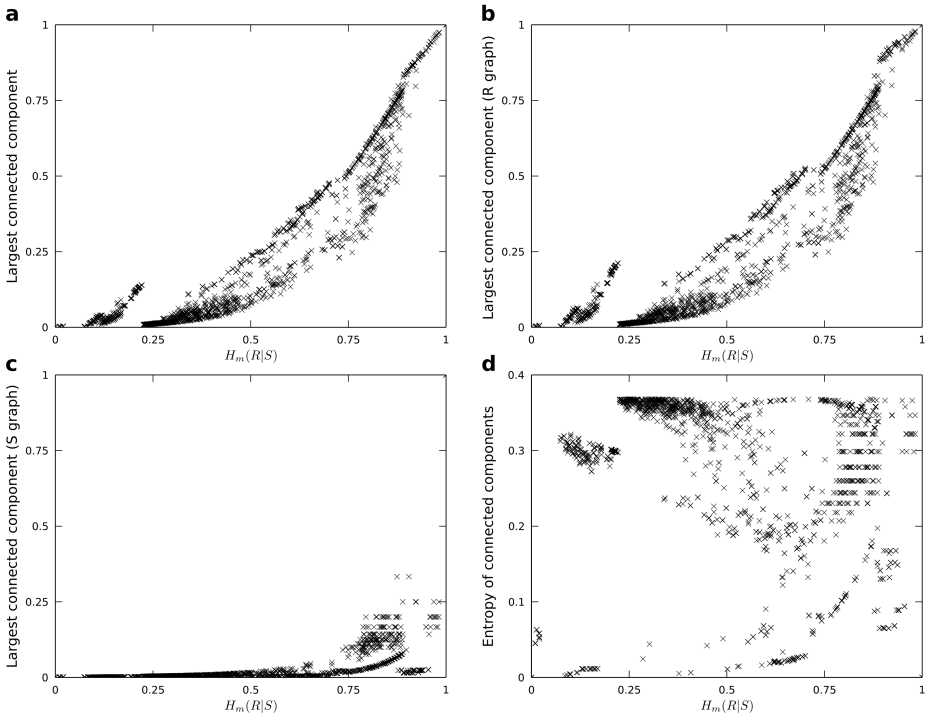


Figure 4.9: **Studying network connectivity along the Pareto front.** **a** The largest connected component seems to grow steadily as we proceed to the star code due to the ambiguity building up around a few degenerated signals. **b** This results in a similar plot for $\|C_1^R\|$. **c** In Pareto optimal S-graphs every connected component has got one node. Here we plot this relative to the number of clusters, which turns out to be L. Hence **d** represents $1/L$, the inverse of vocabulary size. **d** H_C does not reveal a clear pattern. This highlights again both the diversity of the Pareto front and that similar values of equally Pareto optimal target functions can be obtained with fairly different configurations of codes.

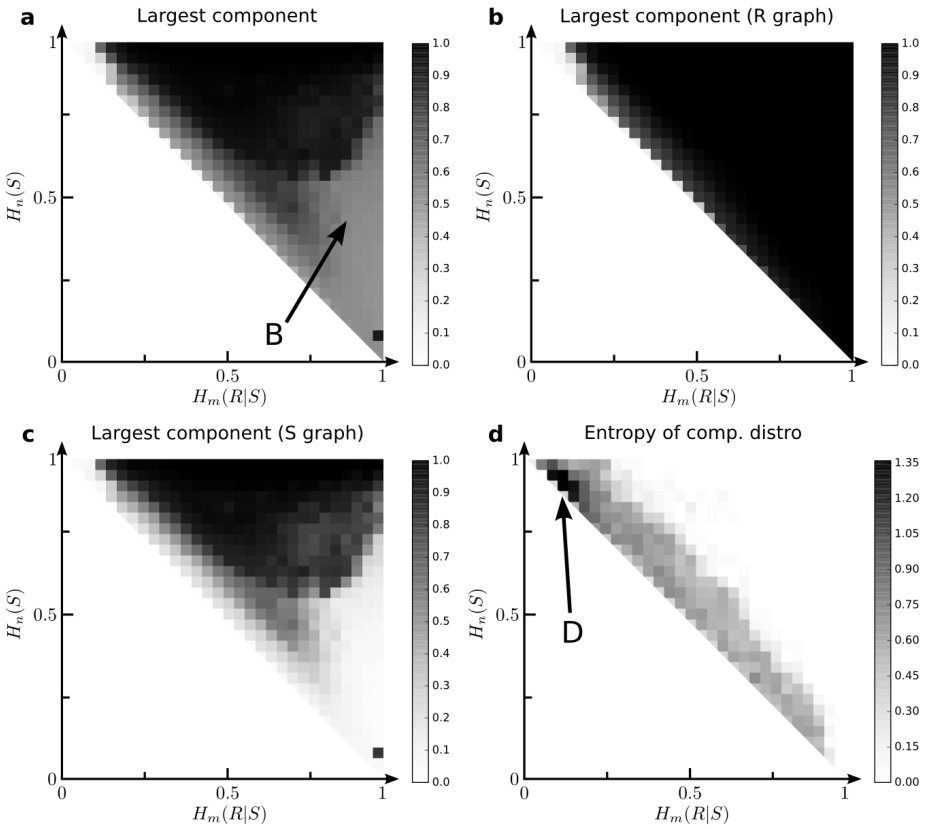


Figure 4.10: **Studying network connectivity over the morphospace.** Most of the morphospace presents well connected networks except along the Pareto front, where the Pareto optimality condition imposes that the code splits into unconnected clusters. The region B also presents smaller connected components. This seems to be the case because those codes do not use most of the signals available. **b** The R-graph does present large connected components in B, while the S-graph is poorly connected again **c**. **d** H_C peaks near the one-to-one mapping and a stripe with large entropy that runs parallel to the front. It vanishes in the interior of the morphospace.

get connected (figure 4.10**b**) while signals remain sparsely linked (figure 4.10**c**), which is consistent with the small vocabulary size and large polysemy there. The entropy of the distribution of connected components is trivially low for well connected networks (figure 4.10**d**). In the transition towards fully connected code graphs (a stripe parallel to the Pareto front) H_C gets large as synonyms start to connect different clusters in a random manner – note the similarity of this process with percolation. H_C peaks near the one-to-one mapping (marked D in figure 4.10**d**). This suggests that, in order to get a varied distribution of cluster sizes, it is better to start off with small clusters (the extreme case being the one-to-one mapping) and let synonyms randomly add up structure, instead of departing from an already diverse collection of isolated clusters (such as those found along the Pareto front). Again, H_C^R and H_C^S (not shown) are similar to H_C

Complexity from codes as a semantic network

Memory and conversation often evolve thanks to spontaneous associations. Words, concepts, and objects in the real world constitute an abstract semantic web whose structure shall be imprinted into (or stem from) our brains [159, 160]. It is often speculated that semantic networks must be easy to navigate, which in turn is related to the presence of a small-world underlying structure [313, 301].

Let us consider our codes as a toy generative model. Starting with an arbitrary signal or object we implement a random walk moving into adjacent objects or signals respectively. This generates strings of characters of arbitrary length associated to elements of the sets R or S. The network structure shall condition the frequency $f(r_j)$ and $f(s_i)$ with which different objects r_j and signals s_i are visited. From these frequencies we get:

$$\begin{aligned} H_R &= -\sum_{j=1}^{m} f(r_j) log_m(f(r_j)), \\ H_S &= -\sum_{i=1}^{n} f(s_i) log_L(f(s_i)). \end{aligned} \tag{4.35}$$

These entropies will be large if R or S are evenly sampled, and they should present low values if a bias towards certain objects or signals exists. Hence in this case low entropy is a measure of non-trivial structure arising form our toy generative model. We also recorded 2-grams (couples of objects or signals that happened one after the other during the random walk) and computed the corresponding entropies H_{2R} and H_{2S}.

An extreme case happens if we are sampling from a very small connected component. A code's bipartite graph might have several large connected components and we might end up in a small one because of an unlucky initial condition. This would result in low H_R and H_S despite the remaining structure of the network. To avoid this situation we imposed that our generative model would implement a random jump when an object was repeated twice since the last random jump. This way we could avoid getting trapped in small connected components all together. (We also interrupted the random walk when signals, instead of objects, were repeated – the results were largely the same.)

Note an interesting effect of this mechanism at the two extremes of the Pareto front. In star codes, one signal is always picked and objects just sample R randomly, hence H_R is maximal (figure 4.11**a**) and the structure over R is minimal. The structure over S seems maximal (minimum H_S, figure 4.11**b**) because there is only one signal in this code. In the one-to-one mapping, a signal is picked up randomly and a same object is always sampled twice in a row, so a disruption happens at every step of the random walk. This results in an equally large H_R, again indicating little structure, and also a large H_S. In the middle of the Pareto front, where clusters of different size exist, we expect to observe relevant biases in R resulting in a drop in H_R. This does not happen in a uniform manner (figure 4.11**a**): consistently with previous findings we see diverse values of H_R for a given $H_m(R|S)$. The codes that impose a less trivial structure in our generative model (i.e. those with lower H_R) happen surprisingly close to the star graph. H_S drops as well, but due to a diminishing vocabulary size, so its variation is more consistent (figure 4.11**b**). H_{2R} presents a challenging structure (figure 4.11**c**). It is large again near the star and the one-to-one codes, and the nets with lower H_{2R} are the same as for H_R.

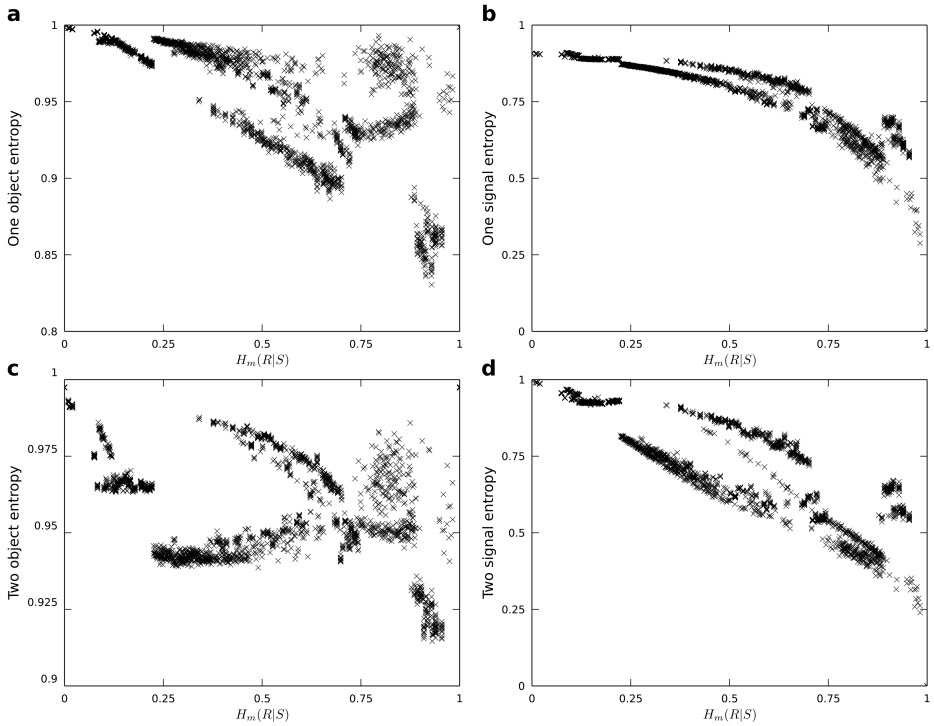


Figure 4.11: **Complexity of random walks in language graphs along the Pareto front.** **a** The random walks were designed so that both extremes of the front present the largest entropy possible. (The point corresponding to H_R = 1 for the star graph is hard to see.) For intermediate values of $H_m(R|S)$, codes exist that present very different H_R. The lowest values happen very close to the star code. **b** H_S decays naturally as the vocabulary size diminishes. **c** Low H_{2R} for codes near the star graph is reproduced, as well as the maximums at the extremes of the front. The rest of the codes seem to cluster together around characteristic entropy values, but a clear trend cannot be appreciated. **d** While the 2-signal entropy still decays as the vocabulary size diminishes, different kinds of languages seem to be appreciated in the intermediate regions of the Pareto front. This results in different H_{2S} for a same $H_m(R|S)$. This also indicates that H_{2S} captures some structure in Pareto optimal codes that H_S misses – we will see that this is not the case in the rest of the Pareto front.

Other intermediate codes tend to cluster together around different entropy values in a quite irregular fashion across $H_m(R|S)$ values. A clear trend does not exist. While a trend is observed in H_{2S}, this is not as clear as the one for H_S, again showing the great variability of codes within the front (figure 4.11**d**).

Extending the analysis to the rest of the morphospace we appreciate two regions (E and F) in which H_R drops (figure 4.12**a**). The underlying code graphs around E and F must have got some canalizing properties that break the the symmetry between objects. H_S also drops, but in a region (G) at the frontier between E and F, as indicated by the level curves in figure 4.12**b**. This mismatch between low H_R and H_S reveals that certain structures of the bipartite networks bias the sampling towards certain signals while keeping the sampling of R fairly regular while some other bipartite structures can do the opposite. A good question is whether some bipartite graphs in E or F may correspond to graphs in G with objects and signals interchanged. We must also consider the possibility that the bias in the sampling of S arises because only a few signals are used (as it happens to codes near the star). However, codes in G have fairly large vocabularies.

H_{2R} (figure 4.12**c**) reveals a stripe with non trivial structure that runs parallel to the front (similarly to the region with large H_C, just a little bit more towards the interior of the morphospace). This stripe includes region E. H_{2S} (figure 4.12**d**) shows values similar to those of H_S suggesting that the later already captures all the relevant structure that random walks induce over H_S.

Zipf and other power laws

Zipf’s law is one of the most notable statistical patterns in human language [351]. Despite important efforts [66, 67], the reasons why natural language should converge towards this distribution of word frequency are far from definitive. Detailed research of diverse written corpora suggest that under certain circumstances (e.g. learning children, military jargon, diseased speakers) the distribution of words presents a power-law distri-

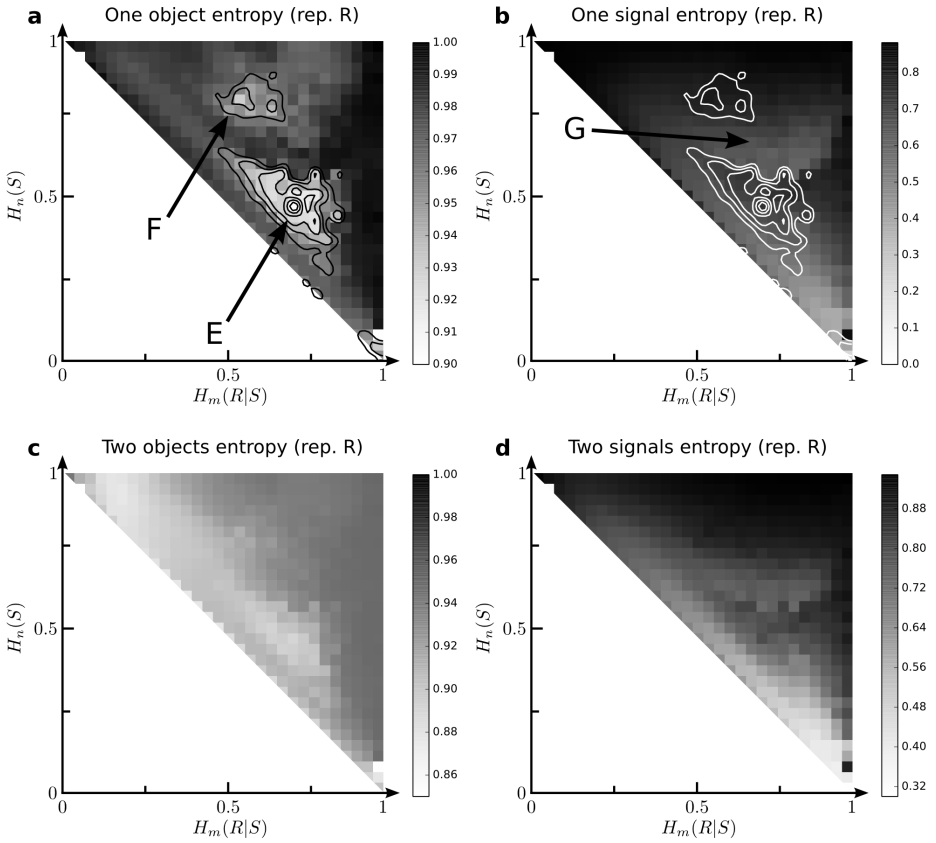


Figure 4.12: **Complexity of random walks in language graphs over the morphospace.** **a** The entropy of single object frequency in random walks presents a non-trivial structure within the morphospace. Two different regions (containing well differentiated A matrices) are capable of breaking the symmetry in the random-walk sampling of objects. **b** Codes in yet another region do the proper to the sampling of signals, but this area does not correspond to any of the previous ones. **c** The entropy of 2-grams is low in a region that correlates roughly with a large entropy of the connected component distribution (H_C). **d** The entropy of 2-grams in signals largely reproduces that of single signals, suggesting that there is not any new structure that H_{2S} can uncover.

bution with a generalized exponent [114, 13]. In the original account of the least- effort model introduced in section 4.1.5 it was found that Zipf's law laid just at the transition point between the star and one-to-one codes [112]. While this is true (it is possible to find Zipf's law just at that transition), it was later shown that Zipf's law is not the most frequent code under those circumstances [251, 265]. We corroborated this in section 4.2.1.

The Pareto front (which consists of all solutions available at the critical point) presents very different solutions including (discrete versions of) power-laws with arbitrary exponent (figure 4.6). Considering all solutions along the front (not only those manually generated power laws), we applied the Kolmogorov-Smirnov (KS) goodness of fit test to check whether the frequency of signals followed Zipf's law (figure 4.13**a**). (To compute the frequency of signals we returned to the original model by Ferrer i Cancho and Solé [112] that assumes that objects need to be recalled with the same frequency.) We find a nadir of low KS score around the expected region (near $H_m(R|S) = 1/2 = H_n(S)$ according to [122] and in agreement with our manually crafted power laws from figure 4.6). But we also find that a similar value of $H_m(R|S)$ can be achieved with codes that do not reproduce correctly Zipf's law. We also tested if alternative power law distributions were a good fit for the frequency of signals. We used the standard test described in [57]. It was found a low KS-score in the expected region (lower half of the Pareto front in figure 4.6, which corresponds to $H_m(R|S) > 0.5$ in figure 4.13**b**). Again, codes exist with bad agreement with power-law distribution and yet similar $H_m(R|S)$ values. Surprisingly, the methods in [57] render quite regular values of the best exponent (figure 4.13**c**), which seem to cluster along two branches as a function of $H_m(R|S)$.

We wonder whether Zipf's law can be found anywhere else in the morphospace. The KS-test suggests a broad region in the middle upper part of the target space (figure 4.14**a**). The goodness of fit to a generic power law (figure 4.14**b**) renders a region adjacent to that with good fit in the Pareto front (along its lower half). There is a large region in which Zipf's law is a better fit than other power-law, even if the methods in [57]

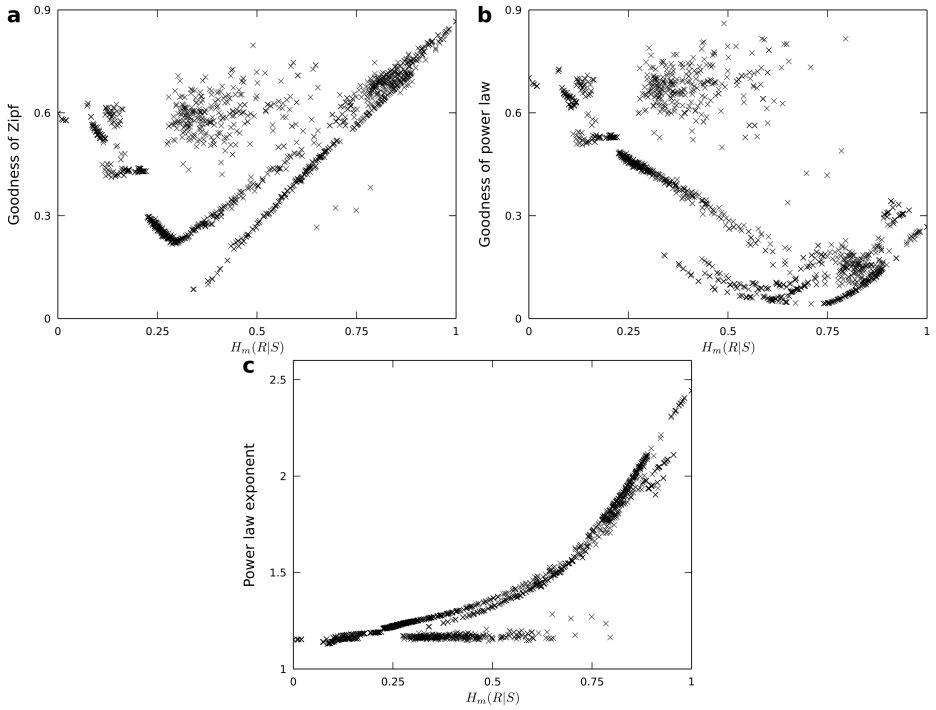


Figure 4.13: **Goodness of power laws in describing the frequency of signal use along the Pareto front.** **a** The goodness-of-Zipf test (which is better if the KS-score is lower) has its lowers value near $H_m(R|S) \sim 0.5$, where we have seeded actual Zipf distributions. Discrepancies shall be attributed to the numerical methods employed. **b** A broad region with $H_m(R|S) \geq 0.5$ presents good fit to power laws. This corresponds to the parts of the Pareto front where actual power laws, preferential attachment, and the dice have been seeded. **c** The exponents around these regions vary between $\gamma \sim 1.1$ and $\gamma \sim 2.5$, with codes seemingly belonging to one of two well differentiated branches.

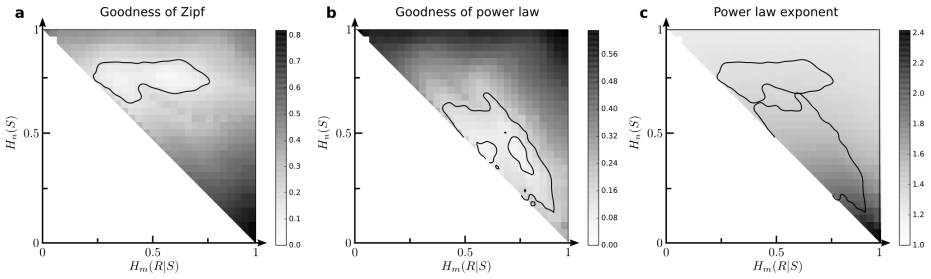


Figure 4.14: **Goodness of power laws in describing the frequency of signal use across the morphospace.** **a** The Kolmogorov-Smirnov test reveals a region of the morphospace in which Zipf's law accounts fairly enough for the frequency of signal use. **b** Power law distributions with arbitrary exponent render a better fit in a broad region near the lower half of the front, where good fits to power laws are also observed. **c** The preferred exponent raises from ~ 1.2 to ~ 2.5 as we approach the star code, but most of these exponents do not correspond to good fits.

suggest an alternative (though close to 1) exponent.

4.2.4. Code archetypes and real languages

In the previous section we have made several different measurements over the matrices A that represent communication codes in our least-effort language model. Both codes along the Pareto front and over the morphospace have been characterized. Thanks to these measurements we can search for archetypal codes that present similarities among them.

We performed a principal component analysis both over the whole morphospace and over the Pareto front alone. The later required 4 principal components to explain more than 90% of the variance of the data. The morphospace required 5 principal for a similar performance. Using all principal components we run several times a k-means algorithm [190]. Using the data of the morphospace we converged consistently into the same five clusters shown in figure 4.15. These five code archetypes

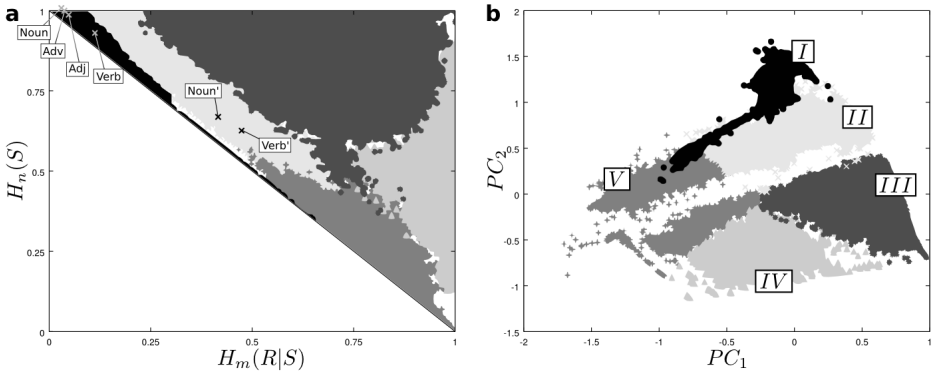


Figure 4.15: **Clustering of languages across the morphospace.** k-means clustering using all principal components reveals a consistent structure in the morphospace. Five cluster are shown here. Real languages fall within cluster I, close to the one-to-one mapping proper of animal communication systems. The real matrices are marked: *Adj* for the adjectives, *Adv* for the adverbs, *Noun* for the nouns, and *Verb* for the verb. If certain grammatical words are included (named with an apostrophe: *Noun'* for nouns and *Verb'* for verbs) they move into cluster II and towards the center of the morphospace, relatively close to the Pareto front. **b** All clusters get further segregated in two principal component space. This space appears interrupted by a stripe along which no codes exist.

correspond roughly to:

I Codes close to the one-to-one mapping and the upper two thirds of the Pareto front. It includes those graphs with the largest H_C values (figure 4.10**d**).

II Codes located along a stripe parallel to the Pareto front. This cluster overlaps largely with the stripes where large H_C (figure 4.10**d**) and low H_{2R} (figure 4.12**c**) were found.

III An interior region of the morphospace mostly consisting of codes

with large connected components and large vocabulary sizes. It includes region F from figure 4.12**a** with low H_R, which indicated a non-uniform sampling of R through random walks over the code graphs.

IV This is seemingly region B from figures 4.7 and 4.10**a**. It consists mostly of codes with large polysemy and small vocabularies, so that communication using these codes would require exhaustive use of contextual cues.

