SCUOLA
NORMALE
SUPERIORE

Classe di Scienze
Corso di perfezionamento in
Data Science
XXXIV ciclo

Candidato
dr. Andrea SOMAZZI

Challenges in Data Science for Complex Systems

by

Andrea Somazzi

Supervisors:

prof. Diego Garlaschelli
prof. Paolo Ferragina

Abstract

In the world of complex systems, which are omnipresent in various domains including economics, biology, and human-engineered systems, understanding their behavior poses significant challenges. The crux of comprehending these systems lies in the effective analysis of the data they produce, whose methodologies are provided by data science. However, a notable challenge in this realm is the confrontation with partial information which, if not addressed judiciously, can lead to biased interpretations or misconceptions.

This thesis is structured into five main chapters: the first provides a broad introduction to the main topics of this work. The second chapter studies opinion dynamics across various social media platforms by defining an opinion dynamics model on a multiplex network, highlighting the interplay of multiple platforms in shaping opinions. It underscores the importance of considering the different network layers, corresponding to social media platforms, when analyzing how users interact and shape their opinions. I find that empirical studies focusing on a single platform, neglecting interactions on other layers, can result in misleading conclusions. Moreover, by considering the richer picture given by this multi-platform opinion dynamics model, segregation of extreme from moderate users emerges. The subsequent chapter concerns the Generalized Maximum Entropy Principle (GMEP), a general principled technique for treating partial information. I will introduce the uninformativeness axiom, which when applied to the Uffink-Jizba-Korbel or the Hanel-Thurner families of entropies selects only Rényi entropy as viable, bridging the consistency between the GMEP and the Maximum Likelihood (ML) principles. I will also showcase the potential of ML in estimating the entropic parameter characterizing Rényi entropy, providing numerical examples supporting my theoretical findings. The fourth chapter regards nonlinear data compression, where I will intro-

duce a generalized Arithmetic Coding scheme to encode sequences in order to minimize the exponential average codeword length. Moreover, I will provide a simple yet general justification for the employment of the exponential average, instead of the linear one. Namely, if the main interest is to reduce the probability of exceeding a given codewords' length threshold, I find that the exponential average is the target quantity to minimize. All my theoretical findings will be supported and confirmed by applications on both simulated i.i.d. and real correlated data. In the last chapter, I will briefly summarize my results.

In essence, this thesis addresses the challenges posed by complex systems to data science, offering insights and methodologies to treat complex-systems-generated data, which are often fragmentary.

Contents

Chapter 1

Introduction

Many systems arising from human interactions and behavior, such as those emerging in the realms of economics and social sciences, as well as ecological, biological or communication systems, pose big challenges in terms of their description and comprehension. The reason is that they are composed by a huge number of units interacting in a nonlinear and often non-stationary fashion, exhibiting particular properties such as self-organization, emergence of power-law distributions, spontaneous order, nonlinear feedbacks and phase transitions [1, 2, 3, 4]. Such intricacy has led to the coining of the term *Complex Systems* to aptly describe the nature of these systems. The basis for our understanding of complex systems is the analysis of the data generated by their behavior. Designing effective experiments, collecting data, visualizing them, extracting meaningful information, recognizing noise and designing models are all tasks for which *data science* provides tools [5]. It encompasses a broad spectrum of methodologies, from statistical learning to data compression, equipping us with instruments necessary to tease out meaningful insights from the data.

However, in many instances, the data we gather from complex systems are only partial, representing only a fraction of the intricate web of interactions and components. Such *partial information* poses the risk of introducing biases, misconceptions, or oversimplifications that may distort our understanding of the system under consideration [6, 7]. It is, therefore, imperative to approach the analysis with caution, considering that data might not be exhaustive and that unobserved correlations can exist. Ignoring or downplaying the potential influence of these hidden dependencies could lead to skewed interpretations and misguided decisions.

A powerful tool to approach the challenges posed by having access to

partial observation is the *Maximum Entropy Principle*. Originating from the realm of physics [8], this principle has found its rigorous footing in information theory and a fertile application terrain in the domain of data science, particularly within the confines of statistical inference [9, 10]. Its strength lies in its unbiased approach. The maximum entropy principle allows to assign a probability distribution to the microscopic states of a system, which is compatible with the available macroscopic data and maximally non committal with respect to missing information [8]. Its applicability ranges from constructing appropriate null models to test hypotheses, to probabilistically reconstruct the properties of systems based on limited or fragmented information [11]. Central to this technique is the maximization of an entropy functional. In its classical formulation, such functional is represented by the *Shannon entropy*, which quantifies the inherent randomness of system's realizations [12]. Its theoretical foundations lie on the four Shannon-Khinchin (SK) [13] axioms or the five Shore-Johnson (SJ) [9] axioms, so, while it is undeniably useful and appropriate in many situations, its applicability is bounded by the conditions imposed by these axioms. In practice, many complex systems do not respect all the SK or SJ axioms because of the aforementioned dependencies. This incompatibility has motivated the exploration of entropies beyond the classical Shannonian framework [10, 14]. In fact, by relaxing some of the traditional axioms, scholars have obtained generalized definitions of entropy, finding resonances across physics, information theory, and statistical inference. Among the generalized entropies, two names stand out, both for their distinct approaches and their broad applications: Rényi entropy [15] and Tsallis entropy [16]. These two entropies provide unique perspectives and tools, and are appropriate in situations where the classical Shannon entropy might fail.

Another pivotal challenge in data science revolves around the effective storage and transmission of data generated by complex systems. Examples include time series, quantum systems, and human-engineered system, as well as specialized applications like DNA coding or transmissions with limited buffer capacity [17, 18, 19, 20]. In the context of data compression, Shannon entropy has been a pillar, representing the minimum attainable average codeword length (i.e. length of a symbol or sequence after compression). Nonetheless, in many of these applications, it is more desirable to minimize a generalized average codeword length. Specifically, the emphasis shifts

towards the *exponential* average (whose precise meaning will be clarified later), where the Shannon entropy is replaced by the Rényi entropy, the latter being the lowest achievable exponential average codeword length.

In this work I will cover the challenges posed by complex systems to Data Science by analyzing three distinct aspects. First, I will give an example on how partial information can lead to deceptive conclusions in the realm of opinion dynamics. Then, I will outline the Generalized Maximum Entropy Principle, a more general and principled method to deal with such partial information. Furthermore, I will introduce, validate and motivate a nonlinear compression algorithm which achieves the Rényi entropy with arbitrary precision. The rest of the thesis is structured as follows.

In Chapter 2, based on [21], I will provide exhaustive evidence of how partial information can lead to misleading conclusions. The focal point of this chapter is the study of how the interplay of various social media platforms shapes opinion dynamics. While a multitude of research predominantly focuses on users' interactions within a single platform (for instance, Facebook), such approach may overlook the fact that users distribute their attention (or time) across multiple platforms. The evolution of their opinions and believes is then shaped by interactions on various platforms, not just one. To encapsulate this complex interplay, I will introduce a model of opinion dynamics on a multiplex network, where each layer corresponds to a social platform. Herein, I will explore how opinions evolve due to users interacting on different social media. I will challenge a study from Facebook which posits that their recommendation algorithm, promoting like-minded content, has negligible influence on political attitudes. This study suggests that political polarization may not be caused by an homophilic recommendation system. However, I will show that the existence of numerous platforms with distinct algorithms can perpetuate polarization. For instance, while Facebook might be diversifying its content, strong homophilic recommendation systems on another platform (like Twitter) can sustain polarized views. Hence, even if one platform shifts its approach, the broader ecosystem of interconnected platforms can maintain certain biases and trends. Moreover, the multiplex opinion dynamics model unveils a segregation of users among different platforms, where the extremes separate from the moderates. This effect, which is in qualitative agreement with the literature, shows how the picture enriches by considering the different layers over which interactions

can take place.

In Chapter 3, based on [22], I will focus on the practical approach for handling partial information, delving into the problem of inferring statistical properties of systems when only partial information about them is available. I will present a brief review of the classical maximum entropy principle. This principle inherently assumes that if two separate pieces of information pertain to distinct parts of a system, these parts are then statistically independent. However, such assumption often falls short in the realm of complex systems. As a result, generalized entropy families are introduced to account for these intricacies. Within this context, I will introduce an *uniformativeness axiom*: namely, no entropic parameter characterizing generalized entropies can be inferred from a uniform probability distribution. This axiom, when applied to an entropic family, distinctly identifies a member from within that family. When applied to the widely recognized Uffink-Jizba-Korbel or Hanel-Thurner entropy families, it selects the Rényi entropy. This axiom serves a dual purpose — not only does it pinpoint the Rényi entropy, but it also retrieves the consistency between the maximum entropy and maximum likelihood principles. Moreover, I will illustrate how the maximum likelihood principle can be adeptly employed to estimate the entropic parameter that characterizes Rényi entropy.