V Codes along the lower third of the Pareto front and a large area adjacent to this part of the front. This group overlaps partly with the region in which fit to power-law distributions has got a low KS-score (i.e. the KS-test indicates good fit to power laws, figure 4.14 – however, part of the region with good fit to power laws falls in the bottom part of group II).

We also ran k-means clustering on the principal components of the measurements performed over Pareto codes alone. For $k = 3$ we found three relatively stable clusters that corresponded to i) a few codes near the one-to-one graph, ii) a few others near the star network, and iii) all remaining codes along the front. However, the boundaries between the clusters changed notably after different initializations of the algorithm. With $k = 5$, the clusters found were not stable at all, meaning that different instantiations of k-means would lump codes together in very different ways. Those cluster would also overlap when plotted along the Pareto front – this is opposed to the analysis in the whole morphospace, whose clustering is very well segregated in target space. This instability of the clusters along the front suggest that we should take with care the previous classification of Pareto optima within clusters I and V. We take these inconsistencies in the analysis of Pareto optimal codes as one final indication of the very different nature of the codes along the Pareto front and hence of the diversity at the critical point that it represents.

Some optimal solutions to the least-effort model proposed in [112] were widely analyzed in the literature from a theoretical perspective [112,

251, 265, 301, 286]. These studies focused on the phase transition, on the existence of Zipf's distribution right at the transition point (which, as we certify here, happens to be a critical point), and on mechanisms that could explain the convergence towards this distribution among all others available. Based on such analyses it was speculated that human language should lay at the transition point, since either extreme was not suitable to describe the flexibility of our communication systems. One-to-one mapping, associated to animal codes, was deemed rather rigid. This is also a code demanding extensive memory usage, which would be used to argue that ambiguity would be the price to pay to alleviate this memory cost. On the other hand, the star code makes communication impossible unless all the information is explicit in the context.

Something missing in the literature is the assessment of real languages using the precise model form [112]. This owes perhaps to the difficulty of building matrices A out of linguistic corpora. WordNet [107, 213] contains a huge database with different semantic relationships, as discussed in section 4.1.4. The database file of the WordNet project includes manually annotated relationships between words and real world objects or abstract concepts. A few examples:

```
ape (...) 02470325 09964411 09796185
car (...) 02958343 02959942 02960501 02960352 02934451
complexity (...) 04766275
rugby (...) 00470966
```

The parenthesis stays for additional information that is not relevant to us right now. Each word is associated to a series of codes. Each numeric code identifies a unique, unambiguous physical object or abstract concept. For example, code 02959942 refers to the car of a railway while code 02960501 refers to the gondola of a funicular or cable car. Hence the word “car” appears associated to these two meanings among others – i.e. it is polysemous. This information is available for four separate groups: adjectives, adverbs, nouns, and names.

We built the corresponding matrices out of the WordNet database and evaluated $H_m(R|S)$ and $H_n(S)$ for each of the four groups of words. De-

spite the work invested in making WordNet accurate, we must keep in mind that we are elaborating a crude mapping of a very complex semantic network into a binary model. After doing so, we readily noted that for all four groups there were more signals than objects. Hence there would be synonyms, which implies that real languages do not lay in the Pareto front. Many models models argue that synonyms should not exist owing to optimality conditions [229, 112, 265]; but synonymy shall be real in folk language usage, which is what worries us here. Besides, synonymy might have degrees and different linguists shall dissent about whether two terms name the precise same concept or not. This is the kind of information lost due to our coarse mapping into binary matrices.

In figure 4.15**a** we plot all four groups in the code morphospace (labeled by *Adj*, *Adv*, *Noun*, and *Verb* respectively and without an apostrophe). While they are not Pareto optimal, they appear fairly close to the front. The most worrying aspect of these real sets of words is that they appear very close to the one-to-one mapping of animal codes located at the upper left corner. This suggests that human language is not such a great departure from those codes in the morphospace. This also contradicts most of the arguments in the literature about least effort language and surprisingly leaves unexplored the huge morphospace available.

However, note that the WordNet database does not contain grammatical words such as pronouns. Some proper names are included in the *Noun* database (e.g. Ada and Darwin), but the pronouns ‘she’ and ‘he’ do not appear anywhere. Neither does the pronoun ‘it’. Note that ‘she’ can substitute any feminine proper name, while ‘it’ can represent any common noun. Similarly, in English it is possible to substitute most verbs for the appropriate form of ‘to do’ or ‘to be’ – e.g. “She plays rugby!” becomes “Does she play rugby?” and eventually “She does!”. We do not discus whether similar mechanisms exist in other languages.

Appending these words to the corresponding matrices would account for adding a signal that can name almost every object, if not all of them. This shifts human languages right into the bottom part of cluster II in morphospace (figure 4.15**a**, points marked *Noun’* and *Verb’* with apostrophes), in a region close to the actual center of the Pareto front. This

suggests that grammatical words might bear all the weight in opening up the morphospace of codes to human languages, with most semantic words conforming a not-so-outstanding network close to the one-to-one mapping and still demanding huge memory usage.

Chapter 5

COMPRESSING SYMBOLIC DYNAMICS

Both this and the next chapters deal with achieving good descriptions of the universe at a low cost. Most of the interesting scientific content is discussed in chapter 6, which deals with the implications of effective modeling for the emergence and persistence of complex life. However, in that chapter we abandon the Pareto optimality strategy for one which is more immediate and flexible to our goals. This is why here, before moving on, we discuss what kind of tradeoffs are posed by the modeling of our physical reality and what kind of questions we would be able to address from the framework introduced in chapter 2. We hypothesize that relevant ‘levels’ or ‘scales’ in the scientific description of nature should stand out in the corresponding Pareto front. We analyze this hypothesis by developing a hierarchy of models for data from natural languages. However, this intends to be an illustrative exercises only. The scientific implications might be wider than we are in a position to address right now. They will be investigated by this author in the future.

5.1. Ants and meaningful levels of description

In *The Mind's I* [151], Douglas Hofstadter presents an interesting debate around the nature of reality and of the human mind. Are they reducible or emergent? Or both? Involved in this discussion we find Achilles, Mr. Tortoise, Mr. Crab, and, more interestingly, Dr. Anteater [152].

This last character is an specialist in ant nests, of course. He has the expected relationship with ants... but a quite singular one with colonies. Over the years he has developed an ability to interpret the swarming patterns, how they adapt to landforms, and how they integrate information about circumstances that hey have met in the past. This sort of memory is encoded by the concentrations of different ant castes [233, 133] across the nest or elsewhere, and is glued together by chemical signals that also serve a kind of short term memory, e.g., when computing optimal routes to sources of food [118].

Dr. Anteater understands that he can interact with the ants too, mainly by changing the landscape. Because of this, he can affect long term memories of the nest in a meaningful way and gets to converse with Aunt Hillary, as the colony is named. Individual ants still fear his presence, but as they recede back underground; emergent, dormant patterns with memories of Dr. Anteater subdue the insects. Altogether, Aunt Hillary enjoys his presence and allows him to *intervene* in her. The Anteater removes selected ants off Aunt Hillary. As a myrmecologist he is not only well-versed in ants but also performs as a “colony surgeon” who specializes “in correcting nervous disorders of the colony by the technique of surgical removal” [152].

A sad yet intriguing episode happened before Aunt Hillary ever existed. The Anteater was friends with Johant Sebastiant Fermant – the former dweller of Aunt Hillary's place, a gifted mathematician and amateur musician. During a hot summer day all his ants and patters where laying under the sun when an unexpected thunderstorm washed them away. It washed the patterns away! Most of the ants survived, some of them with the help of Dr. Anteater, but J. S. F. was gone for good. Over the months,

the surviving ants and others that came anew reconstructed a coherent whole that became Aunt Hillary, a remarkable colony without the faintest idea about music or maths – that is to say, she did not resemble J. S. F. at all.

Why do we assign an entity to individual ants and colonies? We suspect that they are not conscious really, and that they cannot enjoy music or engage in complicated mathematical computations beyond a few hard-wired optimization problems. If not ants, then: why do we assign an entity to neurons and minds? Why do we tend to overlook the possibility that the cerebellum is conscious although it has got more neurons and synapses than the neocortex [18]? Why insist in finding *computational units* for brains [157, 158, 195, 145]? Why not develop a brute-force theory of brains as a whole instead? Why cells, tissues, and organs? Is the hippocampus an organ of itself? Is it V1? Why? Why not?

If we dig deeper: why electrons, protons, or quarks? Quantum mechanics guarantees that these mathematical artifacts – if they have an objective, physical existence – are smeared all over the place and that they become part of abstract unified wholes as the Hamiltonians of different particles interact. Why decompose those Hamiltonians the way we do [322], separating the energetic contributions of each particle, if we know that other decompositions of the Hilbert space work as well? Why do base our models on those abstract particles and not others?

There is an intuitive answer: these choices allow us to develop very informative theories that render myriad details about the reality while being easy to deal with, analytically or by crunching numbers. In cases such as quantum mechanics the eigenvalues and eigenstates of the Hamiltonian suggest that we have selected quite good building blocks (rather, they redefine those building blocks that we had historically been working with). But, more generally, how can we be sure that we are using a good – optimal? – description of our physical reality? We must be confident, since whole branches of sciences are devoted to some of the selected units. Should we confer entity to Aunt Hillary or to J. S. F. but not to the transitory tangle of patterns mediating between them? This case is easy

because both established colonies appear conscious to Dr. Anteater – so it is good to have a utilitarian measure of what *wholes* must do. But often we have not got such a criterion available: we cannot question cellular automata with natural language and, so far, conversing with *real* ant colonies, fish schools, bacterial swarms, or bird flocks is off the table. In these examples, are these *wholes* the only relevant level of description above the individuals? Or are we missing some important, mid-scale phenomenology? *Should* we assign an entity to some lasting patterns of birds or fishes as we do with synapses or synfire chains [1, 137]? *Would* these intermediate structures aid in a theory of swarms or colonies? *Must* a complete theory account for all these levels?

5.2. Information theoretical approaches to level identification

Let us bring these questions to a simpler realm: that of symbolic series. We assume that the whole universe consists of a bi-infinite, one-dimensional series of symbols $\overleftrightarrow{x}$. By bi-infinite we mean that at any arbitrary *present* time we can decompose the symbolic string into past ($\overleftarrow{x}$) and future ($\overrightarrow{x}$) histories. A model of such a universe would step in right at this (or any other arbitrary) present time and *forecast* a future ($\overrightarrow{X}$) given the past. This operation is often probabilistic. Our model $P(\overrightarrow{X}|\overleftarrow{x})$ defines classes of equivalence:

$$\epsilon(\overleftarrow{x}) \;=\; \{\overleftarrow{x}', P(\overrightarrow{X}|\overleftarrow{x}') = P(\overrightarrow{X}|\overleftarrow{x})\}. \tag{5.1}$$

The classes of equivalence ϵ work as *causal states* in our model of the universe. If the model is comprehensive and $P(\overrightarrow{X}|\overleftarrow{x})$ matches always the empirical frequency $F(\overrightarrow{X}|\overleftarrow{x})$ with which symbolic sequences follow each other in $\overleftrightarrow{x}$, then the model reveals the actual causal structure of the process. Following [316], we say that the partition of the space of existing

pasts $\overleftarrow{\mathbf{X}}$ by the equivalence class just described reveals the processes's causal states $S = F(\overleftarrow{X}, \overrightarrow{X})/\sim$.

These causal states are the basis of *computational mechanics* [73, 74, 75], a very powerful approach to extract causal relationships from symbolic series. Computational mechanics introduces a neat representation of causal states and the probabilistic transitions between them. They are called *epsilon machines* (ϵ-machines or ϵ-M, figure 5.1**a-c**) and it is easy to extract from them quantities such as the excess entropy or, more importantly, the *algorithmic complexity* of the process; among others (figure 5.2). Algorithmic complexity is defined as the entropy of the probability distribution over causal states:

$$C_\mu = -\sum_{i=1}^{||S||} P(\epsilon_i) log(P(\epsilon_i)), \tag{5.2}$$

and it constitutes yet another measure that attempts to capture *meaningful information* in complex systems[1].

The representation of an ϵ-M in figure 5.1**a-c** reminds us of a Markov process but there are some important differences. Symbolic dynamics generated by a Markov chain can always be written as ϵ-machines while the opposite is not true. It can be shown that ϵ-M are the unique, smallest, unifilar predictors of $\overrightarrow{X}$. They have a property called *causal shielding* [75, 316]:

$$P(\overleftarrow{X}, \overrightarrow{X}|S) \quad = \quad P(\overleftarrow{X}|S)P(\overrightarrow{X}|S). \tag{5.3}$$

This implies that they capture all the mutual information existing between past and future histories $I(S; \overrightarrow{X}) = I(\overleftarrow{X}; \overrightarrow{X})$, also known as the excess entropy (h_μ).

[1] These kind of measurements capture the intuitive notion of complexity that does not peak at the most disordered or chaotic region, but somewhere towards the *edge of chaos*. This is opposed to the entropy or to Kolmogorov's complexity that have their maximum in the most disordered region, suggesting that they do not capture the *meaningful* part of information that complex systems science aims at.

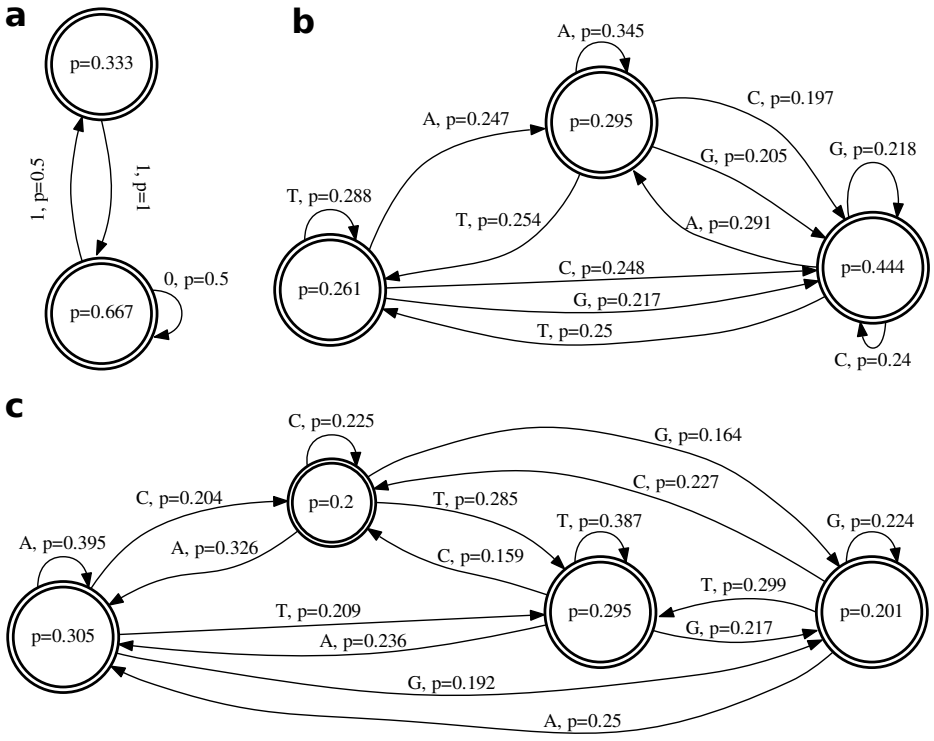


Figure 5.1: **Epsilon machines**. In epsilon machines, causal states (represented by a double circle) are visited randomly. At every causal state we choose between the outgoing arrows according to the probabilities indicated. When an arrow is chosen we emit the symbol associated to that arrow. Enclosed within each causal state we find the frequency with which it shows up in the long term limit. The entropy of the probabilities of the causal states is the algorithmic complexity C_μ. **a** Actual ϵ-M of the golden mean process, which consists of all binary strings that do not present two consecutive zeros. **b** ϵ-M inferred from the genetic sequence of a laboratory construct. **c** ϵ-M inferred from the complete genome of *Mycoplasma pneumoniae*. The increasing size of the ϵ-machines is correlated with an increase of algorithmic complexity in this case – this is not necessarily true.

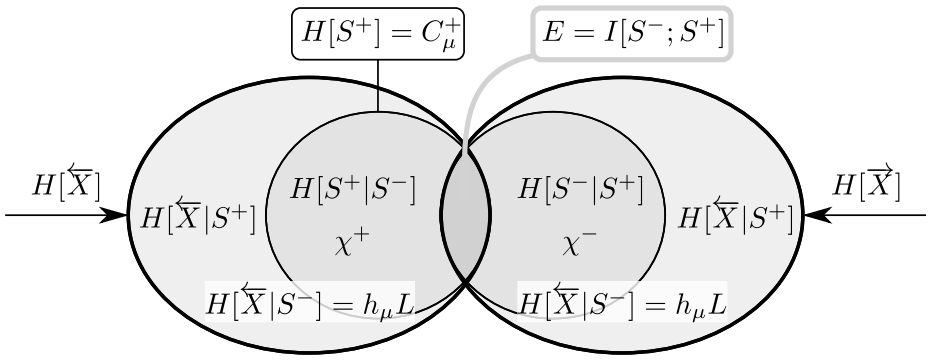


Figure 5.2: **Information theoretical measurements in symbolic dynamics**. A relevant quantities are the entropy of the future dynamics given the causal states ($h_\mu L$), where h_μ represents the entropy production rate – so this number grows linearly with the distance into the future L. The crypticity (χ^+) relates the causal states of forward and backwards dynamics. More importantly for us, the algorithmic complexity C_μ (equation 5.2) captures the complexity of the symbolic dynamics through the entropy of its causal states. This is one of those complexity measurements that is low for deterministic or maximally random processes, but that peaks somewhere near the *edge of chaos* [175].

The implementation of actual ϵ-machines shall require a huge amount of memory – keep in mind that $\overleftrightarrow{x}$ is bi-infinite and its causality can be arbitrarily complex. Several tools are available to reconstruct ϵ-M's from symbolic series [59, 41, 42, 60]. One of the caveats that these numerical techniques come across is straightforwardly related to the next relevant tradeoff that we are interested in.

Certain processes might have small causal states. Consider the *Golden Mean Process* which consists of arbitrary, infinite successions of bits in which there are never two consecutive zeros. We only need to keep track of the last bit to solve this process faithfully. This results in a simple ϵ-M (figure 5.1**a**). Other symbolic dynamics might require that we retain

in memory arbitrarily large *words*. Under some circumstances we might prefer a less comprehensive model even if we loss some of the causal structure. This decision lays in a tradeoff between the memory cost and the loss due to unfaithful predictions. In a nutshell, we wish to come up with models whose states maximize the information that they carry about the future while they minimize the information that they preserve about the past [316, 324].

5.2.1. The information bottleneck method

This optimization problem was addressed explicitly and with examples by Susanne Still et al. [316, 317, 314]. They used *distortion rate* theory [290, 291], a formalism that led them to a single objective optimization:

$$\min_{R} \left\{ I(\overleftarrow{X}; R) + \beta I(\overleftarrow{X}; \overrightarrow{X} | R) \right\}, \tag{5.4}$$

where R is the model – i.e. the variable over which we minimize – and β is the Lagrange multiplier that weights the importance of each target in the optimization. This fits naturally in the framework discussed in chapter 2, so we know that a Pareto front must exist whose cavities and non-analyticities result in first and second order phase transitions in the models by Still et al.

Such transitions are briefly mentioned in the different examples [315, 317, 314], but their nature (first or second order) is not discussed. Equation 5.4 only renders the convex hull of the underlying Pareto front (figure 5.3**a**), so we cannot reconstruct the order of those transitions. Accordingly, those transitions are not pinned down to the shape of the underlying front.

In [316] a class of processes dubbed *predictively reversible* are depicted. They are defined by:

$$P(\overrightarrow{x} | \overleftarrow{x}) = \delta_{\overrightarrow{x}, f(\overleftarrow{x})}, \tag{5.5}$$

where f is invertible, and it is found that the solutions of equation 5.4 trace a straight line (figure 5.3**b**). This is deemed the “worst possible

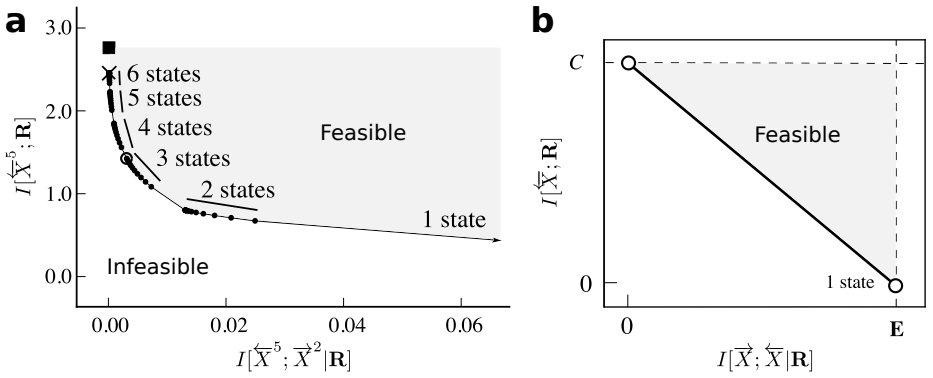


Figure 5.3: **The information bottleneck method reconstructs the convex hull of the Pareto front**. **a** A simple non-deterministic source (see [316]) was reconstructed by models that predicted 2 steps into the future given 5 into the past. Candidate models had different number of causal states (marked in the figure) and perform differently. They trace the rate-distortion curve, that should correspond to the convex hull of an underlying Pareto front. Phase transitions between models take place, but without the actual front we cannot assess their order. **b** Predictively reversible processes present a rate-distortion curve that is a straight line. If this is the actual front of the process, it corresponds to a critical phenomenon according to the framework in section 2.3 of chapter 2. (Figures adapted from [316].)

case for causal compression" [316] and it is noted that every bit gained in predicting the future entails precisely one extra bit of memory to be tracked about the past. Unluckily, we cannot know whether this straight line corresponds to the actual Pareto front or just to the convex hull of a fully concave front (like that of figure 2.4**a** in chapter 2). If the former is true, predictively reversible processes would correspond to the extremely critical case studied in chapter 2, section 2.3. Interesting questions are whether other, similarly critical processes exist and how do they look like.

5.2.2. Conditions for good coarse-grainings of symbolic dynamics

The set of causal states S constitutes a coarse graining of the original dynamics, and a very special one. The dynamics represented by $\overleftrightarrow{x}$ consists of a series of symbols extracted from a certain alphabet $x(t) \in \chi$ (with the index $t \in \{-\infty, \dots \infty\}$). We could use an alternative alphabet Σ that contains a symbol to represent each of the causal states in S. The ϵ-machine consists of this set S and a series of transition probabilities between states (figure 5.1**a-c**) that guarantees that the dynamics in $\overleftrightarrow{x}$ can be reproduced by some other dynamics $\overleftrightarrow{x}_S$ that uses $x_S(t) \in \Sigma$ as coarse-grained variables without losing any causal relationship.

It is fair to wonder whether a similar mapping, in which no causality is lost, exists for other coarse-grainings. ϵ-machines happen to be quite special representations and preserving all causality might be a very stringent condition, hence different authors analyzed what properties (if not causal closure) might be inherited by good coarse-grained dynamics – see [289, 161, 134, 244] among others and [347] for an extensive review on the topic. Finding such well-behaved coarse-grained models may lead us to the identification of relevant levels of description – perhaps correlated with ontological classes that we had already identified in our reality such as the bird flocks, fish schools, or electrons and protons as discussed above. The paradigmatic case of efficient coarse graining are the macrostates of statistical mechanics [289]. Alternatively, if such ontological levels had not been uncovered before, the search for appropriate coarse-grainings offers us a systematic procedure for scientific discover, as suggested in [347].

Following [244] and the notation introduced in section 5.2.1, we identify (figure 5.4**a**) some microscopic dynamics $\overleftrightarrow{x}$ and, more precisely, the mapping ϕ that brings us from the symbol $x(t)$ into the symbol $x(t + \Delta t)$ at successive times. A coarse-graining is induced by a mapping π that projects elements of the alphabet χ into elements of another alphabet χ_R. Finally an arbitrary map ψ may be established at the coarse grained level to evolve symbols $x_R(t)$ into $x_R(t + \Delta t)$. The mapping ϕ corresponds

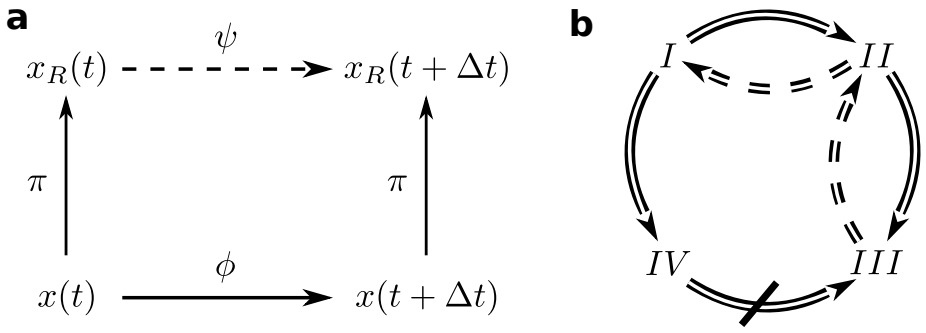


Figure 5.4: **Conditions for good coarse-grainings**. **a** To build a model of a symbolic series we need i) a mapping π that translates the fine grained symbols of alphabet χ into a lumped alphabet χ_R and ii) a mapping ψ that establishes a dynamics in the coarse-grained variables. The original mapping (ϕ) is often unknown. If known, it constitutes the most faithful description of the dynamics over χ, but this is often very costly to simulate. **b** Good mappings π and ψ are such that, once a state has been translated into the coarse-grained symbols, information about the fine grained dynamics cannot give us further information about the coarse grained variables (condition I). This condition implies that ψ enforces Markovian dynamics on χ_R (condition IV) and that the mappings π and ψ commute (condition II). If π is deterministic, some backward implications apply. Markovianity is revealed as the weakest condition.

to the detailed dynamics that we would like to understand, and it might be completely unknown to us. Besides ϕ, any other of the elements involved in figure 5.4**a** might be stochastic or not, and nothing guarantees a priori that they are consistent with each other. Finding those collections $\{\chi_R, \pi, \psi\}$ that make the graph in figure 5.4**a** consistent is our goal.

Pfante et al. [244] analyze four closure conditions:

I Informational closure: Happens if knowledge of the precise microstate ($x(t)$) cannot improve the prediction about the evolution of the macrostate $X_R(t)$. In such case, the coarse-grained dynamics

are informationally closed and there is no information flow from the lower to the higher level.

II **Observational commutativity**: If we coarse grain the variable and then run the macroscopic dynamics we obtain the same result as if we allow the microscopic dynamics to evolve and then perform the coarse-graining.

III **Commutativity**: It exists a mapping ψ such that the previous condition holds. Note the subtle difference between this condition and the previous one. For **commutativity** to hold we just need that it exists a mapping such that the graph in figure 5.4 commutes, but we do not need to use it actively.

IV **Markovianity**: It holds true if the dynamics implemented by ψ are Markovian. This condition has often been considered as the one explicit goal of good coarse-grained models – e.g. [289, 134].

A series of implications are found between these conditions in [244] (figure 5.4**b**, where solid arrows indicate "imply", dashed arrows indicate "imply if π is deterministic", and a crossed arrow indicates that there is not any kind of "imply" relationship). The different properties are arranged in a hierarchy. **Markovianity** is revealed as the weakest one. **Informational closure** stands on the top of the hierarchy.

We could take the A matrices of communication codes studied in chapter 4 as a model for the coarse-graining π. The matrices introduced there related objects to signals. We can take these objects to represent elements of χ and signals to represent elements of χ_R. If an object has *synonym* representations in χ_R some randomness must enter the model in order to decide its mapping. This implies that deterministic models cannot have synonyms, which is the defining feature of Pareto optimal codes. In other words, if and only if π is Pareto optimal (in the sense studied in chapter 4 that simultaneously minimizes the effort of hearers and speakers – i.e. of coders and decoders) all the implication relationships from figure 5.4**b** become valid.

In an alternative approach, Wolpert et al. assign costs to π and ψ, and also consider a penalty due to unfaithful predictions based on the coarse grained model [347]. These costs are weighted linearly as in the information bottleneck method (equation 5.4 in section 5.2.1 and [315, 317, 314]) so that a simple utility function determines when a model is worth implementing or not – irrespective of whether it is Markovian or if commutativity or information closure holds. Wolpert et al. often attempt to find the strict computational cost of each of the mappings involved so that their results would connect with thermodynamics through the Landauer principle that states how much heat is necessary to erase a bit of information.

Important issues remain open: How does Wolpert's et al. utility function relate to the hierarchy identified in figure 5.4**b** and [244]? We know that a Pareto front underlies the phase transitions and criticality conditions of models in [315, 317, 314, 347]. Does the shape of the front relate in a unequivocal way to the information closure hierarchy? Unluckily, causality is not mentioned in [244] and it does not show up in their hierarchy. How does *causal shielding* (equation 5.3) fit in this scheme?

While these are relevant questions, the research attempting to address them is still in very early stages. Instead of delving into these issues, in the next section we decided to focus on a more practical example. Pareto optimality issues just introduced will come into focus again, as well as some concepts and methods from this section.

5.3. A naive empirical approach to the structure of human language

In *More than Nature Needs*, Derek Bickerton defends the idea that the so called *language organ* is present since our birth, but it must *grow* and *develop* until we attain full fledged linguistic capabilities. Of course, a fundamental part of this growing (or learning) must be through example. We could consider language as an *inference organ* that, within the limitations dictates by its rules, gradually internalizes those aspects of our environment that become more relevant to us. We will find that many of

the metaphors that follow work both for language acquisition as we age and for the development of language itself over a proper time scale. Let us take a very brief walk through the phenomenological garden of language.

Most colors are learned in a prefixed order disregard of the tongue that we are raised in: light versus dark first, followed by red, then green or yellow, followed by blue, then brown, and finally purple, pink, orange, and gray [247]. Certain groups do not make use of concepts beyond light and dark, so their languages present a binary divide in color issues. These people, however, can learn new categories without much effort when they are exposed to them. They also learn faster in this predefined order [264, 247].

This language development would affect not only the semantics, but also grammatical aspects. Genders are a great example: besides those more familiar in western cultures (feminine, masculine, and neuter), other groups have developed other gender systems, perhaps because those alternative categories assist them better throughout their *Umwelt*. (Note that gender is not the superfluous distinction between body morphologies, but a linguistic mechanism that enforces a series of coherent structures in our speech.)

It has been proposed that syntax could have emerged ‘for free’ [116, 296], right out of the causal relationships that *Homo sapiens* came across its everyday life a long time ago. In a protolanguage stage, concepts might appear together in a natural way just following their correlations in the physical world. Nim Chimpsky’s [225] largest sentence (“Give orange me give eat orange me eat orange give me eat orange give me you”) is a perplexing example in which ‘eat’, ‘orange’, ‘give’, ‘you’, and ‘me’ appear close to each other due to the natural relationships in which they are involved. This example also shows the failure of a primitive grammar to account for the other, more subtle (and well beyond semantic) structure that must be imposed on sentences in order to communicate in an effective way.

The very core of human language, notably missing in Nim Chimsky’s discourse, might be universal and invariant across cultures [55, 56, 144]. It is speculated that two computational operators (*merge* and *nest* [121])

suffice to generate the remarkable communication system of our species. Building upon these operators, a very rich symbolic dynamics is generated. Each tongue incorporates the constraints arising at multiple levels in its own idiosyncratic way. The few phenomena just mentioned, among many others, make a holistic approach extremely difficult. Linguistics needs to break down into more modest fields including phonetics, syntax, grammar, and semantics. The boundaries between these areas are often diffuse.

In this section we propose an extremely naive and radically empirical approach to study language structure. Ideally, we would start out from a position in which different levels of description (phonetics, grammar, etc) do not exist yet. Using the concepts introduced in section 5.2, we consider the empirical symbolic dynamics that we observe in language corpora. From them we work out a hierarchy of models. If we implement this correctly, our models should proceed from simple and poorly predictive sketches to very detailed descriptions of real languages. Rules belonging in the different branches of linguistics should be incorporated in a parsimonious way. The proposal faces obvious shortages – e.g. while grammatical classes should emerge naturally from our analysis, we needed to make use of some existing ones so that we could work out the simple case presented here. Once these problems are overcome, an advantage of this research program is that all aspects of human language could potentially fall together within one only quantitative theoretical framework.

5.3.1. General problem

The idea is relatively easy to frame but, as it unfolds, non-trivial complexities appear that need to be dealt with. Consider a linguistic corpus – e.g. this PhD Thesis. Then choose the finest grained level of description possible (T^0) for that corpus. Ideally, this is the level at which we have to give the whole text as a description of itself and every word $w_i^0 \in T^0$ ($i = 1, \ldots, ||T^0||$) constitutes a symbol of the dynamics. The alphabet χ^0 is the set of all unique words that appear in the text. We refer to the elements $c_j^0 \in \chi^0$ ($j = 1, \ldots, ||\chi^0|| \leq ||T^0||$) as *classes* – i.e. at this fine grained

level every unique word constitutes a class of itself. This text does not establish a bi- infinite dynamics as those in section 5.2, but we can still use the ideas introduced there. We seek good coarse-grained descriptions of T^0. Historically it has been found that dividing words in grammatical classes allows us to unmask regularities and combinatorial rules through syntactic analysis. An undesired side effect is that we miss most semantic information. This is the fundamental tradeoff that we face again: *simplicity* versus *structure accounted for* by the model. Upon the Pareto front of this problem we shall find the most informative linguistic levels of description.

To find out this optimal tradeoff we need to build coarse-grained models in a systematic way that allows us to quantify how much information is lost and how much compression is gained at every step. We wish to establish arbitrary mappings similar to those in figure 5.4**a** that would connect the fine grained description (our corpus T^0) with some coarse grained version T^l, where the superscript l indexes all possible lumping of words into arbitrary classes.

Following figure 5.4**a**, let us characterize our coarse-graining by two elements: i) the mapping $\pi^l : \chi^0 \to \chi^l$ (such that $\pi^l(c^0_j) = c^l_{j'}$) which lumps many words of the largest alphabet χ^0 into a class of a shorter alphabet $c^l_{j'} \in \chi^l$ ($j' = 1, \ldots, ||\chi^l|| < ||\chi^0||$), and ii) the set of rules represented by $\psi^l : \chi^l \to \chi^l$ that tells us how symbols of the alphabet χ^l follow each other. An example of such a rule is that "In English determinants are usually followed by adjectives or nouns". Our mapping ψ^l could be an arbitrary list of similar statements. This is what most grammar and syntax books consist of. Alternatively, ψ^l might be a mathematical description of these kind of rules, often a stochastic one. Under this light, our grammar or syntax (embodied by ψ^l) acquires an interesting shape that reminds us of similar objects found by generative and construction grammars [54, 305].

5.3.2. Computational and Statistical Mechanics of human language

One possibility to implement this scheme in a systematic way is by means of *Computational Mechanics*. As we have seen in section 5.2, ϵ-machines are a very powerful tool to extract the causal structure of our symbolic dynamics. Given a mapping π^l, we just need to rewrite our corpus using the alphabet χ^l and infer the most likely ϵ-M behind the coarse-grained series of symbols. That machine constitutes our grammar ψ^l. Alternative dynamics $\tilde{\psi}^l$ could be established using the same mapping π^l, but ϵ-machines capture all causal states, so we expect ψ^l to be optimal in a certain sense.

We can use ψ^l to generate synthetic samples and compare them (e.g. by measuring some statistical properties) to the original corpus. Say we come up with a distance $E(\psi^l)$ between the original and synthetic symbols. Besides agreeing with T^0 as much as possible (which translates in low $E(\psi^l)$), good coarse-grainings must be simple. We can measure this, e.g., through the algorithmic complexity $C_\mu(\psi^l)$ of the ϵ-M. These two conflicting targets ($E(\psi^l)$ and $C_\mu(\psi^l)$) allow us to compare different ϵ-machines ψ^l and ψ^m associated to different mappings (π^l and π^m) between χ^0 and arbitrary alphabets χ^l and χ^m.

An evolutionary approach could be established similar to that used in chapter 3 to find Pareto optimal complex networks. The corresponding Pareto front would comprise a hierarchy of ϵ-machines. We expect that cavities and non-analyticities of the front reveal important linguistic scales or shifts between descriptive paradigms (e.g. from semantics to syntactics). We also expect that different corpus generated by one same tongue present reproducible structures. Departures from the standards of a language would provide a basis for empirical comparative studies.

Statistical mechanics is the twin approach to this problem, and the chosen one for the study case presented below. We will use maximum entropy (MaxEnt) models – a tool that was introduced in chapter 3, section 3.2.1. MaxEnt guarantees that we generate the model that introduces less

additional hypotheses given a series of observations. In the example that follows, our observations are given by cross-correlations between pairs of classes.

We assume, as before, that we have a lumping scheme π^l. To build the MaxEnt models we treat each class $c_j^l \in \chi^l$ as a collection $\overrightarrow{\sigma}_j^l = \{\sigma_{j,k}^l, k = 1, \ldots, N_s\}$ of $N_s = ||\chi^l||$ spins. A word $w_i^0 \in T^0$ that is mapped into $c_{j'}^l$ presents $\sigma_{j,k=j}^l = 1$ and $\sigma_{j,k\neq j}^l = 0$. We can now take pairs of words (we could take n-grams of arbitrary length) to produce concatenated collections of spins $\left\langle \overrightarrow{\sigma}^l(i) | \overrightarrow{\sigma}^l(i+1) \right\rangle$, where the index i refers to the position of the words in T^0 and indicates that we take pairs of successive words from the original corpus. This results in a collection of $2N_s$-sized strings of spins.

Let us run a little example. Assume π^l lumps every noun into class c_1^l, every verb into c_2^l, and every other word or symbol ('.', '%', '#', etc) into class c_3^l. The phrase "colorless green ideas sleep furiously" is lumped into $\{c_3^l, c_3^l, c_1^l, c_2^l, c_3^l\}$, which becomes: $\{001, 001, 100, 010, 001\}$. We want to find out couplings between consecutive pairs of words, therefore we group up those bits into: $\{001001, 001100, 100010, 010001\}$. In the general case each of these bit strings will have size $2N_s = 2||\chi^l||$. Large corpora will produce huge collections of bits similar to the previous one. We can summarize them by giving the empirical frequency $F\left(\left\langle c_j^l | c_{j'}^l \right\rangle\right)$ with which each $2N_s$-sized bit string shows up.

The most natural model for such collections of bits are spin glass models. Our task is to find the matrix J^l of couplings between spins that minimizes the distance between $F\left(\left\langle c_j^l | c_{j'}^l \right\rangle\right)$ and the probability distribution over words generated by the model. In our examples we used Minimum Probability Flow Learning (MPFL, [295]), a fast and reliable method that infers J^l given a sufficient sample.

We took a sample of 49 newspaper articles from the Corpus of Contemporary American English [61]. We selected the corpus so that they did not contain foreign (non-English) symbols. We substituted by a period every punctuation symbol that indicated the end of a sentence and removed any other punctuation symbols except for the apostrophes indicating a contraction (e.g. 'don't') or a genitive (e.g. 'someone's'). These

are already important simplifications. In the future we should be able to incorporate those fringe cases too.

Earlier we argued that we should start our analysis on the finest level possible in which every unique word constitutes a class of its own. Relevant grammatical classes should emerge naturally from an efficient modeling. For this simplified example we started from a less ambitious level. This reduced the memory and computational efforts needed. We used Python's Natural Language Processing Toolkit [226] that mapped every word into one of the grammatical classes presented in table 5.1. This preliminary coarse grainining constituted our T^0.

Conjunction	Adverb
Cardinal number	Adverb, comparative
Determiner	Adverb, superlative
Existential there	to
Preposition	Interjection
Adjective	Verb, base form
Adjective, comparative	Verb, past tense
Adjective, superlative	Verb, gerund or present participle
Modal	Verb, past participle
Noun, singular	Verb, non-3rd person singular present
Noun, plural	Verb, 3rd person singular present
Proper noun, singular	Wh-determiner
Proper noun, plural	Wh-pronoun
Predeterminer	Possessive wh-pronoun
Possessive ending	Wh-adverb
Personal pronoun	None of the above
Possessive pronoun	‘.’

Table 5.1: **Grammatical classes present in our corpora**.

Following the instructions explained above, we produced a binary sample from our corpora and fed it into MPFL to infer the couplings $J^0_{kk'}$ for T^0. This model for the most fine-grained level worked as a reference point to compare other models. The indexes k and k' run over

$k, k' = 1, \ldots, 2N_s$. The couplings $J^0_{kk'}$ allow us to associate an energy to pairs of classes:

$$V^0(c^0_j, c^0_{j'}) = \frac{1}{2}\sum_{k,k'} \sigma^0_{j,k} J^0_{kk'} \sigma^0_{j',k'}. \tag{5.6}$$

This in turn tells us the frequency with which we should observe each pair of words according to the model:

$$P\left(\left\langle c^0_j | c^0_{j'}\right\rangle\right) = \frac{1}{Z} e^{\beta V^0(c^0_j, c^0_{j'})}, \tag{5.7}$$

with β an inverse temperature that is also obtained during the inference process and Z the partition function. If our model is good, then the Kullback-Leibler divergence

$$D_{KL}\left(F\left(\left\langle c^0_j || c^0_{j'}\right\rangle\right) | P\left(\left\langle c^0_j | c^0_{j'}\right\rangle\right)\right) \tag{5.8}$$

is low. Tools like MPFL guarantee that this is so at least before we further coarse grain the corpus.

We repeated this same analysis for a series of mappings π^l that gradually contracted grammatical classes into a broader category. For this example the contractions were based on statistical properties of the existing classes so that we would always attempt to lump together classes that looked alike (figure 5.6**a**). In figure 5.5**a** we observe the spin-spin interaction for one of these intermediate models (one in which only 19 classes exist, versus the original 34). This matrix presents a clear box structure:

$$J^l = \left[\begin{array}{c|c} h^l & \overrightarrow{\partial}^l \\ \hline \overleftarrow{\partial}^l & \hat{h}^l \end{array}\right]. \tag{5.9}$$

This is so because the first N_s spins belong to the first class and the last N_s spins belong to the second class in each couple present in the sample. Hence h^l and $\hat{h}^l$ play a role similar to the diagonal biases in the Ising model. These blocks correlate with the frequency with which individual words show up in a corpus.

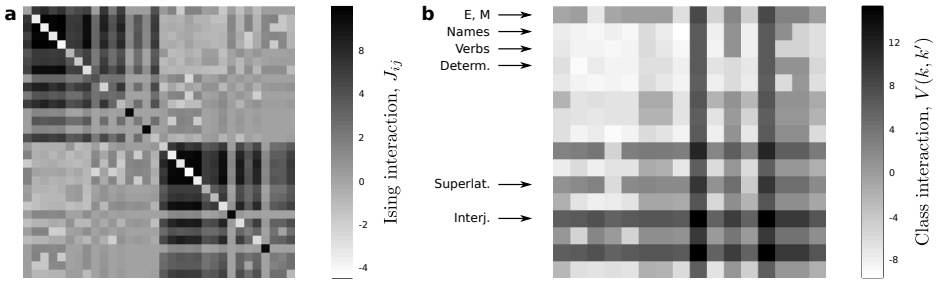


Figure 5.5: **Interactions between spins and word classes**. **a** A first crude model with spins encloses more information than we need for the kind of calculations that we wish to do right now. **b** A reduced version of that model gives us an *interaction energy* between words or classes of words. These potentials capture some non-trivial features of English syntax – e.g. the existential ‘there’ in “there is” or modal verbs (marked E and M respectively) have a lower interaction energy if they are followed by verbs. Interjections present fairly large interaction energy with any other word, perhaps as a consequence of their independence within sentences.

More interesting for us are the interaction terms stored in $\overrightarrow{\partial}^l$ and $\overleftarrow{\partial}^l$. The inference method used guarantees that $\overrightarrow{\partial}^l = (\overleftarrow{\partial}^l)^T$. These bear the non-trivial structure also captured by $V(c^l_j, c^l_{j'})$ (figure 5.6**b**) that tell us, e.g., that the existential there (as in ‘there is’) and modal verbs (both grouped up together by π and presented in the first row in figure 5.6**b**) tend to be followed by some other kind of verb. This description based on interaction energies constitutes our grammars in the proposed approach.

It is possible to use $V(c^l_j, c^l_{j'})$ to generate a synthetic sample $\tilde{T}^l$ and evaluate its energy $E^0(\tilde{T}^l)$ using the most fine grained model J^0. Also, classes at the l-th level of description present a larger entropy if π^l lumps many words together. We can measure this entropy $H^0(\tilde{T}^l)$ over the synthetic sample too. Using $E^0(\tilde{T}^l)$ and $H^0(\tilde{T}^l)$ as optimization targets (that must be respectively minimized and maximized) our models get duly mapped into target space (figure 5.6), where dominance and Pareto op-

timality is well defined.

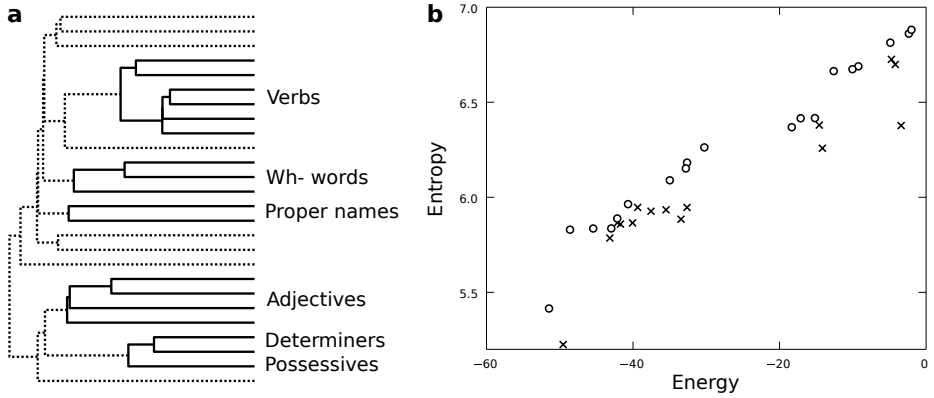


Figure 5.6: **Pareto optimal maximum entropy models of human language**. Among all the models that we try out, we prefer those Pareto optimal in energy minimization and entropy maximization. **a** These reveal a hierarchy of models in which different word classes group up at different levels. The clustering reveals a series of grammatical classes that belong together owing to the statistical properties of the symbolic dynamics, such as possessives and determiners which appear near to adjectives. **b** A first approximation to the Pareto front of the problem. A more accurate one is likely to display cavities along the front corresponding to jumps in the complexity.

For the arbitrary models considered here we already appreciate some π^l more Pareto optimal than others. In elaborating a theory of human language we could dismiss the dominated models knowing that some other one captures at least as much information being even simpler. In the future we will implement evolutionary dynamics that converge to the most Pareto optimal collection of models possible.

5.3.3. Discussion

We have reported a series of preliminary numerical experiments to put forward an ambitious line of research. One immediate computational goal of this theory is to find a hierarchy of models for linguistic corpora. Whether we use computational or statistical mechanics (or any alternative that we might come up with), each of these models constitutes a *generative grammar*. These grammars are extracted empirically from each corpus and, as indicated above, we expect that corpora generated by speakers of a same language result in similar hierarchies of models. Because those models live upon a Pareto front, we also expect that its structure (first and second order phase transitions and critical points) correlate with significant levels of description that have been historically pinned down by the different linguistic disciplines. While we can only speculate at this point, the proposed framework has the potential of bringing together into the same mathematical (empirical and strictly quantitative) theory very distinct levels of linguistic description. This should not be in conflict with other existing approaches (such as Fluid Construction Grammars [305, 306, 307, 308] – instead, both lines of work can benefit each other).

We expect some important regularities for the Pareto fronts obtained for one same language, but we propose that variations within a language can give us important information too. Since our models derive grammars empirically, we can use the huge available corpora across different countries speaking a same tongue to observe how it evolves in real time at a multitude of levels, possibly observing different linguistic scales influence each other.

Alternatively, we could track the unfolding of language in children. Related to this, it has been defended that protolanguage constitutes a separated stage manifested in infants at some stages, in pidgins before they yield to creoles, and, hypothetically, before our species had acquired full fledged language [33]. Certain universalities of protolanguages are reflected in fairly regular patterns that happen in pidgins despite the mixture of languages that they stem from originally [32, 33]. It might be possible to capture this underlying order through the hierarchy of models in

the corresponding Pareto front. If our theory is sensible enough, it will allow a quantitative window to peer into the purported universality of protolanguage versus the huge structural diversity shown by fully developed language.

The communication system of our species is the manifestation of models of the real world elaborated by the brain. Important semantic properties are deeply wired in our cortex, sometimes leading to notable anatomical distinctions [159, 160]. The hierarchy of theories inferred in this chapter allows us not only to peer into this structure at diverse levels, but also to reverse engineer the different relationships between classes of words potentially moving towards more human-like linguistic capabilities in computers.

Chapter 6

INFORMATION THEORY, PREDICTABILITY, AND THE EMERGENCE OF COMPLEX LIFE

In this chapter we explore the consequences of an efficient (informative yet cheap) modeling of the universe for the complexity of life. We note that bare replicative success should not account for the complexity observed in the biosphere. We hypothesize that a tradeoff between predictability of the environment and the cost of the necessary inference machinery might shed some light about this conundrum. We propose a minimal toy model where complexity and meaningful bits can be quantitatively assessed, and in which it is easy to observe *living* entities segregating in different environments according to their complexity. This model may also offer a naive mathematical characterization of Daniel Dennett's *free floating rationals* through which he argues that evolution through natural selection is the origin of all the meaning that we observe.

6.1. Introduction

Simple life forms dominate our biosphere [135] and define a lower bound of embodied, self-replicating systems. But life displays an enormously broad range of complexity levels, affecting many different traits of living entities, from their body size to their cognitive abilities [36]. This creates somewhat a paradox: if larger, more complex organisms are more costly to grow and maintain, why is not all life single-celled? Several arguments help to provide a rationale for the emergence and persistence of complex life forms. Major innovations in evolution involve the appearance of new types of agents displaying cooperation while limiting conflict [207, 319]. A specially important innovation involved the rise of cognitive agents, namely those capable of sensing their environments and reacting to their changes in a highly adaptable way [162]. These agents were capable of dealing with more complex, non-genetic forms of information.

The advantages of such cognitive complexity become clear when considering their potential to better predict the environment, thus reducing the average hazards of unexpected fluctuations. As pointed by Francois Jacob, an organism is *a sort of machine for predicting the future – an automatic forecasting apparatus'* [163, 338, 124]. The main message is that to foresee the future is a crucial aspect of coping with uncertainty. If the advantages of prediction overcome the problem of maintaining and replicating the costly structures needed for inference, the previous view suggest that more complex information-processing mechanisms might be favored under the appropriate circumstances.

Here we aim at providing a minimal model that captures these evolutionary tradeoffs. Specifically, we adopt an information-theoretic perspective in which agents are inference devices interacting with a Boolean environment. For convenience, this environment is represented by a tape with ones and zeros, akin to non-empty inputs of a Turing machine (figure 1**a**). The agent G locates itself in a given position and tries to predict each bit of a given sequence of length n – hence it is dubbed an *n-guesser*. Each attempt to predict a bit involves some cost c, while a reward r is re-

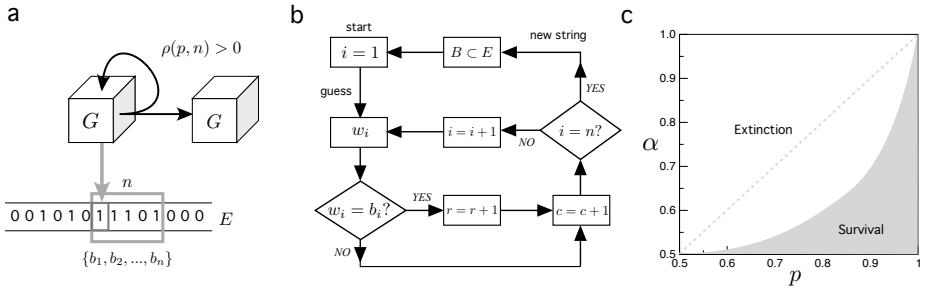


Figure 6.1: **Predictive agents and environmental complexity. a** An agent G interacts with an external environment E that is modeled as a string of random bits. These bits take value 0 with probability p and value 1 otherwise. The agent tries to guess a sequence of n bits at some cost, with a reward bestowed for each correctly guessed bit. The persistence and replication of the agent can only be granted if the balance between reward and cost is positive ($\rho_E^G > 0$). **b** For a machine attempting to guess n bits, an algorithmic description of its behavior is shown as a flow graph. Each loop in the computation involves scanning a random subset of the environment $B = (b_1, ..., b_n) \subset E$ by comparing each $b_i \in B$ to a proposed guess w_i. **c** A mean field approach to certain kind of 1-guesser (modeled in the text through equations 6.1, 6.2, and 6.3) in environments of infinite size renders a boundary between survival ($\rho_E^G > 0$) and death ($\rho_E^G < 0$) as a function of the cost-reward ratio (α) and of relevant parameters for the 1-guesser model (p in this case). Note that for $\alpha < 0.5$ every 1-guesser survives for free.

ceived for each successful prediction. 1-guessers are simple and assume that all bits are uncorrelated, while $(n > 1)$-guessers seek correlations and can get a larger benefit if some structure happens to be present in the environment. A whole n-bit prediction cycle can be described as a program (figure 1**b**). A survival function ρ depends on the number of attempts to guess bits and the number of successful predictions. Successful guessers have a positive balance between reward and prediction cost. These get to replicate themselves and pass on their inference abilities. Otherwise, the agent fails to replicate and eventually dies.