In Chapter 4, based on [23], I will delve into the intricate matter of nonlinear data compression. While the conventional objective in encoding is to minimize the average length of encoded symbols (i.e. codewords), aiming at achieving a value close to the Shannon entropy, there are situations where the goal can shift towards minimizing the Kolmogorov-Nagumo exponential average codeword length. The optimum in such scenarios is given by the Rényi entropy. I will introduce an operational scheme, built upon a generalized version of the well-established Arithmetic Coding algorithm, to encode sequences of symbols in order to minimize the exponential average codeword length. I will provide rigorous analytical proofs, demonstrating that the proposed algorithm achieves the Rényi entropy with vanishing error. The chapter will be enriched with comprehensive examples, shedding light on the method's efficacy, regarding both i.i.d. and correlated symbols, the latter coming from real English language. Moreover, I will provide a further explanation for minimizing the exponential average. In particular, if the encoder's priority is to minimize the probability that the codewords'

length exceed a certain threshold, the exponential average emerges naturally due to its connection with the cumulant generating function of the source distribution, in turn related to the probability of large deviations.

Finally, in Chapter 5, I will briefly review the main results of my work.

Chapter 2

Social media battle for attention: opinion dynamics on competing networks

This chapter is based on: A. Somazzi, G. M. Ferro, D. Garlaschelli, and S. A. Levin. *Social media battle for attention: opinion dynamics on competing networks.* Available at `https://arxiv.org/abs/2310.18309`.

In the age of information abundance, attention is a coveted resource. Social media platforms vigorously compete for users' engagement, influencing the evolution of their opinions on a variety of topics. With recommendation algorithms often accused of creating "filter bubbles", where like-minded individuals interact predominantly with one another, it's crucial to understand the consequences of this unregulated attention market. To address this, we present a model of opinion dynamics on a multiplex network. Each layer of the network represents a distinct social media platform, each with its unique characteristics. Users, as nodes in this network, share their opinions across platforms and decide how much time to allocate in each platform depending on its perceived quality. Our model reveals two key findings. i) When examining two platforms — one with a neutral recommendation algorithm and another with a homophily-based algorithm — we uncover that even if users spend the majority of their time on the neutral platform, opinion polarization can persist. ii) By allowing users to dynamically allocate their social energy across platforms in accordance to their homophilic preferences, a further segregation of individuals emerges. While network fragmentation is usually associated with "echo chambers", the emergent multi-platform segregation leads to an increase in users' satisfaction without the undesired increase in polarization. These results underscore the significance of acknowledging how individuals gather information from a multitude of sources. Furthermore, they emphasize that policy interventions on a single social media platform may yield limited impact.

2.1 Introduction

In our contemporary landscape, online social networks have evolved into pivotal platforms for the acquisition of political news [24, 25]. These modern communication platforms have several advantages for democracy: they simplify access to information, boost citizen participation, enable individuals to express their views, counteract misinformation, and enhance transparency and responsibility in political actions. Ideally, individuals can tap into social media to encounter a range of ideological perspectives and consequently make more informed choices [26, 27, 28]. An extensive literature, empirical [29, 30, 31, 32, 33, 7] and theoretical [34, 35, 36, 37], pertains to how social media influence opinion dynamics, fostering (or not) the appearance of "echo-chambers" where individuals are mainly connected with like-minded peers. Echo chambers are often believed to stem from the recommendation systems utilized by social media platforms, which tend to link individuals with similar views. This phenomenon contributes to the rise of opinion polarization [31]. Furthermore, these automated algorithms interact with the cognitive limitations of individuals, who tend to gravitate towards information that aligns with their existing beliefs and actively avoid contradictory information [38]. It is thus crucial to distinguish between the influence of algorithms and inherent human tendencies when examining the genesis of echo chambers. A recent empirical study [7] has shown how increasing the amount of cross-cutting content on Facebook does not significantly alter political opinions, thereby suggesting that social media recommendation algorithms may not contribute to opinion polarization. They note however that political information on Facebook is mainly incidental (6.7% of the total news consumption), and thus people might get political information from other sources. More in general, the above cited studies focus on the effect of only one social media platform, while contemporary news consumption is characterized by its reliance on a multitude of sources [39]. People exhibit different news repertoires depending on a variety of needs [40, 41, 42]. Several scholars [43, 44, 45] stress that an adequate study of political opinions must consider the interplay between news repertoires and political communication processes.