As a simple illustration of our approach, consider a 1-guesser living in an infinitely large environment E where uncorrelated bits take value 0 with probability p and 1 with probability $1 - p$. The average performance of a guesser G when trying to infer bits from E is given by $\bar{p}_E^G$, the likelihood of emitting a correct guess:

$$\bar{p}_E^G = p^G(0)p + p^G(1)(1-p), \tag{6.1}$$

where $p^G(k)$ is the frequency with which the guesser emits bit value $k \in \{0, 1\}$. A strategy that uses $p^G(0) = p$, $p^G(1) = 1 - p$ (i.e. a guesser that mimics the environment) makes in average

$$\bar{p}_E^G = 2p^2 - 2p + 1 \tag{6.2}$$

successful predictions. Its survival function reads:

$$\rho_E^G = (2p^2 - 2p + 1)r - c. \tag{6.3}$$

This curve trivially dictates the average survival or extinction of 1-guessers as a function of the cost-reward ratio $\alpha \equiv c/r$. Note that any more complex guesser (like the ones described below) would always fare worst in this case: they would potentially pay a larger cost to infer some structure where none is to be found. Note also that the tunable parameter α codes for the severity of the environment.

The idea of autonomy and the fact that predicting the future implies performing some sort of computation suggests that a parsimonious theory of life's complexity needs to incorporate reproducing individuals (and

eventually populations) and information (they must be capable of predicting future environmental states). These two components define a conflict and an evolutionary tradeoff. Being too simple means that the external world is perceived as a source of noise. Unexpected fluctuation can be harmful and useful structure cannot be harnessed in your benefit. Becoming more complex (hence able to infer those larger structures, if they exist) implies a risk of not being able to gather enough energy to support and replicate the mechanisms for inference. As will be shown below, it is possible to derive the critical conditions to survive as a function of the agent's complexity and to connect these conditions to information theory.

6.2. Evolution and Information Theory

Key aspects of information theory relate deeply to formulations in statistical physics [166, 167] and there have been several calls to further integrate information theory in biological research [206, 170, 230, 182, 339, 171]. This theory shall play important roles in population or ecosystems dynamics, in regulatory genomics, and in chemical signal processing among others [210, 318, 274, 275, 80, 28, 81, 184, 94, 95, 261, 2, 124, 327, 146, 106], but a unifying approach is far from complete. Given its generality and power, information theory has also been used to address problems that connect Darwinian evolution and far from equilibrium thermodynamics [224, 98, 130, 105, 242]. In its original formulation, Shannon's information theory [290, 291] considers symbols being conveyed from a transmitter to a receiver through a channel. Shannon only deals with the efficiency of the channel (related to its noise or reliability) and the entropy of the source. This theory ignores the content of the emitted symbols, despite the limitations of such assumption [206, 69].

A satisfactory connection between natural selection and information theory can be obtained by mapping our survival function ρ into Shannon's transmitter-receiver scheme. To do so we consider replicators at an arbitrary generation T attempting to “send” a message to (i.e. getting replicated into) a later generation $T + 1$. A successful “message” thus

requires the transmission of information through the survival of the offspring. The former generation acts as a transmitter, the later becomes the receiver, and the environment and its contingencies constitute the channel through which the embodied message must be conveyed (figure 6.2**a**). In other words, we can think of a genotype as a generative model (an algorithm) that produces the message that must be transmitted. That message would be embodied by a phenotype and it includes every physical process and structure dictated by the generative model. It should entail a physical realization of the genotype too since, according to [337], any replicating machine must pass on a physically embodied copy of its instructions[1]. Finally, any evolutionary pressure (including the interaction with other replicating signals) can be included as contrivances of the channel.

Following a similar idea of messages being passed from one generation to the next one, [206] proposes that the replicated genetic message carries *meaningful information* that must be protected *against* the channel contingencies. Let us instead depart from a replicating message devoid of meaning. We realize that the channel itself would convey more reliably those messages embodied by a phenotype that better deals with the environmental (i.e. channel) conditions. Dysfunctional messages are removed due to natural selection. Efficient signals get more space in successive generations (figure 6.2**b**). Through this process *meaningful* bits of environmental information are *pumped* into the replicating signals, such that the information in future messages will *anticipate* those channel contingencies. Meaningful information is not protected from the channel, but emerges naturally from it.

Our view is related to Dennett's interpretation of evolution through natural selection as the ultimate source of meaning [87] – indeed, we are suggesting a mathematical model of this idea. Through his *free floating rationales*, Dennett proposes that there are “reasons that exist that are nobody's reasons”, and that evolution through natural selection captures them. Just what bits of information constitute those reasons? Put

[1]Note that many of the phenotypic structures build in order to get replicated are later dismissed. However, they are passed onto the next generation because the new generative model should be able to replicate those structures again

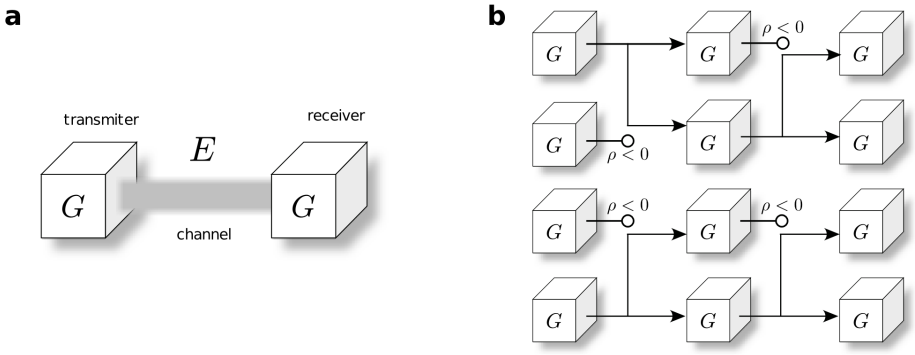


Figure 6.2: **Information and evolution through natural selection. a** A simple diagram of the underlying evolution of a population of bit guessers. The survival and replication of a given agent G is indicated by branching whereas failure to survive is indicated with an empty circle as and endpoint. **b** The propagation of a successful replicator can be understood in terms of a Shannon-like transmission process from one generation to the next.

otherwise, what bits of information are meaningful? And what is the mechanism that renders them meaningful? This is what we model in the next sections in which we argue that meaningfulness emerges out of the tradeoff between replicative vs. inference efficiency.

6.2.1. Messages, channels, and bit guessers

Let us first introduce our implementation of environments (channels), messages, and two classes of agents: transmitters and receivers. Our replicating agents will be dubbed *bit- guessers* because efficient transmission will be equivalent to accurately predicting channel conditions. The notation may seem arid, so it is good to retain a central picture (figure 6.3): Guessers G posses a generative model Γ^G that must fare well in an environment E. Therefore, as G explores subsets of the environment

$B \subset E$, Γ^G produces a behavior W^G roughly comparable to a biological phenotype.

We consider m-environments: strings consisting of m sorted random bits (so that it matters what bits follow each other). We might consider one given m-environment – i.e. one realization (E) of m sorted random bits ($e_i \in E$, $i = 1, \ldots, m$, $e_i \in \{0, 1\}$). Alternatively, we might work with an ensemble (E_m) of m-environments – i.e. all possible environments of the same size ($(e_{i,l} \in E_l) \in E_m$, $l = 1, \ldots, 2^m$, $i = 1, \ldots, m$) or a sample of them ($(E_l \in \hat{E}_m) \subset E_m$, $l = 1, \ldots, ||\hat{E}_m||$). We might evaluate the performance of replicating signals and bit guessers in single m-environments or in an ensemble. As defined, m-environments constitute the conditions of the channel through which messages propagate. To attempt to transmit an n-bit message (W, usually with $n < m$), we extract an n-sized word ($B \subset E$) from the corresponding environment by picking up the bit at a random position in E and the successive $n - 1$ bits. The $b_i \in B$ are compared to the $w_i \in W$ (with $b_i, w_i \in \{0, 1\}$) and each w_i gets through if $w_i = b_i$. Hence attempting to get messages through an environment effectively becomes an inference task: if a guesser can anticipate what bit follows, it has a greater chance of moving into a later generation.

There is a compromise worth investigating between the fidelity of the message that an agent tries to convey and its ability to react to environmental conditions in real time. The generative model Γ^G of the guessers is endowed with a minimal ability to react to the environment on the spot. Thus, rather than conveying a fixed string, they build W as a function of the broadcast history:

$$w_i = w_i(w_1, \ldots, w_{i-1}; b_1, \ldots, b_{i-1}).$$

Because of this, we are rather evaluating ensembles of messages ($W \in W^G$) produced by a same strategy Γ^G. These production strategies are discussed in detail later.

Usually, a guesser attempts to emit an n-bit word many (N_g) times through the same channel. For each one of these broadcasts, a new n-sized word $B^j \subset E$ (with $b_i^j \in B^j$ for $j = 1, \ldots, N_g$ and $i = 1, \ldots, n$) is extracted from the same environment – i.e., as in real life, channel conditions vary.

The manufactured message ensemble becomes W^j with $w_i^j \in W^j$.

We can calculate different frequencies with which the environments or the guessers present bits with value $k, k' \in \{0, 1\}$:

$$p^G(k;i) = \frac{1}{N_g}\sum_{j=1}^{N_g}\delta(w_i^j, k), \tag{6.4}$$

$$p_E(k';i) = \frac{1}{N_g}\sum_{j=1}^{N_g}\delta(b_i^j, k'), \tag{6.5}$$

$$p_{G,E}(k,k';i) = \frac{1}{N_g}\sum_{j=1}^{N_g}\delta(w_i^j, k)\delta(b_i^j, k'), \tag{6.6}$$

$$p_E^G(i) = \frac{1}{N_g}\sum_{j=1}^{N_g}\delta(w_i^j, b_i^j) \Rightarrow \tag{6.7}$$

$$\Rightarrow \bar{p}_E^G = \frac{1}{n}\sum_{i=1}^{n} p_E^G(i); \tag{6.8}$$

with $\delta(x, y)$ being Dirac's delta. Note that $p^G(k;i)$ has a subtle dependency on the environment (because G may react to it) and that $\bar{p}_E^G$ indicates the average probability that guesser G successfully transmits a bit through channel E.

For every bit that attempts to traverse the channel a cost c is paid. A reward $r = c/\alpha$ is cashed in only it that bit is successfully received. α is an external parameter that controls the payoff. The survival function reads:

$$\rho_E^G(\alpha) = (\bar{p}_E^G - \alpha)r. \tag{6.9}$$

As a rule of thumb, if $\bar{p}_E^G > \alpha$ the given guesser fares well enough in the proposed environment.

It is useful to quantify the entropy per bit of the messages produced by G:

$$H(G) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k} p^G(k;i) log\left(p^G(k;i)\right), \tag{6.10}$$

and the mutual information between the messages and the environment:

$$I(G:E) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k,k'} p_{G,E}(k,k';i) \times \log\left(\frac{p_{G,E}(k,k';i)}{p^G(k;i)p_E(k';i)}\right). \quad (6.11)$$

To evaluate the performance of a guesser over an ensemble $\hat{E}_m$ of environments (instead of over single environments) we attempt N_g broadcasts over each of N_e different environments ($E_l \in \hat{E}_m$, $l = 1,\ldots,N_e \equiv ||\hat{E}_m||$) of a given size. For simplicity, instead of labeling $b^j_{i,l}$, we stack together all $N_g \times N_e$ n-sized words W^j and B^j. This way $b^j_i \in B^j$ and $w^j_i \in W^j$ for $i = 1,\ldots,n$ and $j = 1,\ldots,N_gN_e$. We have $p^G(k;i)$, $p_{\hat{E}_m}(k';i)$, $p_{G,\hat{E}_m}(k,k';i)$, $p^G_{\hat{E}_m}(i)$, and $\bar{p}^G_{\hat{E}_m}$ defined just as before, only with j running through $j = 1,\ldots,N_gN_e$. Also as before, we average the payoff across environments to determine whether a guesser's messages get successfully transmitted or not given the length of the environment m and α:

$$\rho^G_{\hat{E}_m}(\alpha) = (\bar{p}^G_{\hat{E}_m} - \alpha)r. \quad (6.12)$$

Importantly,

$$I(G:\hat{E}_m) = \frac{1}{n}\sum_{i=1}^{n}\sum_{k,k'} p_{G,\hat{E}_m}(k,k';i) \times log\left(\frac{p_{G,\hat{E}_m}(k,k';i)}{p^G(k;i)p_{\hat{E}_m}(k';i)}\right) \quad (6.13)$$

is different from

$$\langle I(G:E)\rangle_{\hat{E}_m} = \frac{1}{N_e}\sum_{l=1}^{N_e} I(G:E_l). \quad (6.14)$$

We will come back to this important difference later. We use $\langle\cdot\rangle_{\hat{E}_m}$ to indicate averages across environments of an ensemble $\hat{E}_m$.

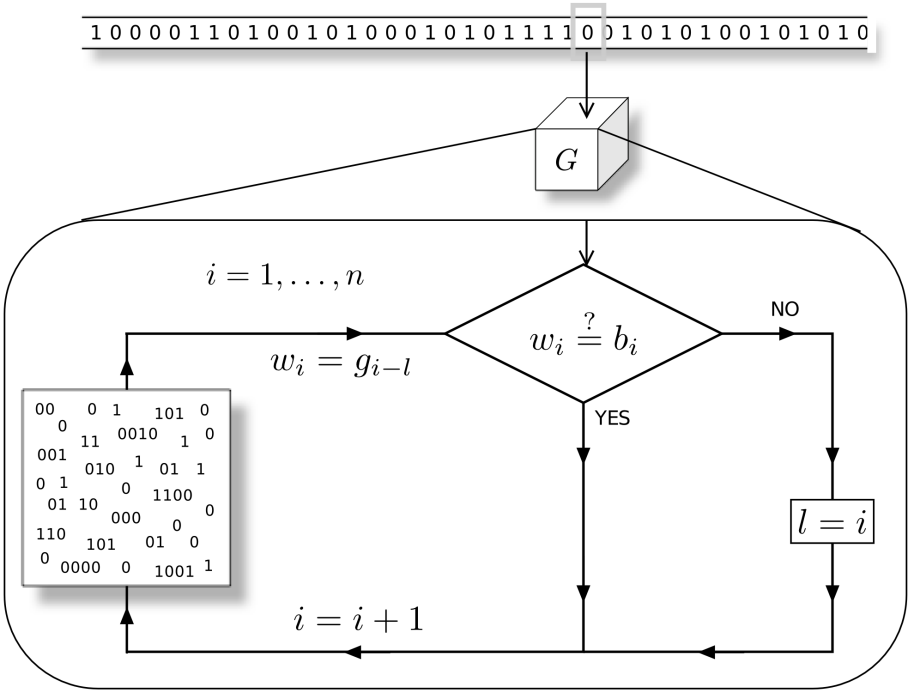


Figure 6.3: **From a generative model to inference about the world.** A diagrammatic representation of the algorithmic logic of the bit guessing machine. Our n-guesser contains a generative model (represented by a pool of words) from which it draws guesses about the environment. If a bit is successfully inferred, the chosen conjecture is pursued further by comparing a new bit. Otherwise, the inference is reset.

The message-generation process Γ^G embodied by our n-guessers could be implemented in different ways, including Artificial Neural Networks (ANNs) [155], spiking neurons [194], Bayesian networks [239, 168], Turing machines [331], Markovian chains [199], ϵ-machines [73], Random Boolean Networks (RBNs) [175], among others. These devices elaborate their guesses through a series of algorithms (e.g. back-propagation, message passing, or Hebbian learning) provided they have access to a sample of their environment.

In the real world, trial and error and evolution through natural selection would be the algorithm wiring the message generation processes into our agents. This would establish some dynamics out of which the relevant inference would emerge. This is a very interesting process (it is, indeed, the implementation of the process that extracts meaningful bits of information through natural selection). However,in this paper we aim at understanding the limits imposed by a channel's complexity and the cost of inference; specially when the performance of the guessers is almost optimal. Therefore, we will assume that our agents perform the near-best inference possible. This best inference will be hard-wired in the guesser's generative model of their environment Γ^G as explained right ahead.

A guesser's generative model usually depends on the environment where it is deployed, so we note $\Gamma^G \equiv \Gamma^G_E$. This Γ^G_E will consist of a pool of bits $g_i \in \Gamma^G_E$ (figure 6.3) and a series of rules dictating how to emit those bits: either in a predetermined order or as a response to the channel's changing conditions. Whenever we pick up an environment $E = \{e_i, i = 1, \ldots, m\}$, the best first guess possible will be the bit (0 or 1) that shows up with more frequency. Hence:

$$\Gamma^G_E(1) \equiv g_1 = \max_{k'} \{p_E(k'; 1)\}\,; \tag{6.15}$$

If both 0 and 1 appear equally often we choose 1 without loss of generality. If the agent succeeds in its first guess, its safest next bet is to emit the bit (0 or 1) that more frequently follows g_1 in the environment. We proceed accordingly if the first two bits have been correctly guessed, if the first three bits have been correctly guessed, etc. Defining $p_{B|\Gamma}(k; i)$ as the probability of finding $k = \{0, 1\}$ at the i-th position of the B^j word

extracted from the environment provided that the guess so far is correct:

$$p_{B|\Gamma}(k';i) \quad = \quad \frac{1}{Z(i)} \sum_{j=1}^{m} \delta(b_i^j, k') \prod_{i'=1}^{i-1} \delta(b_{i'}^j, g_{i'}), \tag{6.16}$$

where j, in this case, labels all n-sized words within the environment $(b_i^j \in B^j) \subset E$ and $Z(i)$ is a normalization constant that depends on how many words in the environment match Γ_E^G up to the $(i-1)$th bit:

$$Z(i) \quad = \quad \sum_{j=1}^{m} \prod_{i'=1}^{i-1} \delta(b_{i'}^j, g_{i'}). \tag{6.17}$$

It follows:

$$\Gamma_E^G(i = 2, \ldots, n) \equiv g_i \quad = \quad \max_{k'} \left\{ p_{B|\Gamma}(k';i) \right\}. \tag{6.18}$$

Note that the pool of bits in Γ_E^G consists of an n-sized word, which is what they try to emit through (i.e. it constitutes the guess about) the channel. If a guesser would not be able to react to environmental conditions, the word W that is actually generated at every emission would be the same in every case and $w_i^j = g_i$ always; but we also allow our guessers a minimal reaction if one of the bits fails to get through the channel (i.e. if one of the guesses is not correct). This minimal reaction capacity by our guessers results in:

$$w_i^j \quad = \quad \Gamma_E^G(i-l) = g_{i-l}, \tag{6.19}$$

where l is the largest i at which $w_i^j \neq b_i^j$. This means that a guesser restarts the broadcast of Γ_E^G whenever it makes a mistake[2].

[2]Note that more elaborated guessers would not only reset their guess. They might browse through a tree with conditional instructions at every point. Besides an extended memory to store the growing number of branches, they would also require nested *if-else* instructions. How should these be compared to plain memory bits? What kind of additional costs do these conditional instructions incur in? On the other hand, ANNs or Bayesian networks might implement such tree-browsing without excessive *if-else* costs.

All together, our guesser consists of a generative model Γ^G that contains a pool of bits and a simple conditional instruction. This is reflected in the flow chart in figure 6.3. Coming back to biology, the generative model is akin to a genotype and the resulting bits emitted as guesses about the channel would make up a phenotype.

We have made a series of choices regarding how to implement environmental conditions. These choices about implementation affect the way some randomness enters the model (reflected in the fact that, given an environment E, a guesser might come across different words $B \subset E$) and also how we implement our guessers (including their minimal adaptability to wrong guesses). We came up with a scheme that codes guessers, the environment (or channel), and the messages transmitted as bit strings. This allows us a direct measurement of information-theoretical features which are suitable for our discussion, but the conclusions at which we arrive should be general: survival will depend on a replicator's ability to extract useful information from the channel (encoded in the message that it conveys from one generation to the next one), and on the cost-efficiency tradeoff related to *how meaningful* bits are.

Because of the minimal implementation discussed, all guessers of the same size are equal. Environmental ensembles of a given size are considered equivalent as well. Hence, the notation is not affected if we identify guessers and environments by their sizes. Accordingly, in the following we substitute the labels G and E by the more informative ones n and m respectively. Hence $\rho^G_{E_m}(\alpha)$ becomes $\rho^n_m(\alpha)$, $\bar{p}^G_E$ becomes $\bar{p}^n_m$, etc.

6.3. Life complexity is tuned by the predictability-replication tradeoff

We report a series of numerical experiments. Some of them deal with guessers in environment ensembles of fixed size, others allow guessers to switch between environment sizes to find a place where to thrive. The important message is that the complexity of the guessers that can populate a

given environment (i.e. the complexity of the most successful messages at traversing a given channel) is tuned by the predictability of the environment. Environment sizes exist at which simple guessers die off but in which more complex life flourishes – thus explaining real-life excursions beyond bare replicative success. Our scheme allows us to quantify the different aspects involved in the tradeoff.

Finally, some of the simulations engage n-guessers with different n in direct competition or explore what may happen if resources could get exhausted. This is suggestive of how the current approach will be expanded in future research.

6.3.1. Guessers isolated in environments of fixed size

Figure 6.4 shows $\bar{p}^n_m$, the average probability that n-guessers correctly guess 1 bit in m-environments. The 1-guesser (that lives off maximally decorrelated bits given the environment) establishes a lower bound. More complex machines will guess more bits in average, except for infinite environment size $m \to \infty$, at which point all guessers have equivalent predictive power.

As m grows, environments (which consist of collections of random bits) get less and less predictable. Importantly, the predictability of shorter words decays faster than that of larger ones, thus enabling guessers with larger n to survive where others would perish. There are 2^n possible n-words, of which m are realized in each m-environment. When $m >> 2^n$, the environment implements an efficient, ergodic sampling of all n-words – thus making them maximally unpredictable. When $n \lesssim m < 2^n$ the sampling of n-sized words is far from ergodic and a non-trivial structure is induced in the environment because the symmetry between n-sized words is broken – they cannot be equally represented due to finite size sampling effects.

This allows that complex guessers (those with the ability to contemplate larger words, keep them in memory, and make choices regarding information encoded in longer strings) can guess more bits, in average, than more simple entities. In terms of messages crossing the channel,

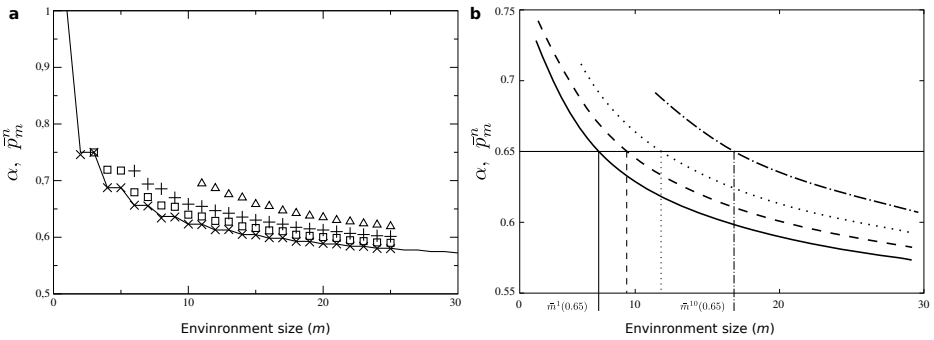


Figure 6.4: **Probability of correctly guessing a bit in environment ensembles of constant size.** $\bar{p}_m^n$, average probability that n-guessers correctly guess 1 bit in m-environments for n = 1 (crosses, solid line), n = 2 (squares, dashed line), n = 5 (pluses, dotted line), and n = 10 (triangles, dot-dashed line). $\bar{p}_m^1$ can be computed analytically (solid line in the main plot) and marks an average, lower predictability boundary for all guessers. In the inset, the data has been smoothed and compared to a given value of α, represented by a horizontal line. At the intersection between this line and $\bar{p}_m^n$ we find $\bar{m}^n(\alpha)$, the environment size at which n-agents guess just enough bits as to survive given α. Notice that n guessers are evaluated only in environments of size $m \geq n$.

while shorter words are meaningless and basically get transmitted (i.e. are correctly guessed) by chance alone, longer words might contain meaningful, non-trivial information that get successfully transmitted because they cope with the environment in the adequate way.

Note that this symmetry breaking to favor predictability of larger words is just a mechanism that allows us to introduce correlations in a controllable and measurable way. In the real world this mechanism might correspond to asymmetries between dynamical systems in temporal or spatial scales. Although our implementation is rather ad hoc (suitable to our computational and conceptual needs), we propose that similar mechanisms might play important roles in shaping life and endowing the uni-

verse with meaningful information. Indeed, it might be extremely rare to find a kind of environment at which words of all sizes become non-informative simultaneously.

Back to our experiments, the mutual information between a guesser's response and the environments (i.e. between broadcast messages and channel conditions) further characterizes the advantages of more complex replicators. Figure 6.5**a** shows $I(G:E_m)$ and $\langle I(G:E)\rangle_{E_m}$. As we noted above, these quantities are not the same. Let us focus on 1-guesser for a moment to clarify what these quantities encode.

Given an environment, 1-guessers have got just one bit that they try to emit repeatedly. They do not react to the environment – there is not room for any reaction within one bit, so their guess is persistently the same. Hence the mutual information between the emitted bit and the arbitrary words $B \subset E$ that the guesser comes across is precisely zero, as shown in the inset of figure 6.5**a**. Hence, $\langle I(G:E)\rangle_{E_m}$ captures the mutual information due to the slight reaction capabilities of guessers to the environmental conditions.

While the bits emitted by 1-guessers do not correlate with $B \subset E$, they do correlate with each given E since they represent the most frequent bit in the environment. Accordingly, the mutual information between a 1-guesser and the aggregated environments (reflected by $I(G:E_m)$) is different from zero (figure 6.5**a**). To this quantity contribute both the reaction capability of guessers and the fact that they have hard-wired a near-optimal guess in Γ^G_E, as explained in section 6.2.1.

We take the size of a guesser n as a crude characterization of its complexity. n represents the number of bits needed to encode a guesser and larger guessers can store more complex patterns, so this is justified. $\langle H(G)\rangle_{E_m}$ indicates that more complex guessers look more entropic than less complex ones (figure 6.5**b**). Larger guessers come closer to the entropy level of the environment (black thick line in figure 6.5**b**), which itself rapidly tends to $log(2)$ per bit. This is remarkable: better performing guessers would appear more disordered to an external observer even if they are better predictors when considered within their context. Note that $\langle H(G)\rangle_{E_m}$ is built based on the bits actually emitted by the

guessers, meaning that this quantity correlates with the complexity of the phenotype. For guessers of fixed size n, we observe a slight decay of $\langle H(G)\rangle_{E_m}$ as we proceed to larger environments. This quantity should tend to $\langle H(G)\rangle_{E_\infty}$: a value different from zero and, in general, characteristic of each size. This quantity might be put forward as an intrinsic measure of complexity for n-guessers.

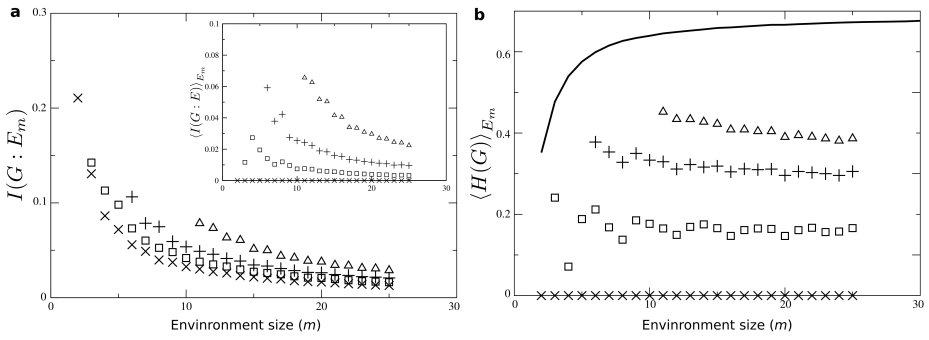


Figure 6.5: **Mutual information and entropy.** Guessers with $n = 1$ (crosses), $n = 2$ (squares), $n = 5$ (pluses), and $n = 10$ (triangles) are presented. **a** $I(G : E_m)$ and $\langle I(G : E)\rangle_{E_m}$ (inset) quantify the different sources of information that allow more complex guessers to thrive in environments in which simpler life is not possible. **b** The entropy of a guesser's message given its environment seems roughly constant in these experiments despite the growing environment size. This suggests an intrinsic measure of complexity for guessers. Larger guessers look more random even if they might carry more meaningful information about their environment. The thick black line represents the average entropy of the environments (which approaches $log(2)$) against which the entropy of the guessers can be compared.

6.3.2. Evolutionary drivers

The key question is whether the payoff may be favorable for more complex guessers provided that they need a more costly machinery in order to get successfully reproduced. As discussed above, the cost of units of ANNs or the number of states in Bayesian networks would enter this evolutionary game if such implementations of bit guessers were chosen. To keep the discussion simple, and without loss of generality, bit guessers incur only in a cost proportional to the number of bits that they try to transmit. Equation 6.12 captures all the forces involved: the cost of transmitting larger messages versus the reward of a successful transmission that comes after more complex environments could be apprehended.

Guessers of a given size survive in an environment ensemble if, in average, they can guess enough bits of the environment or, alternatively, if they can convey enough bits through the channel (in any case, if $\bar{p}^n_m > \alpha$, which implies $\rho^n_m > 0$). Setting fix a value of α we find out graphically $\bar{m}^n(\alpha)$, the largest environment at which n-guessers can survive (figure 6.4, inset). Because m-environments look more predictable to more complex guessers we have that $\bar{m}^n(\alpha) > \bar{m}^{n'}(\alpha)$ if $n > n'$. This guarantees that for $\alpha > 0.5$ there always exist m-environments from which simple life is banned while more complex life can thrive. For $\alpha \leq 0.5$ any guesser survives in average.

We establish a dynamics that must gravitate around $\bar{m}^n(\alpha)$. After evaluating an n-guesser $N_g \cdot N_e$ times in an arbitrary m-environment, the guesser is promoted to $m + 1$ if $\hat{\rho}^n_m(\alpha, N_g, N_e) > 0$, where $\hat{\rho}$ represents an empirically accumulated reward instead of the ensemble average. If $\hat{\rho}^n_m(\alpha, N_g, N_e) < 0$, the guesser is demoted to $m - 1$. The steady state of this dynamics is characterized by a distribution $P^n(m; \alpha, N_g, N_e) \equiv P^n(m, \alpha)$, i.e. the frequency with which n-guessers are found in environments of a given size (figure 6.6**a**). The overlap and gaps between $P^n(m, \alpha)$ for different n suggest that: i) some guessers would engage in harsh competition if they needed to share environments of a given kind and ii) there is room for different guessers to get segregated into environ-

ments of increasing complexity. The average

$$\hat{m}^n(\alpha) \;=\; \sum_m m P^n(m,\alpha) \tag{6.20}$$

should converge to $\hat{m}^n(\alpha) \simeq \bar{m}^n(\alpha)$ under the appropriate limit – i.e. if we evaluate the guessers numerically enough times as to approach these mean field values. Figure 6.6**b** shows dynamically-derived averages $\hat{m}^n(\alpha)$ and some deviations around them as a function of α.

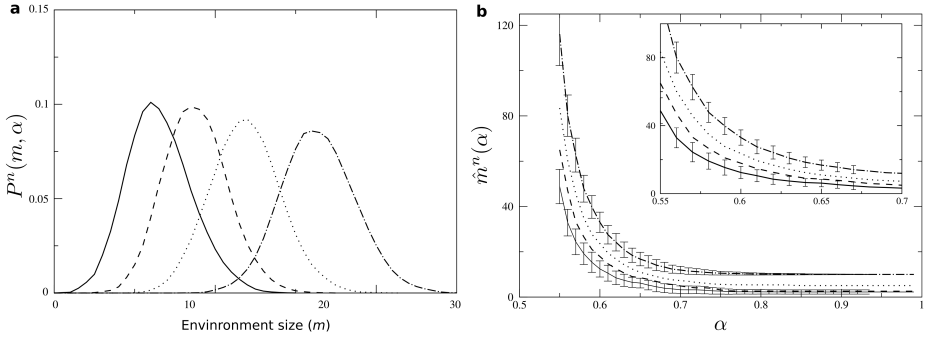


Figure 6.6: **Dynamics around $\bar{m}^n(\alpha)$.** Again, guessers with $n = 1$ (solid line), $n = 2$ (dashed line), $n = 5$ (dotted line), and $n = 10$ (dot-dashed line). **a** $P^n(m,\alpha)$ tells us how often do we find n-guessers in m-environments when they are allowed to roam constrained only by their survival function ρ_m^n. The central value $\hat{m}^n$ of $P^n(m,\alpha)$ must converge to $\bar{m}^n(\alpha)$ and oscillations around it depend (through N_g and N_e) on how often do we evaluate the guessers in each environment. **b** Average $\hat{m}^n$ for $n = 1, 2, 5, 10$ and standard deviation of $P^n(m,\alpha)$ for $n = 1, 10$. Deviations are not presented for $n = 2, 5$ for clarity. The inset represents a zoom in into the main plot.

It is easily justified that guessers drop to simpler environments if they cannot cope with the complexity. It is less clear why they should seek more complicated environments if they thrive in a given one. This might happen spontaneously if simpler environments get crowded or if resources get exhausted, as we study right ahead.

To simulate a competition dynamics, n-guessers with $n = 0, 1, 2, 3,$ and 4 were randomly distributed occupying 100 environments of fixed m and were assigned an initial $\hat{\rho}_i(t = 0) = n\rho_0$ with $i = 1, \ldots, 100$ labeling the guesser at each environment. Larger guessers start out with larger $\hat{\rho}_i(t = 0)$ representing that they come into being with a larger metabolic load satisfied. A 0-guesser represents an unoccupied environment. New empty environments might appear only if actual ($n \neq 0$) guessers die, as we explain below. We tracked the population using $P_m(n, t)$, the proportion of 0-, 1-, 2-, 3-, and 4-guessers through time. (These experiments where the more computationally demanding, that is why we took $n = 1, 2, 3, 4$ instead of the values $n = 1, 2, 5, 10$ used throughout the paper. The insights gained from the simulations do not depend on the actual values of n.)

At each iteration, a guesser (say the i-th one) was chosen randomly and evaluated with respect to its environment. Then the wasted environment was replaced by a new, random one of the same size. When choosing a random guesser for testing, we took care that every guesser attempts to guess the same amount of bits in average. This means, e.g., that 1-guessers are tested twice as often as 2-guessers, etc. If after the evaluation $\hat{\rho}_i(t + \Delta t) < 0$, then the guesser died and it was substituted by a new one. The n of the new guesser was chosen randomly and proportionally to the current distribution $P_m(n, t)$. If $\hat{\rho}_i(t + \Delta t) > 2n\rho_0$, the guesser got replicated and shared its $\hat{\rho}_i$ with its daughter, who overrode a random slot $i' \in \{1, \ldots, 100\}$. This replication at $2n\rho_0$ represents that parents must satisfy a metabolic load that grows with the size n of the guessers. There is a range ($0 < \hat{\rho}_i < 2n\rho_0$) within which guessers are alive but do not replicate. Figure 6.7**a** and **b** show $P_m(n, t = 10\,000)$ with $\alpha = 0.6$ and 0.65. Note that for large environments all guessers combined do not add up to 100 – i.e. mostly empty slots remain and most guessers get extinguished. The most abundant guesser after $10\,000$ iterations as a function of m and α is shown in figure 6.7**c**.

An alternative evolutionary pressure is introduced if the bits in the environment represent resources that might get exhausted. Thinking from

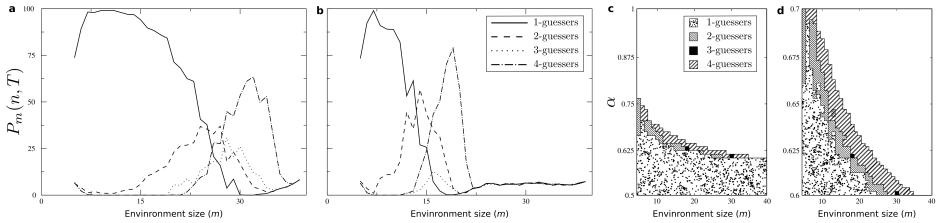


Figure 6.7: **Evolutionary drivers: competition.** Coexisting replicators will affect each other's environments in non-trivial ways which may often result in competition. We implement a dynamics in which 1-, 2-, 3-, and 4-guessers exclusively occupy a finite number of environments (transmission channels) of a given size (fixed m). The 100 available slots are randomly occupied at $t = 0$ and granted to the best replicators as the dynamics proceed. We show $P_m(n, t = 10\,000)$ for $m = 5, \ldots, 39$ and **a** $\alpha = 0.6$, **b** $\alpha = 0.65$. The most abundant guesser at $t = 10\,000$ is shown for **c** $\alpha \in (0.5, 1)$ and **d** $\alpha \in (0.6, 0.7)$. Once m is fixed, there is an upper value of α above which no guesser survives and all 100 available slots remain empty. Competition and the replication-predictability tradeoff segregate guessers according to the complexity of the environment – i.e. of the transmission channel. Coexistence of different guessers seems possible (e.g. $m = 15$ in **b**), but it cannot be guaranteed that the dynamics have converged to a steady distribution.

the message broadcasting perspective, a spot on the channel might appear crowded if it is engaged in a successful transmission. Assume that every time that a bit is correctly guessed, it gets exhausted (or gets crowded) with an efficiency β so that in average each bit cannot contribute any reward $\beta(\bar{p}_m^n/m)$ of the time. The average reward extracted by a guesser from an ensemble (figure 6.8) becomes:

$$\tilde{r}_m^n = \left(1 - \beta\frac{\bar{p}_m^n}{m}\right)\bar{p}_m^n r. \tag{6.21}$$

Smaller guessers living in very small environments quickly crowd their channels (alternatively, exhaust the resources they depend on). In figure

6.8**b** given some α and β = 1, 1- and 2-guessers can only survive within some under and upper limits (figure 6.8**b**).

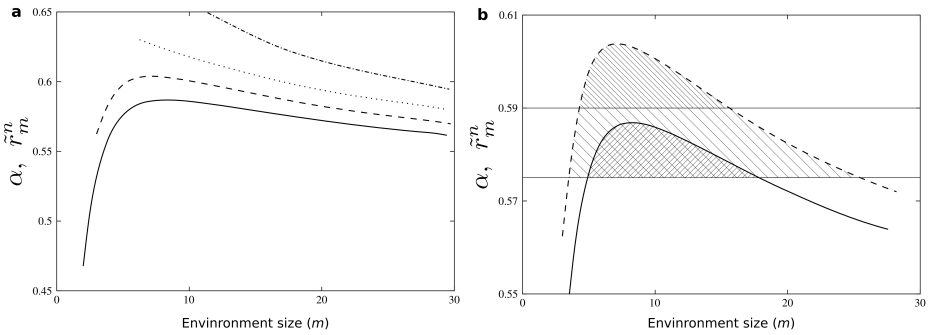


Figure 6.8: **Evolutionary drivers: exhausted resources.** Rather than monopolizing channel slots (as in figure 6.5), we can also conceive individual bits as valuable, finite resources that get exhausted whenever they are correctly *guessed*. Then a successful replicator can spoil its own environment and new conditions might apply to where life is possible. **a** Average reward obtained by 1-, 2-, 5-, and 10-guessers in environments of different sizes when bits get exhausted with efficiency β = 1 whenever they are correctly guessed. **b** Given α = 0.575 and α = 0.59, 1- and 2-guessers can survive within upper and lower environment sizes. If the environment is too small, resources get consumed quickly and cannot sustain the replicators. In message transmission language, the guessers crowd their own channel. If the environment is too large, unpredictability takes over for these simple replicators and they perish.

6.4. Discussion

Our models show how the complexity of life that thrives in a given kind of environment is dictated by a replication-predictability tradeoff. More complex environments look more unpredictable to simpler replicators. Agents that can keep a larger memory and make inferences based

on more elaborated information can extract enough valuable bits from the environment as to survive in those more challenging situations. Despite the inevitable cost inherent to the cognitive machinery, a selection process towards more complex life is shown to occur. In our study we identify a transmitter (replicators at a given generation), a receiver (replicators at the next generation), and a channel (*any* environmental conditions) through which a message (ideally instructions about how to build newer replicators) is passed on. Darwinian evolution follows naturally as effective replicators transit a channel faster thus getting more and more space in successive generations. The inference task is implicit as the environment itself codes for meaningful bits of information that, if picked up by the replicators, boost the fitness of the phenotypes embodied by the right dynamics.

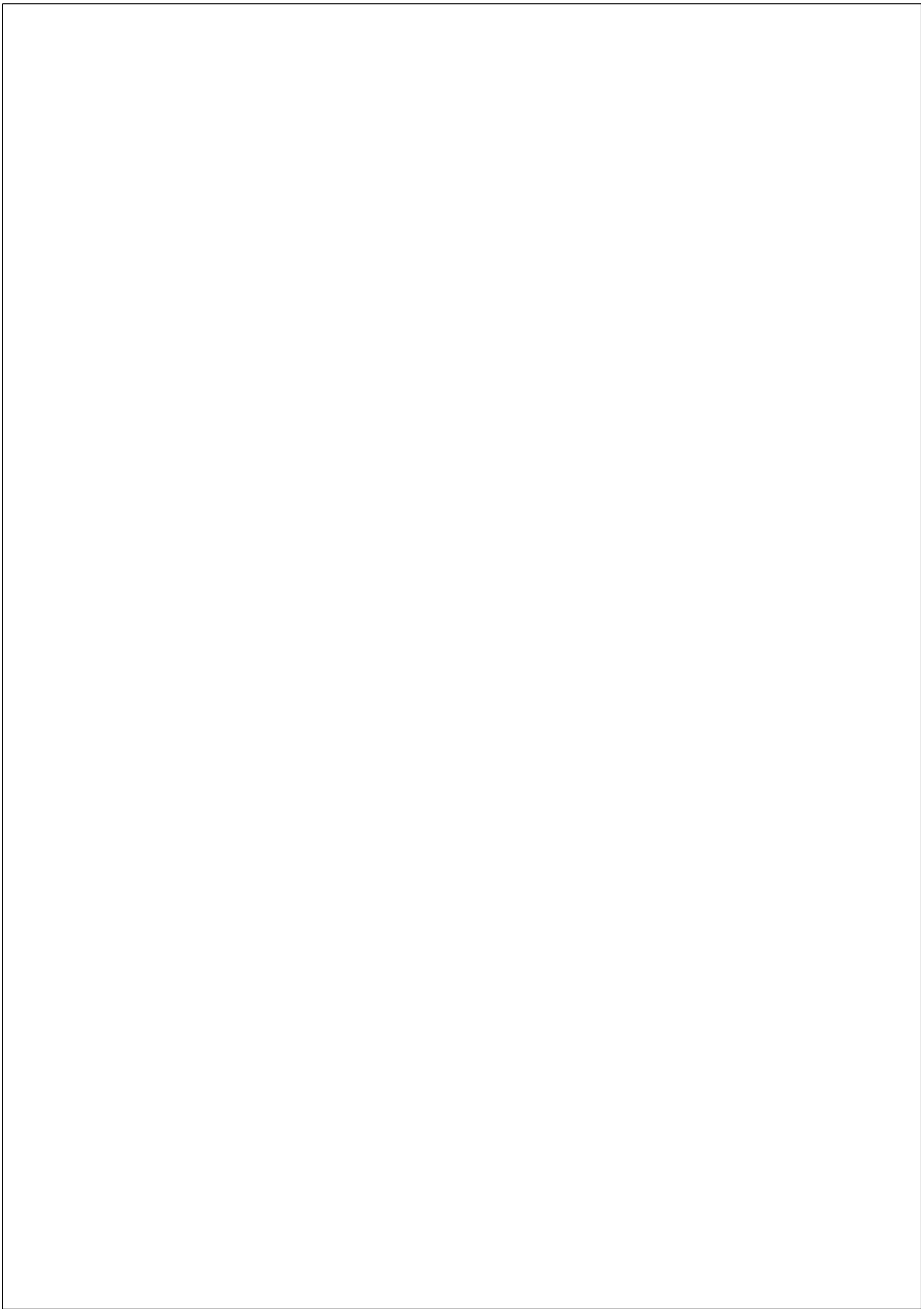
This view is directly inspired by a qualitative earlier picture [206]. That metaphor assigned to a DNA message some external meaning that had to be preserved against the environmental noise. We propose that *all* meaning is contained in the channel and the constraints that it imposes to the propagated messages. Through natural selection, these relevant bits are incorporated into the genome or whatever physical support a replicated message might rely on. This is an elaborated, quantitative approach to Dennett's *free floating rationales* [87] that assign to natural selection the origins of all meaning. The way that we introduce correlations in our scheme (through a symmetry breaking between the information borne by short and larger words due to finite size effects) is compatible with this view. However, interestingly, it also suggests that meaningful information might arise naturally even in highly unstructured environments when different spatial and temporal scales play a relevant role.

This way of integrating information theory and Darwinism is convenient to analyze the questions at hand that concern the emergence of complex life forms. But this also suggests new lines of research. As introduced in this paper, guessers and their transmissible messages might and should shape the transmission channel (e.g., by crowding it, as explored in section 6.3.2). What possible co-evolutionary dynamics between guessers and channels can be established? Are there stable ones, others leading to

extinction, etc? Do some of them, perhaps, imply open-ended evolution? Which ones? A guesser's transmitted message might be considered an environment in itself, opening the door to ecosystem modeling based on rigorous information theoretical processes. It is also suggested the exploration of different symbiotic relationships from this perspective and how they might affect coevolution.

Finally, an important question was left aside that concerns the memory vs adaptability tradeoff of bit guessers. Here we studied guessers with a minimal adaptability to focus on the emerging hierarchy of complexity. Adaptability at faster (say, at behavioral) temporal scales is linked to more complex inferences with richer dynamics. This brings in new dilemmas as to how to weight the different building blocks of complex inference – as advanced above, how do we compare memory and *if-else* or *while* instructions?

However these costs are solved, the nature of this conflict seems to lay on a temporal scale dissociation alone. Its solution depends on the answer to the question: 'Does this meaningful structure happen often enough as to incorporate it in my DNA, or is it convenient to elaborate a costly inference machine that reacts (behaviorally) in real time?' Note that a population of interacting agents that mutually interfere with each other's environments or channels might yield replicating structures that rely not in long-lasting genetic instructions, but in those fast, adaptable inference machines that do not hard-wire their relevant information. This suggests ways to incorporate memetic evolution into the picture and offers a candidate for the origin of the genetic- memetic dissociation. Importantly, it might be possible to derive quantitative conclusions about this problem.



Chapter 7

CONCLUSIONS

In this work we have adopted optimization (and Pareto optimality most of the time) as a general framework to analyze conflicts that arise in a series of complex systems. The problems tackled include graph theory, protein folding, several aspects of theoretical linguistics, and efficient modeling of the world and its implications for the evolution of complex life. Along the way we have stated a series of hypotheses. Some of them, examined in the earlier chapters of this thesis, refer to the mathematical intrinsic nature of optimization problems and should have more general consequences. The rest of the conjectures are more specific to each individual topic, but they often stem from some optimality principle. We review now all these hypotheses and whether (and how) they have been resolved.

Hypothesis 1, **Optimization is a unifying principle in complex systems**:

This has been illustrated with examples in Chapter 1 and throughout the rest of this thesis. The use of optimization methods to localize similar universal features (phase transitions and critical points) in problems across fields further strengthens this pervasive view of optimality.

Hypothesis 2, **Conflicting traits are a hallmark of complex systems, hence Pa-**

reto optimality stands as a natural framework for them:

Without conflicting traits, optimal systems converge to uniform solutions. We have seen how different optimality constraints must be taken into account to describe a series of systems that are often cataloged as complex – such as human language or evolutionary dynamics. The Pareto front locates geometrically, along a tradeoff, the many designs that make up those complex systems.

This point is related to the long-standing motto of complex systems research: ‘*life at the edge of chaos*’ [175]. Very orderly dynamics fall into unchanging – and uninteresting – stable states, while extreme disorder leads to chaos or noise that cannot sustain salient structures. Despite the apparent turmoil, these extremely disordered regimes often accept very simple descriptions (take, as an instance, the most entropic states in thermodynamics). Only at the *edge of chaos*, where both conflicting extremes meet, do we find meaningful structure with non- trivial yet lasting-enough behaviors.

Hypothesis 3, **A series of universal features emanate from the mathematical structure of optimization alone**:

In this thesis we have proved how the Pareto front is a generalization of the Gibbs surface in thermodynamics, and we have shown how it captures phase transitions and critical points [282, 286]. These robust, macroscopically observable features, hence, shall appear naturally in the study of complex systems and are deeply entwined in the mathematics of optimization, beyond the details of each individual problem.

Hypothesis 4, **These universal features should be manifested differently and have distinct implications depending on the nature of each given problem**:

While phase transitions and criticality in different systems emanate from similar accidents in some underlying mathematical description, they have different implications for the different phenomena

modeled. Phase transitions could take the form of radical rearrangements in optimal networks or protein folding, while for the communication tradeoff a phase transition separates memory-demanding communication from a no-communication regime. For complex life, too, a tradeoff might sharply bar the persistence of certain kinds of living systems.

Hypothesis 5, **Complex networks are a perfect testbed to illustrate the different universal features that show up in Pareto optimal systems**:

Using different optimization targets we can manufacture a series of problems that illustrate criticality and first and second order phase transitions [285]. The proposed tasks relate to actual engineering problems, e.g. building efficient communication or transport networks. Locating first order phase transitions can help us anticipate costly reconfigurations in real life. Second order phase transitions might indicate especially robust solutions to problems with conflicting traits. Finally, critical points might show us where degenerate solutions exist and how to harness them if necessary.

Hypothesis 6, **Pareto selective forces always drive certain systems to a critical point**:

Criticality is related to straight stretches of the Pareto front. Pareto selective forces are defined as any algorithmic means that evolves an ensemble of designs – whatever their nature – towards an underlying Pareto front. It follows that whenever the Pareto front is a straight line, Pareto selective forces *always* drive those precise systems (but not others) towards a critical point [284]. We illustrated with three examples involving different kinds of Pareto optimal networks. We discussed several processes that implement Pareto selective forces, from design to ecosystem dynamics to economic markets.

All together, we have exposed a comprehensive and parsimonious framework that brings together criticality with the outcome of optimal evolutionary paths. A series of relevant systems (e.g. those

with Highly Optimized Tolerance – HOT, [45, 46]) whose criticality has been the object of heated debate can now be assessed in a theoretically sound manner.

Hypothesis 7, **The new paradigm enriches the interpretation of existing literature on complex systems**:

While phase transitions appear widely discussed in problems tied to Lagrange multipliers, to the very best of our knowledge they are not mentioned in the extensive literature on Pareto optimal systems. This invited us to revise a series of works on Pareto optimality. The tradeoffs that we have found, under the light of our work, reveal phase transitions or critical points. Our contributions extend beyond this brief revision of literature: they affect any previous work on Pareto optimality and also offer a well-defined framework to interpret future contributions.

In our literature revision we uncovered phase transitions in models of communication networks [132]. We also suggested that these transitions might have physically observable implications for structures that grow based on the efficient delivery of information. We also found phase transitions and evidence of a critical point in ensembles of networks that are robust to random or directed attacks [250]. Finally, our framework has proved useful to guide us through models of Pareto optimal protein folding. In that problem, phase transitions have a definite physical nature. Our criteria clearly indicate that a series of foldings considered in the literature cannot be stable [78].