Here, we answer to the following question: what are the implications of social media platforms competing for users' attention on opinion dynamics?

To address this, we develop a model of opinion dynamics which integrates three main ingredients. i) People can connect on different platforms, each platform represented by a layer in a *multiplex* network. These networks differ in their recommendation algorithms (a single parameter representing its homophily) and in their political focus. ii) Users' opinions evolve according to a well-known non-linear model [37], on the basis of interactions taking place on the multiplex. Opinions depend on the *social interaction strength*, issue controversy and the heterogeneous activity profile on social media. iii) Users allocate their time among the different social media platforms, depending on their personal preferences and limited information processing.

Our model provides two pivotal insights. Firstly, when considering two platforms — one governed by a neutral recommendation algorithm and the other by a homophily-centric algorithm — we find that even with a user majority on the neutral platform, opinion polarization can endure. Secondly, as users dynamically allocate their social engagement across platforms based on their (strong or weak) homophilic inclinations, agents manifest a pronounced separation across platforms. While most associate network fragmentation with "echo chambers", this emergent multi-platform segregation boosts user satisfaction without increasing polarization. These findings highlight the importance of recognizing the multifaceted avenues through which individuals assimilate information.

2.2 Results

As detailed in Sec. 2.4, we consider a system composed by Γ social platforms populated by N agents (see Fig. 2.1) having continuous opinions $x_i(t) \in (-\infty, +\infty), i = \{1, ...N\}$. The opinions evolve according to $\dot{x_i} = -x_i + \sum_{\gamma=1}^{\Gamma} \left(K^{(\gamma)} \sum_{j=1}^{N} A_{ij}^{(\gamma)}(t) \tanh\left(cx_j\right)\right)$, a generalization of a well-known model [46]. In absence of social interactions, i.e. the second addend on the r.h.s. equals zero, all opinions relax to $x^* = 0$. This assumption allows us to study the network influence on opinions, isolating it from the other possible causes of polarization such as identity politics [47], cognitive biases [48, 35] and economic inequality [49]. The $\tanh(cx)$ term reflects the fact that the opinion change for each interaction is limited. The parameter c represents how *controversial* a topic is. For c high enough, all opinions equally contribute to the dynamics since $|\tanh(cx)|$ saturates to 1, meaning that

people are maximally susceptible to be socially influenced. On the other hand, if c is very small, only users with extreme opinions are effectively able to influence others. The platform-dependent parameter $K^{(\gamma)}$ represents the social interaction strength. The larger $K^{(\gamma)}$, the larger opinions change as a consequence of given interactions. Its platform-dependence captures the idea that, for a given topic, users may consider a platform more "appropriate" than another. For instance, since the exposure to political content on Facebook is often incidental [7], while users tend to consume more political news on Twitter [50, 51], it is reasonable to assume that, in the realm of politics, Facebook social interaction strength $K^{(FB)}$ is lower than Twitter $K^{(TW)}$. This results in users giving less credit to the political news they are exposed on Facebook.