Hypothesis 8, **All possible communication codes (within a popular model) define a morphospace over which different communication systems can be located. Least effort communication poses a tradeoff that results in a critical first order phase transition. The complex nature of human language shall be captured by this critical point.** :

A toy model of associations between objects and signals that name

them allow us to study generic communication systems [112]. We explored the whole space characterizing the possible codes through a series of measurements such as the complexity that the codes can produce, or their network structure [288]. This allowed us to cluster communication codes together in a series of archetypes that lay segregated in the morphospace.

That toy model had been studied in the literature from an optimality perspective. Treating as targets the efforts made by *hearers* and *speakers* (i.e. by *coders* and *decoders*) a first order phase transition is revealed [112, 251, 265, 301, 286] between codes that implement a one-to-one mapping between objects and signals and codes that assign a same signal to every object. This transition is informally referred to as critical in the literature. Thanks to the Pareto formalism in chapter 2, we certified that this is indeed a critical transition.

A long held hypothesis, that we shared at the beginning of our work, is that human language lays at this transition and that it benefits from the complex structures that critical points contain. That would account for its flexibility and its capacity to generate complex structures. We assessed this conjecture by fitting real data from the WordNet database into the toy model. The results show that real languages without grammatical words fall in a region of code morphospace near to one-to-one mappings, often associated to animal communication systems [112, 286]. Adding grammatical words radically changes the locus of human languages in the morphospace, suggesting that they contain an important key in exploring alternative communication systems. While in all cases human codes lay fairly near the Pareto front of the efficient communication problem, the strict Pareto optimality condition (that there are not any synonyms) is not met. The complexity of codes near (but not on) the Pareto front suggest that remaining suboptimal may have beneficial effects in this case.

Hypothesis 9, **A tradeoff between accuracy and cost of prediction becomes relevant when modeling our physical reality – be it for the sur-**

vival of living beings or to increase our scientific knowledge. In the later case, a hierarchy of ontological ‘levels’ or ‘scales’ emerges along the corresponding Pareto front. Phase transitions become relevant in guiding the design of theoretical models:

We made a brief literature review of information theoretical advances in the identification of salient levels of description of symbolic dynamics. A series of important elements and methods (such as the information bottleneck models, computational mechanics, ϵ machines, causal shielding, or informational closure) came into focus and proved useful in tackling the tradeoff between the simplicity of the models and the amount of useful information that they can capture. Many of the works in this field present models that reconstruct the convex hull of underlying Pareto fronts. Phase transitions are mentioned in the literature [315, 317, 314] but their nature remains unknown since we do not have access to the whole Pareto front of the problems. Notwithstanding those transitions, along the tradeoff a series of models of increasing complexity are revealed. These are important in efficiently guiding the understanding of our physical reality.

We used the tools from this literature review to propose a research line in linguistics that attempts to find comprehensive yet simple models of the symbolic dynamics produced by human language. We hypothesized that the resulting hierarchy of models and the accidents that the corresponding Pareto front might show shall correspond to relevant levels of description. We further speculate that our methods can be useful for comparative studies and to pin down influences across different scales of language. The proposed theory shall also allows us to validate theories concerning the universality of protolanguage [33].

Both this preliminary analysis of language and the conjectures that we extracted are included to illustrate potential application of the work within this thesis. The promising results invite us to further

these studies, but we must be wary because the conclusions of this work might change as more layers of complexity are included.

Hypothesis 10, **A tradeoff involving replication efficiency and environment prediction complements Darwinian evolution explaining why efficient replicators are not the ultimate biological design**:

Evolution through natural selection may predict that extremely simple yet efficient replicators should wipe out any other (more complex, hence costly) kind of life. Empirical evidence (e.g. Spiegelman's monster [172]) shows that this is the case in certain controlled situations. We propose that an ability to predict the environment suffices as an explanation about why life scales up in complexity despite the associated replicative cost.

We illustrate this with a toy model that roots reproductive fitness in information theory. The model shows parsimoniously how life's complexity adapts to the environmental complexity thanks to the predictive capabilities of the living agents. These agents may coexist in certain environments or may exclude each other (hence generating a segregated hierarchy of complex life) depending on the complexity associated to each niche.

Hypothesis 11, **Evolution through natural selection is the origin of meaningful information in living systems**:

Daniel Dennett proposed *natural selection* as the origin of all the meaning that we perceive. He noted that *free-floating rationales* are reasons that existed before there were any reason representers [88], and he argues that natural selection is the mechanism that first captures those *reasons*. Eventually we, as reason representers, come to know these rationales and recognize their meaning.

The toy mathematical model introduced in chapter 6 illustrates these points. Reproducing species are modeled within a Shannon-like scheme as emitting a message across generations, with the changing environment as the channel conditions. Messages that better

deal with the environmental (i.e. channel) conditions get reproduced faster. In our toy model this is pictured as a dynamics that fixes meaningful bits into the transmitted messages. Coming closer to real biology, meaning would be embodied in the genotypes transmitted between generations. Wondering how that meaning came about, we note that it must originate from the selection imposed by environmental conditions alone.

Appendix A

APPENDICES

A.1. Analytic and numeric approaches for MOO solving

We relied on genetic algorithms to locate the different Pareto fronts. The fully topological case can be solved analytically, but the same genetic algorithm was used to check for good convergence showing very good results. In this appendix we explain in detail the genetic algorithm. Following that explanation, a few disclaimers are in order about the numerical nature of the solutions found – i.e. about the fact that convergence to the Pareto front cannot be guaranteed and the smoothing necessary to render a continuous approximation to the front.

A.1.1. A multiobjective genetic algorithm

MOO relays on the concept of Pareto dominance. Given two networks $\gamma_i, \gamma_j \in \Gamma$, both mapped into $\mathbb{R}^2$ through $t_1(\gamma_{i/j})$ and $t_2(\gamma_{i/j})$, we say that γ_i dominates γ_j (and note it $\gamma_i \prec \gamma_j$) if γ_i is not worst than γ_j regarding any target and it is better than γ_j with respect to at least one target. We can visualize this: Since we deal with minimizations in the $t_1 - t_2$ plane, network γ_i has got a set of axes associated with their origin at $(t_1(\gamma_i), t_2(\gamma_i))$

and every network γ_j laying on the first quadrant of these axes is Pareto dominated by γ_i.

Following the literature on multiobjective genetic algorithms [93, 352, 181] we computed a dominance score: We took the set $D_j \equiv \{\gamma_i | \gamma_i \prec \gamma_j\}$ of solutions from within a given *population* (an arbitrary subset of Γ) that dominate γ_j. The size of this set ($d_j = ||D_j||$, dubbed the *dominance score*), indicates how *fit* γ_j is in terms of Pareto optimality. We proceeded then to minimize this score. We departed from an initial population of N_P networks (either random or designed, see below), selected $N_P/2$ of the population based on the dominance score, chose random pairs among the selected networks to produce $N_P/2$ new networks, and applied random mutations to all but a subset of *elite* networks. We iterated this scheme a fixed number of generations.

Mutations consisted in random appending or deleting edges or totally swapping the connections of two arbitrary nodes. For crossover, from each of the two mating nets we assigned each node and its connections randomly to each of the offspring graphs checking that the same node and connection was not assigned twice to the same *child*. After crossover or edge deletion we checked that all networks remained connected all the time. We completed one missing link whenever connectedness failed and then checked for connectedness again.

All we care about for the current research is good convergence towards the front. This justifies our using of crossover: this is a very good evolutionary operator, though unrealistic if we wanted to study some features of nature. For example, such an operator would not be adequate to study species that do not reproduce sexually. Studying Pareto optimality under constrained conditions – e.g. without crossover – also renders a set of non-dominated solutions. These might converge to the Pareto front or not, and they might be subjected to geometrical constraints in $t_1 - t_2$ that are similar to those studied in [282] and in this paper for Pareto optimal networks. Such constrained evolutionary schemes pose interesting research questions, but here we are concerned with Pareto optimal solutions. This justifies the crossover and a clever initialization of the algorithm. It might be difficult to converge towards some solutions that are

highly non-trivial – e.g. the MST. We know, though, that this solution belongs to the front of all physically grounded problems. If an algorithm would fail in finding this solution, this could hinder convergence towards an interesting (though challenging) region of the Pareto front. Once again, because we are concerned with Pareto optimal solutions and would like to attain the closest convergence possible, it is also fully justified to *seed* the initial population with a few *designed* solutions. We did so by introducing since the very beginning cliques, star-graphs, MSTs, and circle networks (these two are equivalent for the circle) with very slight mutations. We produced $N_P/4$ of each such major topologies at the beginning. The crossover and mutation operators ensure fast exploration of hybrid topologies.

As noted above, we know that some of these networks are Pareto optimal: the clique is always so, and the MST is Pareto optimal in all physically grounded problems. However, and since we started with little variations upon these graphs, the algorithm did not always reach these solutions – but it surely explored the region nearby. There might be other interesting regions of the front that might not have been fully explored and that are impossible to seed without foreknowledge. Although we are concerned with Pareto optimal solutions alone, our methods are numerical at the end and convergence of multiobjective genetic algorithms to the Pareto front cannot be guaranteed. We decided to report on the results of the simulations with as little reinterpretation and further speculation as possible. Notwithstanding, the overall details of the Pareto front seem to be recovered and the theory posed in [282] is properly illustrated.

As for the implementation of the algorithm, we used a population of $N_P = 3000$ connected networks with $N = 50$ nodes – with the initial population seeded as indicated above. The population was evolved during $T = 10000$ generations in every case. Mutation happened with a probability $p_\mu = 0.001$ of appending an extra link to each network, the same probability of deleting an existing edge, and the same probability of swapping the ends of each existing connection. The top $N_e = 50$ networks of the population where considered *elite* and were spared any mutation. As the

algorithm proceeds, many networks reach a dominance score of 0 even if they are not Pareto optimal. Unluckily, this score is the best indicator of Pareto optimality available (not only in the current implementation, but generally). This results in elite members of the population not being objectively better than non-elite members – in terms of Pareto optimality. Because the algorithm sorted the population consistently from one generation to the next, what members of the population are considered elite is largely a matter of antiquity: early members that reach low dominance score and are not overthrown are likely to be preserved during the whole simulation. We repeated 4 times the simulations with scattered nodes to check that the relevant features obtained were not artifacts of some lucky distributions of the nodes.

A.1.2. Smoothing of the front and order parameters

The only speculation that we allowed ourselves is in choosing a relevant scale for analysis, which led to a smoothing of the Pareto front. As noted in Sec. 3.1.3, we deal with a discrete set of networks whose front is necessarily discrete as well. Accordingly, every shift in global optima is a first order phase transition at some scale and global optima remain so for a continuous range of λ, as in second order phase transitions. This does not further our understanding of the problem as much as a coarse-grained analysis that renders noteworthy phase transitions. Since the genetic algorithm only produces a finite set of (ideally) Pareto optimal solutions, we applied a Bezier smoothing to their plot on the t_1 – t_2 plane. We took care that the smoothing did not introduce alien concavities. Because Bezier curves cannot present sharp edges (thus ruling out second order phase transitions), when a sharp edge seemed the best description of the front (section 3.1.3, partly geometrical problem on nodes scattered over a plane), we decided to split the front in its two salient branches and apply two independent smoothing processes that allowed us to recover the transition in great detail.

To locate global optima, we calculated $\Omega(x_\Pi, \lambda)$ for the optimal solutions produced by the genetic algorithm, and for a large sample of points

from the Bezier curves introduced in the previous paragraph. We registered the global optimal for different values of λ. One of the problems pointed at earlier is that, because of the discreteness of the front, global optima are so for several values of λ. This would cause that the plots of order parameters look tiered. For a better illustration of the results, whenever order parameters are plotted we indicate only the first and last values of λ for which each global optima are indeed optima (black crosses in all order parameter plots). The smoothing allows a finner grained sampling so that this is not an issue: the corresponding order parameters (red curves in all order parameter plots) look continuous always.

Following [282], anything *well behaved* that we measure upon global optima are accepted as order parameters. By well behaved we imply that order parameters should not introduce alien divergences into the problem, and that solutions laying at different points over the front should score differently in this parameter. This way we ensure that any feature stemming from the optimization problem does not go unreported and that we do not introduce phase-transition–like behaviors that originated, e.g., on some function diverging to infinity for reasons of its own. Taking these guidelines into account, the target functions themselves are always good order parameters. We use these ($\theta = t_1$ in Sec. 3.1.3), or trivial transformations of them ($\theta = 1 - t_2$ in Sec. 3.1.3). More drastic transformations such as $1/(2 - t_1)$ would be banned: note that this function diverges for $t_1 = 2$ even if this is a perfectly regular point of the front for all problems.

Bibliography

[1] Abeles, M., 2012. *Local cortical circuits: An electrophysiological study* (Vol. 6). Springer Science & Business Media.

[2] Adami, C., 2012. The use of information theory in evolutionary biology. *Ann. N. Y. Acad. Sci.*, **1256**(1), pp.49-65.

[3] Aicardi, F., 2001. On the classification of singularities in thermodynamics. *Phys D*. **158**, 175-196.

[4] Albert, R., Jeong, H. and BarabÃ¡si, A.L., 2000. Error and attack tolerance of complex networks. *Nature*, **406**(6794), pp.378-382.

[5] Albert, R. and Barabási, A.L., 2002. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, **74**(1), p.47.

[6] Altmann, E.G., Cristadoro, G. and Degli Esposti, M., 2012. On the origin of long-range correlations in texts. *Proc. Natl. Acad. Sci.*, **109**(29), pp.11582-11587.

[7] Amaral, L.A.N., Scala, A., Barthélémy, M., and Stanley, H.E., 2000. Classes of small-world networks. *P. Nat. A. Sci.*, **97**(21), p.11149.

[8] Arrow, K.J., Debre, G., 1954. Existence of an equilibrium for a competitive economy. *Econometrica* **22**(3), pp.265-289.

[9] Arrow, K.J., 1963. Uncertainty and the welfare economics of medical care. *Am. Econ. Rev.*, **53**(5), pp.941-973.

[10] Asselmeyer, T., Ebeling, W., Rosé, H., 1997. Evolutionary strategies of optimization. *Phys Rev E.* **56**(1), pp.1171-1180.

[11] Auerbach, F., 1913. Das Gesetz der Bevölkerungskonzentration. *Petermans Geographische Mitteilungen* **59**, pp.74-76.

[12] Avena-Koenigsberger, A., Goñi, J., Solé, R., and Sporns, O., 2015. Network morphospace. *J. R. Soc. Interface*, **12**(103), 20140881.

[13] Baixeries, J., Elvevåg, B. and Ferrer-i-Cancho, R., 2013. The evolution of the exponent of Zipf's law in language ontogeny. *PLoS one*, **8**(3), p.e53227.

[14] Bak, P., Tang, C., Wiesenfeld, K., 1987. Self-organized criticality: An explanation of the 1/f noise. *Phys. Rev. Lett.* **59**(4), p.381.

[15] Bak, P., Tang, C., Wiesenfeld, K., 1988. Self-organized criticality. *Phys. rev. A*, **38**(1), p.364.

[16] Bak, P., Sneppen, K., 1993. Punctuated equilibrium and criticality in a simple model of evolution. *Phys Rev Let*, **71**(24), 4083.

[17] Bak, P., 1996. *How nature works: the science of self-organized criticality*. Springer Science & Business Media.

[18] Balduzzi, D. and Tononi, G., 2009. Qualia: the geometry of integrated information. *PLoS. Comput. Biol.*, **5**(8), p.e1000462.

[19] Balian, R., 2006. *From microphysics to macrophysics: methods and applications of statistical physics (Vol. 2)*. Springer Science & Business Media.

[20] Ball, P., 2004. *Critical mass: How one thing leads to another*. Macmillan; 2004.

[21] Banavar, J.R., Maritan, A., and Rinaldo, A., 1999. Size and form in efficient transportation networks. *Nature*, **399**(6732), p.130.

[22] Banga, J.R., 2008. Optimization in computational systems biology. *BMC Syst. Biol.*, **2**, p.47.

[23] BarthÃ©lemy, M., 2011. Spatial networks. *Phys. Rep.* **499**(1), pp.1-101.

[24] Bartolini, R., Apollonio, M. and Martin, I.P.S., 2012. Multiobjective genetic algorithm optimization of the beam dynamics in linac drivers for free electron lasers. Physical Review Special Topics-Accelerators and Beams, 15(3), p.030701.

[25] Bassett, D.S., Greenfield, D.L., Meyer-Lindenberg, A., Weinberger, D.R., Moore, S.W., Bullmore, T., 2010. Efficient Physical Embedding of Topologically Complex Information Processing Networks in Brains and Computer Circuits. *PLoS Comput. Biol.*, **6**(4), e1000748.

[26] Baxter, G.J., Dorogovtsev, S.N., Goltsev, A.V., Mendes, J.F., 2010. Bootstrap percolation on complex networks. *Phys. Rev. E*, **82**(1), 011103.

[27] Beggs, J.M., Plenz, D., 2003. Neuronal avalanches in neocortical circuits. *J Neurosci.* **23**(35), pp.11167-11177.

[28] Bergstrom, C.T. and Lachmann, M., 2004. Shannon information and biological fitness. In Information Theory Workshop, 2004. IEEE (pp.50-54).

[29] Bertalan, Z., Kuma, T., Matsuda, Y., Nishimori, H., 2011. Ensemble Inequivalence in the Ferromagnetic p-spin Model in Random Fields. *J. Stat. Mech. 1*, P01016.

[30] Beuls, K. and Steels, L., 2013. Agent-based models of strategies for the emergence and evolution of grammatical agreement. *PLoS one*, **8**(3), p.e58960.

[31] Bialek, W., 2013. *Biohysics: searching for principles*. Princeton University Press.

[32] Bickerton, D., 1992. *Language and species*. University of Chicago Press.

[33] Bickerton, D., 2014. *More than Nature Needs*. Harvard University Press.

[34] Blaug, M., 2007. The fundamental theorems of modern welfare economics, historically contemplated. *History of Political Economy* **39**(2), 185-207.

[35] Bonachela, J.A., De Franciscis, S., Torres, J.J. and Muñoz, M.A., 2010. Self-organization without conservation: are neuronal avalanches generically critical?. *J. Stat. Mech.*, **2010**(02), p.P02015.

[36] Bonner, J.T., 1988. *The evolution of complexity by means of natural selection*. Princeton University Press.

[37] Borge-Holthoefer, J., Moreno, Y. and Arenas, A., 2011. Modeling abnormal priming in Alzheimer's patients with a free association network. *PLoS one*, **6**(8), p.e22651.

[38] Bornholdt, S., 1998. Genetic algorithm dynamics on a rugged landscape. *Phys Rev E.*, **57**(4), 3853-3860.

[39] Bouchet, F., Barr, J., 2008. Classification of phase transitions and ensemble inequivalence, in systems with long range interactions. *J. Stat, Phys.*, **118**(5-6), pp.1073-1105.

[40] Branke, J., Deb, K., Dierolf, H. and Osswald, M., 2004. Finding knees in multi-objective optimization. In Parallel Problem Solving from Nature-PPSN VIII (pp. 722-731). Springer Berlin Heidelberg.

[41] Brodu, N., 2011. Reconstruction of epsilon-machines in predictive frameworks and decisional states. *Adv. Complex Syst.*, **14**(05), pp.761-794.

[42] http://nicolas.brodu.net/recherche/decisional_states/index.html

[43] Buldyrev, S.V., Parshani, R., Paul, G., Stanley, H.E. and Havlin, S., 2010. Catastrophic cascade of failures in interdependent networks. *Nature*, **464**(7291), pp.1025-1028.

[44] Callaway, D.S., Newman, M.E., Strogatz, S.H. and Watts, D.J., 2000. Network robustness and fragility: Percolation on random graphs. *Phys. Rev. Let.*, **85**(25), p.5468.

[45] Carlson, J.M., Doyle, J., 1999. Highly optimized tolerance: A mechanism for power laws in designed systems. *Phys. Rev. E*, **60**(2), p.1412.

[46] Carlson, J.M., Doyle, J., 2000. Highly optimized tolerance: Robustness and design in complex systems. *Phys. Rev. Lett.* **84**(11), 2529.

[47] Carvalho, R., Buzna, L., Bono, F., Gutiérrez, E., Wolfram, J., Arrowsmith, D., 2009. Robustness of trans-European gas networks *Phys. Rev. E.* **80**, 016106.

[48] Castellano, C., Fortunato, S., Loreto, V., 2009. Statistical physics of social dynamics. *Rev. Mod. Phys.*, **81**(2), p.591.

[49] Chan, A.S., Butters, N. and Salmon, D.P., 1997. The deterioration of semantic networks in patients with Alzheimer's disease: A cross-sectional study. *Neuropsychologia*, **35**(3), pp.241-248.

[50] Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M., 1983. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, **4**(2), pp.187-217.

[51] Brooks, B.R., Brooks, C.L., MacKerell, A.D., Nilsson, L., Petrella, R.J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S. and Caflisch, A., 2009. CHARMM: the biomolecular simulation program. *J. Comput. Chem.*, **30**(10), pp.1545-1614.

[52] Chen, W.K., 1999. *The VLSI Handbook* CRC Press, Boca Raton, Fl.

[53] Chertkow, H., Bub, D. and Seidenberg, M., 1989. Priming and semantic memory loss in Alzheimer's disease. *Brain. Lang.*, **36**(3), pp.420-446.

[54] Chomsky, N., 1956. Three models for the description of language. *IEEE T. Inform. Theory*, **2**(3), pp.113-124.

[55] Chomsky, N., 1998. *Minimalist inquiries: The framework (No. 15)*. MIT Working Papers in Linguistics, MIT, Department of Linguistics.

[56] Chomsky, N., 2002. An interview on minimalism. *N. Chomsky, On Nature and Language*, pp.92-161.

[57] Clauset, A., Shalizi, C.R. and Newman, M.E., 2009. Power-law distributions in empirical data. *SIAM Rev.*, **51**(4), pp.661-703.

[58] Clune, J., Mouret, J.B., and Lipson, H., 2013. The evolutionary origins of modularity. *Proc. R. Soc. B*, **280**(1755), 20122863.

[59] http://csc.ucdavis.edu/~chaos/courses/ncaso/Software/

[60] http://wissenplatz.org/compmech.html

[61] http://corpus.byu.edu/coca/

[62] Coello, C.A., 2006. Evolutionary Multi-Objective Optimization: A Historical View of the Field. **IEEE Comput. Intell. M**. **1**(1), pp.28-36.

[63] Colizza, V., Banavar, J.R., Maritan, A., Rinaldo, A., 2004. Network Structures from Selection Principles. *Phys. Rev. Lett.*, **92**(19), pp.198701-1.

[64] Corominas-Murtra, B., Valverde, S. and Solé, R., 2009. The ontogeny of scale-free syntax networks: phase transitions in early language acquisition. *Adv. Complex Syst.*, **12**(03), pp.371-392.

[65] Corominas-Murtra, B., Rodríguez-Caso, C., Goñi, J. and Solé, R., 2010. Topological reversibility and causality in feed-forward networks. *New J. Phys.*, **12**(11), p.113051.

[66] Corominas-Murtra, B. and Solé, R.V., 2010. Universality of Zipf's law. *Phys. Rev. E*, **82**(1), p.011102.

[67] Corominas-Murtra, B., Fortuny, J. and Solé, R.V., 2011. Emergence of Zipf's law in the evolution of communication. *Phys. Rev. E*, **83**(3), p.036115.

[68] Corominas-Murtra, B., Goñi, J., Solé, R.V. and Rodríguez-Caso, C., 2013. On the origins of hierarchy in complex networks. *Proc. Natl. Acad. Sci.*, **110**(33), pp.13316-13321.

[69] Corominas-Murtra, B., Fortuny, J. and SolÃ©, R.V., 2014. Towards a mathematical theory of meaningful communication. *Sci. Rep.*, **4**, 4587.

[70] Corominas-Murtra, B., Hanel, R. and Thurner, S., 2015. Understanding scaling through history-dependent processes with collapsing sample space. *Proc. Natl. Acad. Sci.*, 112(17), pp.5348-5353.

[71] Costeniuc, M., Ellis, R.S., Touchette, H., 2005. Complete Analysis of Phase Transitions and Ensemble Equivalence for the Curie-Weiss-Potts Model. *J. Math. Phys.* **46**, 063301.

[72] Cover, T.M. and Thomas, J.A., 2012. *Elements of information theory*. John Wiley & Sons.

[73] Crutchfield, J.P. and Young, K., 1989. Inferring statistical complexity. *Phys. Rev. Let.*, **63**(2), p.105.

[74] Crutchfield, J.P., 1994. The calculi of emergence: computation, dynamics and induction. *Physica D*, **75**(1), pp.11-54.

[75] Crutchfield, J.P. and Shalizi, C.R., 1999. Thermodynamic depth of causal states: Objective complexity via minimal representations. *Phys. Rev. E*, **59**(1), p.275.

[76] Cuntz, H., Forstner, F., Borst, A., Häusser, M., 2010. One Rule to Grow Them All: A General Theory of Neuronal Branching and Its Practical Application. *PLoS. Comput. Biol.*, **6**(8), e1000877.

[77] Cuntz, H., Borst, A., Segev, I., 2007. Optimization principles of dendritic structure. *Theor. Biol. Med. Model.*, **4**, p.21.

[78] Cutello, V., Narzisi, G., Nicosia, G., 2006. A multi-objective evolutionary approach to the protein structure prediction problem. *J. R. Soc. Interface.*, **3**(6), pp.139-151.

[79] da Fontoura Costa, L., Zawadzki, K., Miazaki, M., Viana, M.P., and Taraskin, S.N., 2010. Unveiling the neuromorphological space. *Frontiers Comput. Neurosci.* **4**, 150.

[80] Dall, S.R. and Johnstone, R.A., 2002. Managing uncertainty: information and insurance under the risk of starvation. Phil. T. Roy. Soc. B, **357**(1427), pp.1519-1526.

[81] Dall, S.R., Giraldeau, L.A., Olsson, O., McNamara, J.M. and Stephens, D.W., 2005. Information and its use by animals in evolutionary ecology. *Trends Ecol. Evol.*, **20**(4), pp.187-193.

[82] Darmon, D., Omodei, E., Flores, C.O., Seoane, L.F., Stadler, K., Wright, J., Garland, J. and Barnett, N., 2013. Detecting communities using information flow in social networks. In *Proc. of the CSSS*, Santa Fe Institute.