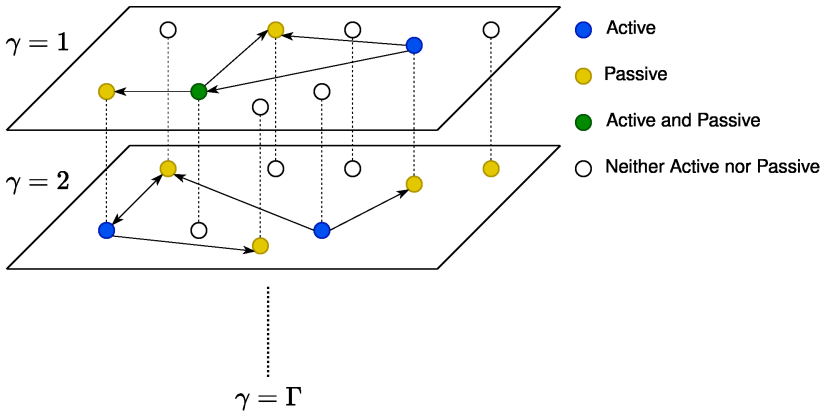


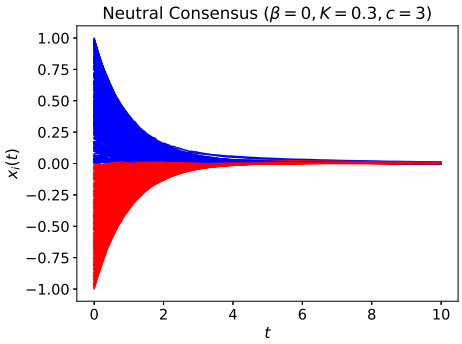
Figure 2.1: Opinion dynamics on a multiplex network. Dashed lines indicate that the same nodes (users) are shared among the layers. In this example, every active user (blue dots) contacts $m = 2$ passive users (yellow dots). A passive user can reciprocate a link with probability r (see Sec. 2.4). A user can be both active and passive on the same platform (green dots), or not engaging at all in social media activity (white dots).

The opinion dynamics model evolves on a multiplex directed network, where each layer represents a single social platform, as pictured in Fig 2.1. At every time step (see Sec. 2.4 and Appendix 2.6.1 for further details), user i can be *active* (news producer), *passive* (news consumer) or both with probability a_i, p_i and $a_i p_i$ respectively. Conditional on being active (resp. passive), user i chooses a platform γ with probability $\rho_i^{(\gamma)}$. Note that he might be active on platform γ and passive on platform γ', each with probability $\rho_i^{(\gamma)}$ and $\rho_i^{(\gamma')}$, respectively. Active users on a platform/layer contact passive users

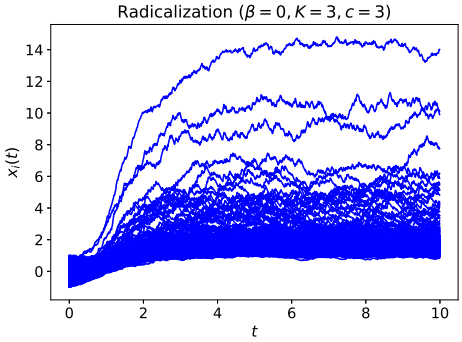
on the same platform/layer. The probability that a γ-active user i contacts a γ-passive user j on platform γ at time t is $q_{ij}^{(\gamma)}(t) \propto |x_i(t) - x_j(t)|^{-\beta^{(\gamma)}}$, leading to $A_{ji}^{(\gamma)}(t) = 1$. The exponent $\beta^{(\gamma)}$ represents the degree of homophily of the recommendation engine of platform γ.

The model with a single platform ($\Gamma = 1$) has been studied in [46]. Figure 2.2 shows the main qualitatively different dynamics that the model exhibits, as a function of the different parameters, obtained initializing the opinions uniformly in $[-1, +1]$. Specifically, when $K^{(1)} = K$ is small, social interaction is negligible and opinions relax to 0 (Fig. 2.2(a)). If instead social coupling is relevant ($K = 3$), but no homophilic recommendation engine is present ($\beta = 0$), a one-side radicalization appears where all the opinions have the same sign (Fig. 2.2(b)). When both K and β are big enough, opinions split into two opposite sides. The intuition is that with $\beta \neq 0$ agents tend to connect only with like-minded peers, an interaction which further polarizes users' stance. This effect makes it even more likely to connect with like-minded individuals; this vicious cycle fosters polarization (a more exhaustive phase diagram for the single-platform model is reported in [46]). Note that this phase manifests only if opinions are initialized with different signs. Otherwise, there is no range of parameters which leads to polarization.