[83] Das, I., 1999. On characterizing the “knee” of the Pareto curve based on normal-boundary intersection. *Struct. Optimization*, **18**(2-3), pp.107-115.

[84] Dawkins, R., 1986. *The blind watchmaker: Why the evidence of evolution reveals a universe without design*. WW Norton & Company.

[85] R. Dawkins, *Climbing mount improbable* (WW Norton & Company, 1997).

[86] Delgado, J. and Solé, R.V., 1997. Noise induced transitions in fluid neural networks. *Phys. Let. A*, **229**(3), pp.183-189.

[87] Dennett, D.C., 1995. *Darwin's dangerous idea.* Touchstone, New York; 1995.

[88] Dennett, D.C., 2013. The Evolution of Reasons. Contemporary philosophical naturalism and its implications, 13, p.47.

[89] Denysiuk, R., Silva, C.J., Torres, D.F.M., 2014. Multiobjective approach to optimal control for tuberculosis model. *Optim. Method. Soft.* doi: 10.1080/10556788.2014.994704

[90] De Weck, O.L., 2004. Multiobjective optimization: History and promise. In Invited Keynote Paper, GL2-2, The Third China-Japan-Korea Joint Symposium on Optimization of Structural and Mechanical Systems, Kanazawa, Japan (Vol. 2, p. 34).

[91] Dickman, R., Moloney, N.R., Altmann, E.G., 2012. Analysis of an information-theoretic model for communication. *J. Stat. Mech.* **2012**(12): P12022.

[92] Dickman, R., Muñoz, M.A., Vespignani, A. and Zapperi, S., 2000. Paths to self-organized criticality. *Braz. J. Phys.*, **30**(1), pp.27-41.

[93] Dittes, F.M., 1996. Optimization on Rugged Landscapes: A New General Purpose Monte Carlo Approach. *Phys. Rev. Lett.* **76**(25): pp.4651-4655.

[94] Donaldson-Matasci, M.C., Lachmann, M. and Bergstrom, C.T., 2008. Phenotypic diversity as an adaptation to environmental uncertainty. *Evol. Ecol. Res.*, **10**(4), pp.493-515.

[95] Donaldson-Matasci, M. C., Bergstrom, C. T., & Lachmann, M 2010 The fitness value of information. *Oikos*, **119**(2), 219-230.

[96] Dorogovtsev, S.N., Goltsev, A.V., Mendes, J.F.F., 2006. K-core organization of complex networks. *Phys. Rev. Lett.* **96**(4), 040601.

[97] Doyle, J.C., Alderson, D.L., Li, L., Low, S., Roughan, M., Shalunov, S., Tanaka, R. and Willinger, W., 2005. The “robust yet fragile” nature of the Internet. *Proc. Natl. Acad. Sci.*, **102**(41), pp.14497-14502.

[98] Drossel, B. 2001 Biological evolution and statistical physics. *Adv. Phys.* **50**(2), 209-295.

[99] Druckmann, S., Banitt, Y., Gidon, A., Schürmann, F., Markram, H., Segev, I., 2007. A novel multiple objective optimization framework for constraining conductance-based neuron models by experimental data. *Front. Neurosci.* **1**(1): 7-18.

[100] Edelman, G.M. and Mountcastle, V.B., 1978. *The mindful brain: Cortical organization and the group-selective theory of higher brain function*. Massachusetts Inst of Technology Pr.

[101] Edgeworth, F.Y., 1881. “Mathematical Psychics: An Essay on the Application of Mathematics to the Moral Sciences.” London: Kegan Paul.

[102] https://en.wikipedia.org/wiki/Francis_Ysidro_Edgeworth

[103] Einstein, A., 1916. Die grundlage der allgemeinen relativitätstheorie. *Annalen der Physik*, **354**(7), pp.769-822.

[104] Ellis, R.S., Haven, K. and Turkington, B., 2000. Large deviation principles and complete equivalence and nonequivalence results for pure and mixed ensembles. *J. Stat. Phys.*, **101**(5-6), pp.999-1064.

[105] England, J.L., 2013. Statistical physics of self-replication. *The Journal of chemical physics*, **139**(12), p.121923.

[106] Evans, J.C., Votier, S.C. and Dall, S.R., 2015. Information use in colonial living. *Biol. Rev.*, DOI: 10.1111.

[107] Fellbaum, Christiane, ed., 1998. WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

[108] Ferrer i Cancho, R. and Solé, R.V., 2001. The small world of human language. *Proc. R. Soc. B*, **268**(1482), pp.2261-2265.

[109] Ferrer i Cancho, R. and Solé, R.V., 2001. Two Regimes in the Frequency of Words and the Origins of Complex Lexicons: Zipf's Law Revisited. *J. Quant. Linguist.*, **8**(3), pp.165-173.

[110] Ferrer i Cancho, R. and Solé, R.V., 2002. Zipf's law and random texts. *Adv. Complex Syst.*, **5**(01), pp.1-6.

[111] Ferrer i Cancho, R. and Solé, R.V., 2003. Optimization in complex networks. In Statistical mechanics of complex networks (pp.114-126). Springer Berlin Heidelberg.

[112] Ferrer i Cancho, R. and Solé, R.V., 2003. Least effort and the origins of scaling in human language. *Proc. Natl. Acad. Sci.*, **100**(3), pp.788-791.

[113] Ferrer i Cancho, R.F., Solé, R.V. and Köhler, R., 2004. Patterns in syntactic dependency networks. *Phys. Rev. E*, **69**(5), p.051915.

[114] Ferrer i Cancho, R., 2005. The variation of Zipf's law in human language. *Eur. Phys. J. B*, **44**(2), pp.249-257.

[115] Ferrer i Cancho, R.F., 2005. Decoding least effort and scaling in signal frequency distributions. *Physica A*, **345**(1), pp.275-284.

[116] Ferrer i Cancho, R., Riordan, O. and Bollobás, B., 2005. The consequences of Zipf's law for syntax and symbolic reference. *Proc. R. Soc. B*, **272**(1562), pp.561-565.

[117] Ferrer i Cancho, R.F. and Díaz-Guilera, A., 2007. The global minima of the communicative energy of natural communication systems. *J. Stat. Mech.*, **2007**(06), p.P06009.

[118] Flanagan, T.P., Letendre, K., Burnside, W., Fricke, G.M. and Moses, M., 2011. How ants turn information into food. In Artificial Life (ALIFE), 2011 IEEE Symposium on (pp. 178-185). IEEE.

[119] Fonseca, C.M., Fleming, P.J., 1995. An Overview of Evolutionary Algorithms in Multiobjective Optimization. *Evol. Comput.* **3**: pp.1-16.

[120] Fontana, W., Schnabl, W., Schuster, P., 1989. Physical aspects of evolutionary optimization and adaptation. *Phys Rev E.* **40**, pp.3301-3321.

[121] Fortuny, A. J. and Corominas-Murtra, B., 2009. Some formal considerations on the generation of hierarchically structured expressions. *Cat. J. L.*, **8**, pp.099-111.

[122] Fortuny, J. and Corominas-Murtra, B., 2013. On the origin of ambiguity in efficient communication. *J. Logic Lang. Inform.*, **22**(3), pp.249-267.

[123] Frank, A. and Jaeger, T.F., 2008. Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the 30th annual meeting of the cognitive science society* (pp. 933-938). Washington, DC: Cognitive Science Society.

[124] Friston, K. 2013 Life as we know it. *J. R. Soc. Interface*, **10**(86), 20130475.

[125] Gafiychuk, V.V., Lubashevsky, I.A., 2001. On the Principles of the Vascular Network Branching. *J. Theor. Biol.* **212**, pp.1-9.

[126] Gastner, M.T., Newman, M.E.J., 2006. Optimal design of spatial distribution networks. *Phis. Rev. E.* **74**, 016117.

[127] Gent, I.P, Walsh, T., 1996 The TSP phase transition, *Artif. Intell.* **88**, pp.349-358.

[128] Gibbs, J.W. , 1873. Graphical Methods in the Thermodynamics of Fluids. *Trans. Conn. Acad.* **2**, pp.309-342.

[129] Gibbs, J.W. , 1873. A Method of Geometrical Representation of the Thermodynamic Properties of Substances by Means of Surfaces. *Trans. Conn. Acad.* **2**, pp.382-404.

[130] Goldenfeld, N., & Woese, C., 2010. Life is physics: evolution as a collective phenomenon far from equilibrium. arXiv preprint arXiv:1011.4125.

[131] Goñi, J., Arrondo, G., Sepulcre, J., Martincorena, I., de Mendizábal, N.V., Corominas-Murtra, B., Bejarano, B., Ardanza-Trevijano, S., Peraita, H., Wall, D.P. and Villoslada, P., 2011. The semantic organization of the animal category: evidence from semantic verbal fluency and network theory. *Cogn. Process.*, **12**(2), pp.183-196.

[132] Goñi, J., Avena-Koenigsberger, A., de Mendizabal, N.V., van den Heuvel, M., Betzel R., and Sporns, O., 2013. Exploring the morphospace of communication efficiency in complex networks. *PLoS ONE* **8**, e58070.

[133] Gordon, D.M., 1999. *Ants at work: how an insect society is organized*. Simon and Schuster.

[134] Görnerup, O. and Jacobi, M.N., 2010. A method for finding aggregated representations of linear dynamical systems. *Adv. Complex Syst.*, **13**(02), pp.199-215.

[135] Gould, S. J. 2011 *Full house*. Harvard, MA: Harvard University Press.

[136] Gregory, T.R., 2008. The evolution of complex organs. *Evol. Ed. Outreach*, **1**(4), pp.358-389.

[137] Griffith, J.S., 1963. On the stability of brain-like structures. *Biophys. J.*, **3**(4), p.299.

[138] Grimme, C., Lepping, J., Papaspyrou, A., 2012. Parallel predator-prey interaction for evolutionary multi-objective optimization. *Nat. Comput.* **11**, pp.519-533.

[139] Haimovici A, Tagliazucchi E, Balenzuela P, Chialvo DR, Brain organization into resting state networks emerges at criticality on a model of the human connectome. *Phys Rev Let* **110**(17), 178101 (2013).

[140] Handl, J. and Knowles, J., 2005. Exploiting the trade-off – the benefits of multiple objectives in data clustering. In Evolutionary Multi-Criterion Optimization (pp. 547-560). Springer Berlin Heidelberg.

[141] Hart, Y., Sheftel, H., Hausser, J., Szekely, P., Ben-Moshe, N.B., Korem, Y., Tendler, A., Mayo, A.E. and Alon, U., 2015. Inferring biological tasks using Pareto analysis of high-dimensional data. **Nat. Methods**, **12**(3), pp.233-235.

[142] Harte J. Maximum entropy and ecology: a theory of abundance, distribution, and energetics. Oxford University Press; 2011.

[143] Hasenstaub A, Otte S, Callaway E, Sejnowski TJ. Metabolic cost as a unifying principle governing neuronal biophysics. P Nat A Sci. 2010;107(27): 12329-12334.

[144] Hauser, M.D., Chomsky, N. and Fitch, W.T., 2002. The faculty of language: what is it, who has it, and how did it evolve? *Science*, **298**(5598), pp.1569-1579.

[145] Hawkins, J. and Blakeslee, S., 2007. *On intelligence*. Macmillan.

[146] Hidalgo, J., Grilli, J., Suweis, S., Muñoz, M.A., Banavar, J.R. and Maritan, A., 2014. Information-based fitness and the emergence of criticality in living systems. *Proc. Nat. Acad. Sci.*, **111**(28), pp.10095-10100.

[147] Hidalgo, J., Pigolotti, S. and Muñoz, M.A., 2015. Stochasticity enhances the gaining of bet-hedging strategies in contact-process-like dynamics. *Physical Review E*, **91**(3), p.032114.

[148] Higuera C, Villaverde AF, Banga JR, Ross J, Mora F. Multi-Criteria Optimization of Regulation in Metabolic Networks. PLoS ONE. 2012;7(7): e41122.

[149] Hilbert, M., 2015. Fitness as Informational Fit: The Communication Channel between the Evolving Population and Its Environment. Available at SSRN 2619963.

[150] Hines, P., Cotilla-Sanchez, E. and Blumsack, S., 2010. Do topological models provide good information about electricity infrastructure vulnerability?. *Chaos*, **20**(3), p.033122.

[151] Dennett, D.C. and Hofstadter, D., 1981. *The mind's I. Fantasies and Reflections on Self and Soul*. Basic Books, New York.

[152] Dennett, D.C. and Hofstadter, D., 1981. *The mind's I. Fantasies and Reflections on Self and Soul*. Basic Books, New York. Ch. 11, pp.149-190.

[153] Hogg T, Huberman BA, Williams CI. Phase transitions and the search problem. Artif Intell. 1996;81: 1-15.

[154] Holanda, A. de J., Torres, I., Kinouchi, O, Souto, A., Seron, E. E., 2004. Thesaurus as a complex network. *Physica A*, **344**(3), pp.530-536.

[155] Hopfield, J. J. 1988 Artificial neural networks. *IEEE Circuits Devices Mag.*, **4**(5), 3-10.

[156] Huang, K., 1987. *Statistical Mechanics, 2nd. Edition*. New York: John Wiley & Sons.

[157] Hubel, D.H. and Wiesel, T.N., 1962. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.*, **160**(1), pp.106-154.

[158] Hubel, D.H. and Wiesel, T.N., 1968. Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.*, **195**(1), pp.215-243.

[159] Huth, A.G., Nishimoto, S., Vu, A.T. and Gallant, J.L., 2012. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, **76**(6), pp.1210-1224.

[160] http://gallantlab.org/semanticmovies/

[161] Israeli, N. and Goldenfeld, N., 2006. Coarse-graining of cellular automata, emergence, and the predictability of complex systems. *Phys. Rev. E*, **73**(2), p.026203.

[162] Jablonka E and Lamb MJ. 2006. The evolution of information in the major transitions. *J. Theor. Biol.* **239**, 236-246.

[163] Jacob, F. 1998 *On flies, mice and man*. Harvard, MA: Harvard University Press.

[164] Jaeger, T.F., 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychol.*, **61**(1), pp.23-62.

[165] Jaeger, T.F. and Levy, R.P., 2006. Speakers optimize information density through syntactic reduction. In *Advances in neural information processing systems* (pp. 849-856).

[166] Jaynes ET, Information theory and statistical mechanics. *Phys. Rev.* **106**, 620-630 (1957).

[167] Jaynes ET, Information theory and statistical mechanics. II. *Phys. rev.*, **108**(2), 171 (1957).

[168] Jensen, F. V. 1996 *An introduction to Bayesian networks*. London: UCL press.

[169] Stanley, J.W., 1911. “Theory of Political Economy.” The MacMillan company, London, England.

[170] Joyce, G. F. 2002 Molecular evolution: Booting up life. *Nature* **420**, 278-279. doi:10.1038/420278a

[171] Joyce, G. F. 2012 Bit by Bit: The Darwinian Basis of Life. *PLoS Biol.* **10**(5), e1001323. doi:10.1371/journal.pbio.1001323

[172] Kacian, D. L., Mills, D. R., Kramer, F. R., & Spiegelman, S. 1972 A replicating RNA molecule suitable for a detailed analysis of extracellular evolution and replication. *Proc. Nat. Acad. Sci.*, **69**(10), 3038-3042.

[173] K. J. Kansky, PhD thesis, University of Chicago, 1963. *Structure of transportation networks: relationships between network geometry and regional characteristics*, PhD Thesis,

[174] Karrer, B. and Newman, M.E., 2011. Stochastic blockmodels and community structure in networks. *Phys. Rev. E*, **83**(1), p.016107.

[175] Stuart A. Kauffman, 1993. *The origins of order: Self organization and selection in evolution*. Oxford University Press, USA.

[176] Ke, J., 2004. Self-organization and language evolution: system, population and individual (Doctoral dissertation, City University of Hong Kong).

[177] Kennedy MC. Functional-structural models optimize the placement of foliage units for multiple whole-canopy functions. Ecol. Res. 2010;25: 723-732.

[178] Kihara T, Midzuno Y, Shizume T. Statistics of Two-Dimensional Lattices with Many Components. J Phys Soc Jpn. 1954;9(5): 681-687.

[179] Kinouchi, O., Martinez, A.S., Lima, G.F., LourenÃ§o, G.M. and Risau-Gusman, S., 2002. Deterministic walks in random networks: An application to thesaurus graphs. *Physica A*, **315**(3), pp.665-676.

[180] Knoll, A.H., 2011. The multiple origins of complex multicellularity. *Annu. Rev. Earth Planet. Sci.*, **39**, pp.217-239.

[181] Konak A, Coit DW, Smith AE. Multi-objective optimization using genetic algorithms: A tutorial. Reliab Eng Syst Safe. 2006;91(9): 992-1007.

[182] Krakauer, D.C., 2011. Darwinian demons, evolutionary complexity, and information maximization. *Chaos*, **21**(3), p.037110.

[183] Krakauer, D., Bertschinger, N., Olbrich, E., Ay, N. and Flack, J.C., 2014. The information theory of individuality. arXiv preprint arXiv:1412.2447.

[184] Kussell, E. and Leibler, S., 2005. Phenotypic diversity, population growth, and information in fluctuating environments. *Science*, **309**(5743), pp.2075-2078.

[185] Lagrange, J.L., Boissonnade, A., Vagliente, V.N. and Galuzzi, M., 1998. "Analytical Mechanics: Translated from the Mecanique analytique, nouvelle edition of 1811." *ISIS-International Review Devoted to the History of Science and its Cultural Influence*, 89(1), pp. 140-140.

[186] Landau, L.D., 1991. *Statistical Physics*. Vol. 5 of Landau and Lifshitz Course of Theoretical Physics.

[187] Landauer R. Irreversibility and heat generation in the computing process. IBM J Res Dev. 2961;5(3): 183-191.

[188] Laumanns M, Rudolph G, Schwefel HP, A spatial predator-prey approach to multi-objective optimization: A preliminary study. In *Parallel Problem Solving from Nature*, 241-249. Springer Berlin Heidelberg (1998).

[189] Lippiello E, Corral Á, Bottiglieri M, Godano C, de Arcangelis L, Scaling behavior of the earthquake intertime distribution: Influence of large shocks and time scales in the Omori law. *Phys. Rev. E* **86**(6), 066119 (2012).

[190] Lloyd, S.P., 1982. Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, **28**(2), pp.129-137.

[191] R. Louf, P. Jensen, and M. Barthélemy, Emergence of hierarchy in cost-driven growth of spatial networks. *P. Nat. A. Sci.* **110**(22), 8824 (2013).

[192] Lu ET, Hamilton RJ, Avalanches and the distribution of solar flares. Astrophys. J. **380**, L89-L92 (1991).

[193] Maass, W., 2014. Noise as a resource for computation and learning in networks of spiking neurons. *Proceedings of the IEEE*, **102**(5), pp.860-880.

[194] Maass, W. & Bishop C.M. 2001 *Pulsed neural networks*. Cambridge, MA: MIT Press.

[195] Maass, W., Natschläger, T. and Markram, H., 2002. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Comput.*, *14*(11), pp.2531-2560.

[196] Mahowald, K., Fedorenko, E., Piantadosi, S.T. and Gibson, E., 2013. Speakers choose shorter words in predictive contexts. *Cogn.*, **126**(2), pp.313-318.

[197] Mandelbrot, B.B. and Hudson, R.L., 2004. “The (mis) Behaviour of Markets: A Fractal View of Risk, Ruin and Reward. ”

[198] Mantegna, R.N., Buldyrev, S.V., Goldberger, A.L., Havlin, S., Peng, C.K., Simons, M. and Stanley, H.E., 1994. Linguistic features of noncoding DNA sequences. *Phys. Rev. Let.*, **73**(23), p.3169.

[199] Markov, A., 1971. Extension of the limit theorems of probability theory to a sum of variables connected in a chain.

[200] Martin, A. and Chao, L.L., 2001. Semantic memory and the brain: structure and processes. *Curr. Opin. Neurobiol.*, **11**(2), pp.194-201.

[201] Martin OC, Monasson R, Zecchina R. Statistical mechanics methods and phase transitions in optimization problems. Theor Comp Sci. 2001;265: 3-67.

[202] Mas-Colell A, Whinston MD, Green JR, *Microeconomic theory (Vol. 1)*. New York: Oxford university press (1995).

[203] Massad, D., Omodei, E., Strohecker, C., Xu, Y., Garland, J., Zhang, M. and Seoane, L.F., 2013. Unfolding History: Classification and analysis of written history as a complex system. In *Proc. of the CSSS*, Santa Fe Institute.

[204] Mathias N, Gopal V. Small Worlds: how and why. Phys Rev E. 2001;63(2): 021117.

[205] Maxwell JC. Theory of Heat. Longmans, Green, and Co.; 1904.

[206] Maynard-Smith, J. 2000 The concept of information in biology. *Philos. Sci.*, **67**(2), 177-194.

[207] Maynard-Smith, J. and Szathmáry, E. 1997. *The major transitions in evolution.* Oxford University Press.

[208] http://www.mcdmsociety.org/

[209] McGhee, G.R., 1999. “Theoretical morphology: the concept and its applications. ” Columbia University Press.

[210] McNamara, J.M. and Houston, A.I., 1987. Memory and the efficient use of information. *J. Theor. Biol.*, **125**(4), pp.385-395.

[211] H. Mengistu, J. Huizinga, J. B. Mouret, and J. Clune, The evolutionary origins of hierarchy. arXiv preprint arXiv:1505.06353 (2015).

[212] Milgram, S., 1967. The small world problem. *Psychol. Today*, **2**(1), pp.60-67.

[213] Miller, G.A., 1995. WordNet: a lexical database for English. *Commun. ACM*, **38**(11), pp.39-41.

[214] Mitzenmacher M, A brief history of generative models for power law and lognormal distributions. *Internet Math.* **1**(2), 226-251 (2004).

[215] Mora T, Bialek W, Are biological systems poised at criticality? *J. Stat. Phys.* **144**(2), 268-302 (2011).

[216] Mora T, Walczak AM, Bialek W, Callan CG, Maximum entropy models for antibody diversity. *Proc. Natl. Acad. Sci.* **107**, 5405-10 (2010).

[217] Motter, A. E., de Moura, A. P. S., Lai, Y.-C., and Dasgupta, P., 2002. Topology of the conceptual network of language. *Physical Review E*, **65**(6), p.065102.

[218] Murray CD. The physiological principle of minimum work. I. The vascular system and the cost of blood volume. Physiology. 1926;12: 207-213.

[219] Nash, J.F., 1950. Equilibrium points in n-person games. *Proc. Nat. Acad. Sci.*, **36**(1), pp.48-49.

[220] Nash, J., 1951. Non-cooperative games. *Ann. Math.*, **54**(2) pp.286-295.

[221] Nehert RA, Shraiman BI. Statistical Genetics and Evolution of Quantitative Traits. Rev Mod Phys. 2011;83(4): 1283-1300.

[222] Newman ME, Power laws, Pareto distributions and Zipf's law. *Contemp. Phys. 46*(5), 323-351 (2005).

[223] Newman MEJ. Networks. Oxford University Press; 2010.

[224] Nicolis, G. & Prigogine, I. 1977 Self-organization in nonequilibrium systems. New York, NY: Wiley, New York.

[225] https://en.wikipedia.org/wiki/Nim_Chimpsky

[226] http://www.nltk.org/

[227] Noor E, Milo R. Efficiency in Evolutionary tradeoffs. Science. 2012;336: 1114.

[228] Nowak, M.A., Plotkin, J.B. and Jansen, V.A., 2000. The evolution of syntactic communication. *Nature*, **404**(6777), pp.495-498.

[229] Nowak MA, Plotkin JB, Krakauer DC, The evolutionary language game. *J. Theor. Biol.*, **200**(2), 147-162 (1999).

[230] Nurse, P. 2008 Life, logic and information. *Nature* **454**(7203), 424-426.

[231] Obst, O., Polani, D. and Prokopenko, M., 2009. Origins of scaling in genetic code. In *Advances in Artificial Life. Darwin Meets von Neumann* (pp. 85-93). Springer Berlin Heidelberg.

[232] Oehlenschläger, F. & Eigen, M. 1997 30 Years Later – a New Approach to Sol Spiegelman's and Leslie Orgel's in vitro EVOLUTIONARY STUDIES Dedicated to Leslie Orgel on the occasion of his 70th birthday. Origins Life Evol. B., **27**(5-6), 437-457.

[233] Oster, G.F. and Wilson, E.O., 1978. *Caste and ecology in the social insects*. Princeton University Press.

[234] Otero-Espinar, M.V., Seoane, L.F., Nieto, J.J. and Mira, J., 2013. An analytic solution of a model of language competition with bilingualism and interlinguistic similarity. *Physica D*, **264**, pp.17-26.

[235] Otero-Muras I, Banga JR. Multicriteria global optimization for biocircuit design. BMC Syst Biol. 2014;8: 113.

[236] H. M. Ozaktas, Paradigms of connectivity for computer circuits and networks. *Opt. Eng.* **31**, 1536 (1992).

[237] Pareto, V., 1906. “Manuale di economia politica (Vol. 13).” Societa Editrice.

[238] https://en.wikipedia.org/wiki/Vilfredo_Pareto

[239] Pearl, J., 1985. Bayesian networks: A model of self-activated memory for evidential reasoning. University of California (Los Angeles). Computer Science Department.

[240] Peixoto, T.P. and Bornholdt, S., 2012. Evolution of robust network topologies: Emergence of central backbones. *Phys. Rev. Let.*, **109**(11), p.118703.

[241] Pérez-Escudero A, de Polavieja GG. Optimally wired subnetwork determines neuroanatomy of *Caenorhabditis elegans*. P Nat A Sci. 2007;104(43): 17180-17185.

[242] Perunov, N., Marsland, R. and England, J., 2014. Statistical physics of adaptation. arXiv:1412.1875.

[243] Petersen, A.M., Tenenbaum, J.N., Havlin, S., Stanley, H.E. and Perc, M., 2012. Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Sci. Rep.*, **2**.

[244] Pfante, O., Bertschinger, N., Olbrich, E., Ay, N. and Jost, J., 2014. Comparison between different methods of level identification. *Adv. Complex Syst.*, **17**(02), p.1450007.

[245] Piantadosi, S.T., Tily, H. and Gibson, E., 2011. Word lengths are optimized for efficient communication. *Proc. Natl. Acad. Sci.*, **108**(9), pp.3526-3529.