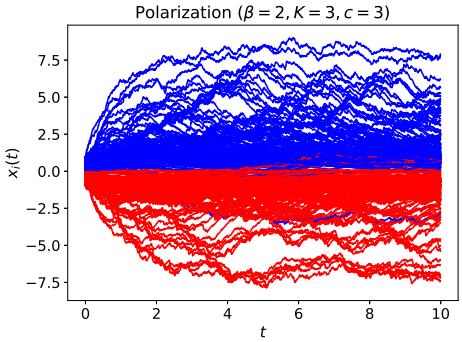
As anticipated in Sec. 2.1, we ask whether polarization can persist when users spend a tiny fraction of their time on politically-oriented social media with an homophilic recommendation engine, while engaging the rest of their time on a politically neutral platform. To explore this scenario, we consider $\Gamma = 2$ platforms and stationary and homogeneous allocation probabilities, i.e. $\rho_i^{(\gamma)}(t) = \rho^{(\gamma)}$ for all i and for $\gamma = \{1, 2\}$. Clearly, $\rho^{(1)} + \rho^{(2)} = 1$. Both the assumptions of stationarity and homogeneity will be later relaxed. Platform 1 has a set of parameters such that, if users were only there, opinions would converge to neutral consensus ($\beta^{(1)} = 0$ and $K^{(1)}$ small). On the other hand, platform 2 is assumed to adopt an homophilic recommendation engine, which translates in $\beta^{(2)} \neq 0$. The social interaction strength $K^{(2)}$ is left as a varying parameter, meaning that platform 2 could have exhibited both neutral consensus or polarization if it were the only platform, depending on its value. We are then interested in exploring the opinion dynamics for different values of $\rho^{(1)}$ and $K^{(2)}$. The former represents how long users engage on platform 1 (the "politically neutral" platform); the lat-



(a)



(b)



(c)

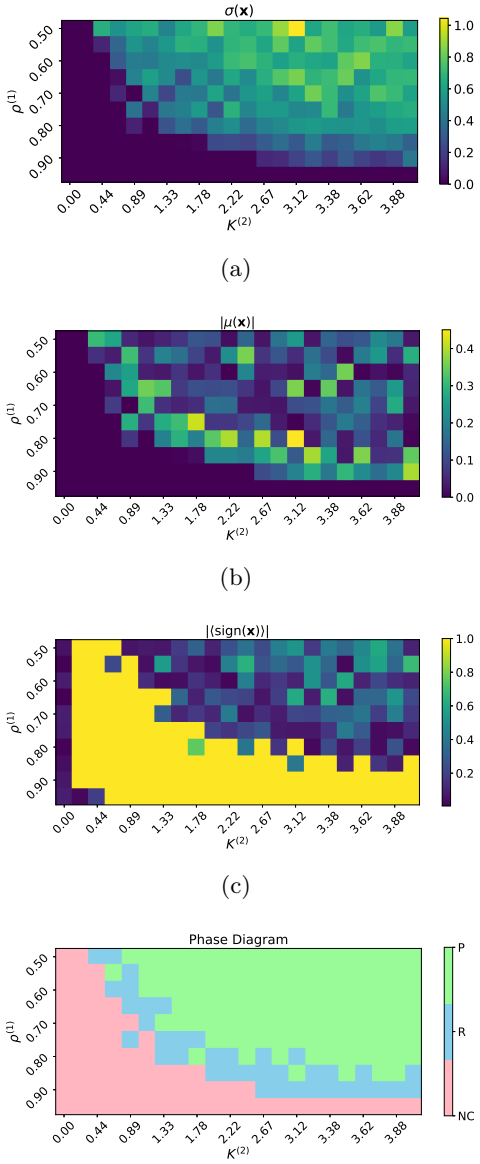
Figure 2.2: (a) Neutral consensus, all opinions converge to zero ($K = 0.3, \beta = 0$). (b) (One-sided) radicalization ($K = 3, \beta = 0$). (c) Opinion polarization, in which opinions split into two opposite sides ($K = 3, \beta = 2$). Topic controversiality and reciprocity were set to $c = 3$ and $r = 0.5$.