[246] Piantadosi, S. T., Tily, H., and Gibson, E., 2012. The communicative function of ambiguity in language. *Cogn.*, **122**(3), pp.280-291.

[247] Pinker, S., 1994. *The language instinct: The new science of language and mind* (Vol. 7529). Penguin UK.

[248] Pinker, S. and Bloom, P., 1990. Natural language and natural selection. *Behav. Brain Sci.*, **13**(4), pp.707-727.

[249] F. R. Pitts, A graph theoretic approach to historical geography. *Prof. Geogr.* **17**(5), 15 (1965).

[250] Priester C, Schmitt S, Peixoto T P. Limits and tradeoffs of Topological Network Robustness. PLoS ONE. 2014;9(9): e108215.

[251] M. Prokopenko, N. Ay, O. Obst, and D. Polani, Phase transitions in least-effort communications. *J. Stat. Mech.* **11**, P11025 (2010).

[252] PrÃ¼gel-Bennett A, Shapiro JL. Analysis of Genetic Algorithms Using Statistical Mechanics. Phys Rev Lett. 1994;72(9): 1305-1309.

[253] Pustejovsky, J. , 1991. The generative lexicon. *Comput. Linguis.*, **17**(4), pp.409-441.

[254] Pustejovsky, J., 1995. *The generative lexicon*. MIT Press, Cambridge, MA.

[255] Raup, D.M., 1966. Geometric analysis of shell coiling: general problems. *J. Paleo.*, pp.1178-1190.

[256] Raup, D.M., 1967. Geometric analysis of shell coiling: coiling in ammonoids. *J. Paleo.*, pp.43-65.

[257] Regot, S., Macia, J., Conde, N., Furukawa, K., Kjellén, J., Peeters, T., Hohmann, S., de Nadal, E., Posas, F. and Solé, R., 2011. Distributed biological computation with multicellular engineered networks. *Nature*, **469**(7329), pp.207-211.

[258] Reif, F., 2009. *Fundamentals of statistical and thermal physics*. Waveland Press.

[259] Reimann A, Ebeling W. Ensemble-based control of evolutionary optimization algorithms. Phys Rev E. 2002;65: 046106.

[260] Rinaldo, A., Rodríguez-Iturbe, I., Rigon, R., Ijjasz-Vasquez, E. and Bras, R.L., 1993. Self-organized fractal river networks. *Phys. Rev. Let.*, **70**(6), p.822.

[261] Rivoire, O. and Leibler, S., 2011. The value of information for populations in varying environments. *J. Stat. Phys.*, **142**(6), pp.1124-1166.

[262] Rodríguez-Iturbe, I. and Rinaldo, A., 2001. *Fractal river basins: chance and self-organization*. Cambridge University Press.

[263] Rosas-Casals, M., Valverde, S. and SolÃ©, R.V., 2007. Topological vulnerability of the European power grid under errors and attacks. *Int. J. Bifurcat. Chaos*, **17**(07), pp.2465-2475.

[264] Rosch, E.H., 1973. Natural categories. *Cognitive psychology*, **4**(3), pp.328-350.

[265] C. Salge, N. Ay, D. Polani, and M. Prokopenko, Zipf's Law: Balancing Signal Usage Cost and Communication Efficiency. SFI working paper: 13-10-033 (2013).

[266] Salinas, S., 2013. *Introduction to statistical physics*. Springer Science & Business Media.

[267] Scalise, S., 1986. *Generative morphology (Vol. 18)*. Walter de Gruyter.

[268] Schindel, D.E., 1990. Unoccupied morphospace and the coiled geometry of gastropods: architectural constraint or geometric covariation. *Causes of evolution: a paleontological perspective*. University of Chicago Press, Chicago, pp.270-304.

[269] Schneider, C.M., Moreira, A.A., Andrade, J.S., Havlin, S. and Herrmann, H.J., 2011. Mitigation of malicious attacks on networks. *Proc. Natl. Acad. Sci.*, **108**(10), pp.3838-3841.

[270] Schuetz R, Zamboni N, Zampieri M, Heinemann M, Sauer U. Multidimensional Optimality of Microbial Metabolism. Science. 2012;336: 601-604.

[271] Schuster, P. 1996. How does complexity arise in evolution? *Complexity*, **2**(1), 22-30.

[272] Schuster P. Optimization of multiple criteria. Complexity. 2012;18: 5-7.

[273] Searls, D. B. , 2002. The language of genes. *Nature*, **420**(6912), pp.211-217.

[274] Segré, D., Ben-Eli, D. and Lancet, D., 2000. Compositional genomes: prebiotic information transfer in mutually catalytic noncovalent assemblies. *Proc. Nat. Acad. Sci.*, **97**(8), pp.4112-4117.

[275] Segré, D., Shenhav, B., Kafri, R. and Lancet, D., 2001. The molecular roots of compositional inheritance. *J. Theor. Biol.*, **213**(3), pp.481-491.

[276] Segura, C., Coello, C.A.C., Miranda, G. and León, C., 2013. Using multi-objective evolutionary algorithms for single-objective optimization. *J. Oper. Res.*, **11**(3), pp.201-228.

[277] P. Sen, S. Dasgupta, A. Chatterjee, P. A. Sreeram, G. Mukherjee, and S. S. Manna, Small-world properties of the Indian railway network. *Phys. Rev. E* **67**(3), 036106 (2003).

[278] Seoane, L.F., Gabler, S. and Blankertz, B., 2015. Images from the mind: BCI image evolution based on rapid serial visual presentation of polygon primitives. *Brain-Computer Interfaces*, 2(1), pp.40-56.

[279] https://www.youtube.com/watch?v=LERyK3z2cYs

[280] Seoane, L. F. and Mira, J., 2016. Modeling the life and death of competing languages from a physical and mathematical perspective. In **Bilingualism and Minority Languages: Current trends and developments**.

[281] Seoane, L.F., Parafita, M. C., Casares, H., Monteagudo, H., Mira, J., 2016. Predicting the evolution of heterogeneous language contact situations: the case of Galician-Spanish bilingualism. In preparation.

[282] Seoane, L. F. and Solé, R., 2013. A multiobjective optimization approach to statistical mechanics. http://arxiv.org/abs/1310.6372

[283] Seoane, L.F. and Solé, R.V., 2013. Synthetic biocomputation design using supervised gene regulatory networks. arXiv preprint arXiv:1310.5017.

[284] Seoane, L. F. and Solé, R., 2015. Systems poised to criticality through Pareto selective forces. arXiv preprint arXiv:1510.08697.

[285] Seoane, L.F. and Solé, R., 2015. Phase transitions in Pareto optimal complex networks. *Phys. Rev. E*, **92**(3), p.032807.

[286] Seoane, L. F., Solé, R., 2015. Multiobjective optimization and phase transitions. In *Proceedings of ECCS 2014*, ch.22.

[287] Seoane, L. F. and Solé, R., 2016. Information theory, predictability, and the emergence of complex life. In preparation.

[288] Seoane, L. F. and Solé, R., 2016. Exploring the morphospace of communication codes. In preparation.

[289] Shalizi, C.R. and Moore, C., 2003. What is a macrostate? Subjective observations and objective dynamics. arXiv preprint cond-mat/0303625.

[290] Shannon, C. E. 2001 A mathematical theory of communication. *Bell Syst. Tech. J.* **27**(3), 379-423. doi:10.1002/j.1538-7305.1948.tb01338.x

[291] Shannon, C. E. & Weaver, W. 1949 *The Mathematical Theory of Communication*. Univ of Illinois Press, 1949.

[292] H. Sheftel, O. Shoval, A. Mayo, and U. Alon, The geometry of the Pareto front in biological phenotype space. *Ecol. Evol.* **3**(6), 1471 (2013).

[293] Shoval O, Sheftel H, Shinar G, Hart Y, Ramote O, Mayo A, et al. Evolutionary tradeoffs, Pareto Optimality, and the Geometry of Phenotype Space. Science. 2012;336: 1157-1160.

[294] Sigman, M. and Cecchi, G. A., 2002. Global organization of the Wordnet lexicon. *Proc. Natl. Acad. Sci.*, **99**(3), pp.1742-1747.

[295] Sohl-Dickstein, J., Battaglino, P.B. and DeWeese, M.R., 2011. New method for parameter estimation in probabilistic models: minimum probability flow. *Phys. Rev. Let.*, **107**(22), p.220601.

[296] Solé, R., 2005. Language: syntax for free?. *Nature*, **434**(7031), pp.289-289.

[297] Solé, R.V. and Delgado, J., 1996. Universal computation in fluid neural networks. *Complexity*, **2**(2), pp.49-56.

[298] Solé, R. V., Corominas-Murtra, B., and Fortuny, J., 2010. Diversity, competition, extinction: the ecophysics of language change. *J. Phys. Soc. Interface*, **7**(53), pp.1647-1664.

[299] Solé, R. and Goodwin, B., 2008. *Signs of life: How complexity pervades biology*. Basic books, 2008.

[300] Solé, R.V. and Miramontes, O., 1995. Information at the edge of chaos in fluid neural networks. *Physica D*, **80**(1), pp.171-180.

[301] Solé, R.V. and Seoane, L.F., 2015. Ambiguity in language networks. *Linguist. Rev.*, **32**(1), pp.5-35.

[302] Stadler, W., 1979. A survey of multicriteria optimization or the vector maximum problem, part I: 1776–1960. *J. Optimiz. Theory App.*, **29**(1), pp.1-52.

[303] Steels, L., 2000. The emergence of grammar in communicating autonomous robotic agents. In *ECAI* (pp. 764-769).

[304] Steels, L., 2003. Evolving grounded communication for robots. *Trends Cogn. Sci.*, **7**(7), pp.308-312.

[305] Steels, L. ed., 2011. *Design patterns in fluid construction grammar (Vol. 11)*. John Benjamins Publishing.

[306] Steels, L., ed., 2012. *Experiments in cultural language evolution. Vol. 3*. John Benjamins Publishing, 2012.

[307] Steels, L. and De Beule, J., 2006, A (very) brief introduction to fluid construction grammar. In *Proceedings of the Third Workshop on Scalable Natural Language Understanding* (pp. 73-80). Association for Computational Linguistics.

[308] Steels, L., De Beule, J., and Neubauer, N., 2005. Linking in Fluid Construction Grammars. In *BNAIC*, pp.11-20.

[309] Stephens GJ, Mora T, Tkačik G, Bialek W, Statistical thermodynamics of natural images. *Phys Rev Let* **110**(1), 018701 (2013).

[310] Stephens GJ, Bialek W, Statistical mechanics of letters in words. *Phys Rev E*, **81**(6), 066119 (2010).

[311] Steuer, R.E., 1986. *Multiple criteria optimization: theory, computation, and applications*. Wiley.

[312] Steuer RE, Na P. Multiple Criteria Decision Making Combined with Finance: A Categorized Bibliographic Study. Eur J Oper Res. 2003;150(3): 496-515.

[313] Steyvers, M. and Tenenbaum, J.B., 2005. The Large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive science*, **29**(1), pp.41-78.

[314] Still, S., 2014. Information bottleneck approach to predictive inference. *Entropy*, **16**(2), pp.968-989.

[315] Still, S., Bialek, W. and Bottou, L., 2003. Geometric clustering using the information bottleneck method. In Advances in neural information processing systems (p. None).

[316] Still, S. and Crutchfield, J.P., 2007. Structure or Noise? Santa Fe Institute working paper: #2007-08-020.

[317] Still, S., Crutchfield, J.P. and Ellison, C.J., 2007. Optimal causal inference. Santa Fe Institute working paper: #2007-08-024.

[318] Szathmáry, E., 1989. The integration of the earliest genetic information. *Trends Ecol. Evol.*, **4**(7), pp.200-204.

[319] Szathmáry, E. and Maynard-Smith, J. 1997. From replicators to reproducers: the first major transitions leading to life. *J. Theor. Biol.* **187**, 555-571.

[320] Szekely P, Sheftel H, Mayo A, Alon U. Evolutionary Tradeoffs between Economy and Effectiveness in Biological Homeostasis Systems. PLoS Comput Biol. 2013;9(8): e1003163.

[321] Szekely, P., Korem, Y., Moran, U., Mayo, A. and Alon, U., 2015. The Mass-Longevity Triangle: Pareto Optimality and the Geometry of Life-History Trait Space. *PLoS Comput. Biol.*, **11**(10), p.e1004524.

[322] Tegmark, M., 2015. Consciousness as a state of matter. *Chaos Soliton. Fract.*, **76**, pp.238-270.

[323] Tendler, A., Mayo, A. and Alon, U., 2015. Evolutionary tradeoffs, Pareto optimality and the morphology of ammonite shells. *BMC Syst. Biol.*, **9**(1), p.1.

[324] Tishby, N., Pereira, F.C. and Bialek, W., 2000. The information bottleneck method. arXiv preprint physics/0004057.

[325] Tkačik G, Marre O, Amodei D, Schneidman E, Bialek W, Berry MJ II, Searching for collective behavior in a large network of sensory neurons. *PLoS Comput Biol* **10**(1), e1003408 (2014).

[326] Tkačik G, Marre O, Mora T, Amodei D, Berry II MJ, Bialek W, The simplest maximum entropy model for collective behavior in a neural network. *J. Stat. Mech.* **2013**(03), P03011 (2013).

[327] Tkačik, G. and Bialek, W., 2014. Information processing in living systems. arXiv preprint arXiv:1412.8752.

[328] Tkačik G, Mora T, Marre O, Amodei D, Palmer SE, Berry MJ II, Bialek W, Thermodynamics and signatures of criticality in a network of neurons. *Proc. Nat. Acad. Sci.* **112**(37), 11508-11513 (2015).

[329] Touchette H, Ellis RS, Turkington B. An introduction to the thermodynamic and macrostate levels of nonequivalent ensembles. Phys A. 2004;340: 138-146.

[330] Trubetzkoy, N. S., 1969. *Principles of Phonology*. Reprinted 1969, University of California Press, Berkeley, CA.

[331] Turing, A. M. 1936 On computable numbers, with an application to the Entscheidungsproblem. *J. of Math* **58**(345-363), 5.

[332] Valverde, S., Cancho, R.F. and Solé, R.V., 2002. Scale-free networks from optimal design. *Europhys. Let.*, *60*(4), p.512.

[333] Valverde, S. and Solé, R.V., 2003. Hierarchical small worlds in software architecture. arXiv preprint cond-mat/0307278.

[334] Varchenko AN. Evolutions of convex hulls and phase transitions in thermodynamics. J Sov Math. 1990;52(4): 3305-3325.

[335] Voltaire, 1759. *Candide, ou l'Optimisme*. Cramer, Marc-Michel Rey, Jean Nourse, Lambert, and others.

[336] von Neumann, J., 1928. Zur theorie der gesellschaftsspiele. *Math. Ann.*, **100**(1), pp.295-320.

[337] von Neumann, J. & Burks, A. W. 1966 Theory of self-reproducing automata. *IEEE Trans. Neural Netw.*, **5**, 3-14.

[338] Wagensberg J. 2000 Complexity versus uncertainty: the question of staying alive. *Biol. Phil.*, **15**, 493-508.

[339] Walker, S. I. & Davies C. W. 2012 The algorithmic origins of life. *J. Phys. Soc. Interface*, **10**: 20120869.

[340] https://en.wikipedia.org/wiki/Léon_Walras

[341] Watanabe, S. and Sakakibara, K., 2005. Multi-objective approaches in a single-objective optimization environment. In Evolutionary Computation, 2005. The 2005 IEEE Congress on (Vol. 2, pp. 1714-1721). IEEE.

[342] D. J. Watts and S. H. Strogatz, Collective dynamics of 'small-world' networks. *Nature* **393**, 440 (1998).

[343] V. J. Wedeen, D. L. Rosene, R. Wang, G. Dai, F. Mortazavi, P. Hagmann, J. H. Kaas, and W.-Y. I. Tseng, The Geometric Structure of the Brain Fiber Pathways. *Science* **335**, 1628 (2012).

[344] West GB, Brown JH, Enquist BJ. A General Model for the Origin of Allometric Scaling Laws in Biology. Science. 1997;276: 122-126.

[345] West GB, Brown JH, Enquist BJ. A general model for the structure and allometry of plant vascular systems. Nature. 1999;400: 664-667.

[346] Wilson LA, Moore MD, Picarazzi JP, Miquel SDS. Parallel genetic algorithm for search and constrained multi-objective optimization. Parallel and Distributed Processing Symposium, Proceedings. 2004.

[347] Wolpert, D.H., Grochow, J.A., Libby, E. and DeDeo, S., 2014. Optimal high-level descriptions of dynamical systems. Santa Fe Institute working paper: #2015-06-017

[348] Wray, Alison, ed., 2002. *The transition to language. Vol. 2.* Peterson’s.

[349] Wu FY. The Potts model. Rev Mod Phys. 1982;51(1): 235-267.

[350] Yapo, P.O., Gupta, H.V. and Sorooshian, S., 1998. Multi-objective global optimization for hydrologic models. Journal of hydrology, 204(1), pp.83-97.

[351] Zipf GK, *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Reading, MA (1949).

[352] Zitzler E. Evolutionary Algorithms for Multiobjective Optimization: Methods and Applications, A dissertation submitted to the Swiss Federal Institute of Technology. PhD Thesis, Eidgenössische Technische Hochschule Zürich. 1999.

Index

$80-20$ principle, 6
ϵ machine, 165, 204
k-means clustering, 143, 145
p-value, 75
More than Nature Needs, 161
Mycoplasma pneumoniae, 154
The Mind's I, 150

Abrupt ending of the Pareto front, 29, 39, 47, 63, 66, 70
Algorithmic complexity, 153–155, 165
Alphabet, 158, 159, 163, 165
Ant colony, 150
Ants, 80, 150
Archetype, 20, 143

Bialek, William, 74
Bickerton, Derek, 161, 171, 204
Bird flock, 151
Brain, 151

Canonical ensemble, 38, 40, 41, 44, 46, 74, 83
Causal closure, 158
Causal shielding, 204
Cavity, 30, 37, 39, 42, 43, 47, 49, 66, 67, 80, 82, 83, 85, 88, 89
Chomsky, Noam, 162
Clique, 58–60, 63–67, 69, 70, 72, 76
Coarse grained model, 158, 159
Communication codes, 124, 126, 127, 129, 143, 160, 202
Communication networks, 2, 17, 18, 80, 202
Commutativity, 159
Competitive market, 77
Complex networks, 17, 53, 71, 102, 201
Complex systems, 3, 6, 32, 53, 71, 73, 74, 77, 153, 199, 200, 202
Complexity, 155
Computational mechanics, 165, 204
Concavity, 27
Constrained optimization, 7, 77

Control parameter, 27, 32, 44, 85, 87–89
Convex hull, 30, 38, 40–44, 89, 156, 157, 204
Convexity, 27, 28, 63, 69, 70
Critical ensemble, 126
Critical point, 23, 32, 35–37, 45, 49, 73, 83, 125, 127, 139, 200–203
Critical process, 156
Critical second order phase transition, 35
Critical system, 156
Criticality, 23, 32, 38, 73, 83, 125, 156, 200, 201, 203
Crypticity, 155

Darwinian evolution, 2, 177, 179, 205
Dennett, Daniel, 3, 150, 173, 178, 196, 205
Diffusion, 80
Distortion rate theory, 156
Dominance, 9, 10, 13–15, 40, 43, 49, 169, 207

Edge of chaos, 155, 200
Edgeworth box, 5, 6
Edgeworth, Francis Ysidro, 4
Edgeworth, Francis Ysidro , 5
Efficient communication, 80, 91, 124, 201, 203
Einstein’s relativity, 1
Emergence, 150
Energy, 1, 24, 25, 27, 30, 34, 37, 38, 40–42, 44, 46, 48, 49, 54–56, 63, 66, 68, 72, 74, 75, 80–83, 85–89, 113, 126, 167, 169, 170, 176
Energy landscape, 24, 30, 31, 56, 63, 66, 80–83, 85, 86, 89
Ensemble inequivalence, 44, 45
Entropy, 1, 34, 37, 38, 40–42, 44, 46, 49, 74, 75, 80, 111, 113, 119, 133, 135–138, 140, 153–155, 169, 170, 177, 181, 189, 190
Epsilon machine, 153–155, 158, 182
Excess entropy, 153, 155

First order phase transition, 30, 34–37, 41, 43–45, 47, 49, 51, 52, 60, 66, 72, 75, 80–85, 89, 114, 116, 124–126, 201–203
Fish school, 151
Fitness, 1, 19, 24, 56, 108, 113, 195, 205
Fluid Construction Grammar, 171
Fluid neural networks, 80
Free energy, 38, 40, 44, 46, 48, 49, 74, 86, 88
Fully geometric problem, 59, 69

Fully topological problem, 58, 60
Fundamental theorems of economic welfare, 77

Generative grammar, 171
Gibbs, 38
Gibbs free energy, 44
Gibbs function, 83
Gibbs potential, 44
Gibbs surface, 23, 38, 40, 44, 80, 200
Gibbs, Josiah Willard, 44
Golden mean process, 154

Hamiltonian, 151
Hamiltonian dynamics, 1
Hierarchy, 14, 17, 160–162, 165, 170, 171, 197, 204, 205
Highly Optimized Tolerance, 73, 77
Hofstadter, Douglas, 150
Hybrid phase transition, 34–37

Indifference curve, 5
Information bottleneck, 156, 157, 204
Information theory, 155, 177, 179
Informational closure, 159, 161, 204
Internal energy, 41, 42
Ising model, 23, 45, 46, 166

Jaynes, Edwin T, 38, 74, 80, 177

Knee, 85
Kolmogorov complexity, 153
Kolmogorov-Smirnov (KS) test, 141–143
Kullback-Leibler divergence, 77, 167

Lagrange multiplier, 24, 75, 156, 202
Lagrangian dynamics, 1
Language networks, 92, 98, 103, 107, 121, 133, 135
Language organ, 161
Least effort language, 108, 119, 124, 160, 202
Linear chain, 60

Macrostate, 158, 159
Markovianity, 159
Maximization, 1, 9, 27, 38, 40, 42, 46, 170
Maximum Entropy models (MaxEnt), 34, 74, 126, 165
Meaningful information, 178, 205
Merge (syntactic operator), 162
Message passing, 80
Microcanonical ensemble, 40, 41, 44, 83
Minimalist program, 162
Minimax theorem, 7
Minimization, 1, 9, 24, 25, 27, 29, 38, 40, 54, 55,

74–76, 87, 89, 113, 117, 118, 121, 124, 170
Minimum Probability Flow Learning, 166
Minimum Spanning Tree, 59, 65, 66, 69
MOO-SOO collapse, 24, 55
Morphospace, 15, 16, 20, 56, 71, 80, 124, 127, 128, 130, 143, 202
Multi Objective Genetic Algorithm (MOGA), 7, 13, 59, 73, 75, 207
Multi Objective Optimization (MOO), 1, 3, 4, 8, 13, 15, 23, 24, 38, 40, 55, 80, 207

Nash equilibrium, 7
Nash, John Forbes , 7
Nest (syntactic operator), 162
Network optimization, 54, 55, 57, 59
Network robustness, 83
Networks, 1, 2, 8, 17, 53, 57, 75, 77, 80, 82, 83, 92, 98, 101–105, 107, 133, 135, 136, 139, 165, 200, 201, 207
Neumann, John von, 7
Neural Darwinism, 80
Neuron, 151

Observational commutativity, 159
Ontogeny, 80, 123, 202
Optimal networks, 2
Optimization, 1–3, 8, 9, 13, 18, 40
Order parameter, 27–29, 31, 32, 34–37, 48, 62, 64, 65, 67, 69–71, 74, 80–82, 114, 126, 210

Panglossianism, 2
Pareto core graph, 60
Pareto front, 3, 9, 10, 14, 15, 23, 40, 44, 56, 63, 66, 69–71, 73, 80, 83, 85, 125–127, 129, 131, 142, 143, 145, 156, 157, 165, 200, 210
Pareto optimal communication codes, 125–127, 129, 160, 203
Pareto optimal complex networks, 71, 165
Pareto optimal complex systems, 53
Pareto optimal ensemble, 75
Pareto optimal grammars, 165
Pareto optimal graphs, 60
Pareto optimal language, 124
Pareto optimal networks, 17, 18, 57, 59, 60, 80, 83, 201, 207
Pareto optimal protein folding, 85
Pareto optimal systems, 202

Pareto Optimality (PO), 1, 3, 7–9, 14, 23, 40, 55, 71, 80, 165, 169, 199, 200, 202
Pareto optimality condition, 131, 132
Pareto principle, 6
Pareto selective forces, 53, 73–75, 77, 132, 201
Pareto, Vilfredo, 4, 6
Partly geometrical problem, 58, 63
Phase transition, 23, 28–30, 38, 46, 49, 53, 71, 156, 157, 200, 202
Phenotype, 19, 20, 177–179, 185, 189, 195
Phenotype space, 20
pLux, 154
Polysemy, 105, 109, 121, 130, 131, 133, 145
Potts model, 23, 45, 49
Power law, 6, 32, 94, 96, 101, 126–128, 139, 141–143, 145
Predictively reversible process, 156, 157
Pressure, 44
Prey-predator model, 77
Principal components analysis (PCA), 143, 145
Protein folding, 11, 85
Protolanguage, 162, 171
protolanguage, 204

Routing, 80

Second order phase transition, 29, 38, 44–47, 57, 63, 64, 66, 70, 72, 76, 201, 210
Self-Organized Criticality, 53, 73
Semantic networks, 104, 105, 136
Sharp edge, 29, 46, 63
Shell coiling, 16
Simple non-deterministic source, 157
Single Objective Optimization (SOO), 24, 156
Spiking neurons, 12
Star graph, 59, 60, 65, 66
Statistical mechanics, 23, 38, 44, 158
Susceptibility, 32, 34, 48, 63
Swarm, 151
Symbolic dynamics, 152, 162
Synonymy, 109, 124, 130, 131, 146, 160
Syntax for free, 162

Target function, 7, 9, 20, 165, 169
Target space, 9, 15, 20, 55, 80, 128, 157, 169
Temperature, 38, 40, 44, 46, 49
Thermodynamics, 1, 23, 38, 44, 80, 85

Utilitarianism, 4

Vascular system, 2
Vocabulary size, 130, 131
Volume, 44

Walras, Léon, 4
Walras, Léon , 5
Welfare economics, 4
WordNet, 105, 107, 146, 147, 203

Zipf’s law, 94–96, 98, 101, 104, 107, 108, 116–118, 123, 126, 139, 142, 143