ter captures how "polarizing" platform 2 is. We define the *rescaled* vector of opinions at equilibrium as $\mathbf{x} = \{\frac{x_1^{(eq)}}{K^{(1)}\rho^{(1)}+K^{(2)}\rho^{(2)}}, \ldots, \frac{x_N^{(eq)}}{K^{(1)}\rho^{(1)}+K^{(2)}\rho^{(2)}}\}$ in

order to compute a set of three metrics which allow us to distinguish different opinion phases. In particular, such metrics are the standard deviation of the opinions $\sigma(\mathbf{x})$, the absolute value of the average opinion $|\mu(\mathbf{x})|$ and the absolute value of the average opinions' sign $|\langle \text{sign}(\mathbf{x})\rangle|$. In Figure 2.3 we show the results of our analysis for $\beta^{(1)} = K^{(1)} = 0$ and $\beta^{(2)} = 3$. We can observe three main phases. i) Neutral Consensus ($\sigma(\mathbf{x}) \approx 0, \mu(\mathbf{x}) \approx 0$), observable when platform 2 is not polarizing enough (i.e. $K^{(2)}$ not high enough) w.r.t. the time spent on the neutral platform 1[1]. ii) Radicalization ($|\mu(\mathbf{x})| >> 0, \langle \text{sign}(\mathbf{x})\rangle = 1$) is evident by the fact that all opinions share the same sign. Such phase is driven by an initial relaxation towards $x_i(t) \approx 0\ \forall i$ due to the neutral platform. Then, when close to 0, opinions start to share the same sign and are progressively amplified by the polarizing (now, rather, radicalizing) platform. iii) Polarization ($\sigma(\mathbf{x}) >> 0, \langle \text{sign}(\mathbf{x})\rangle < 1$) emerges if platform 2 can sustain diverging opinions, i.e. if $K^{(2)}$ is big enough to off-set the time users spend on the neutral platform $\rho^{(1)}$. The take home message is that polarization can persist even when users spend most of their time on a politically neutral platform ($\rho^{(1)} > 0.5$), thus suggesting the importance of considering that users gather information from different sources. In Fig. 2.5 of the Appendix 2.6.2, we show a similar phase diagram for a different value $\beta^{(2)}$. The qualitative picture remains the same.

The above results are obtained by assuming that users' taste for social media platforms may depend on factors which are not captured in the model (e.g. better user interface), therefore we had an homogeneous and stationary allocation probability $\boldsymbol{\rho} = \{\rho^{(1)}, \rho^{(2)}\}$. Hereafter, we suppose that users dynamically *choose* their *Social Media Repertoire* (SMR), depending on the perceived political quality of platforms. Based on the psychological theory of optimal distinctiveness [52], we imagine that each user looks for a trade-off between assimilation (homophily) and differentiation (debate). As detailed in Sec. 2.4.3, we capture such desired balance into the (user-dependent) parameter ϕ_i, which represents the desired fraction of "far" opinions (i.e. contributing to differentiation) user i wants to be exposed to. In particular, inspired by bounded confidence theory [53], user i considers "far" opinions those x such that $|x_i - x| > r$. The others are considered "close" opinions, contributing to assimilation. Thus, while in bounded confidence theory users

[1]The neutral consensus phase can have both $|\langle \text{sign}(\mathbf{x})\rangle| \approx 0$ and $|\langle \text{sign}(\mathbf{x})\rangle| \approx 1$, as the opinions never really reach exactly 0.



(a)

(b)

(c)

(d)

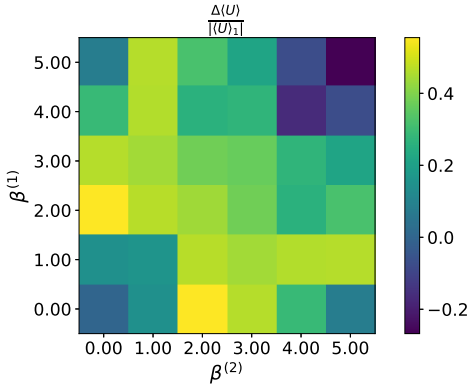
Figure 2.3: Allocation diagram. In all the four panels, on the horizontal axis the social interaction strength of the polarizing platform, and on the vertical axis the time spent on the neutral platform. (a) Standard deviation of the opinions; (b) Absolute value of average opinion; (c) Absolute value of average opinions' sign; (d) Phase diagram. Panel (d) is obtained through "filtering" panels (a)-(b)-(c) according to the conditions detailed in the main text. In short: i) NC corresponds to $\sigma(\mathbf{x}) \approx 0$ and $\mu(\mathbf{x}) \approx 0$). ii) R corresponds to $|\mu(\mathbf{x})| >> 0$ and $\langle \text{sign}(\mathbf{x}) \rangle = 1$. iii) P corresponds to $\sigma(\mathbf{x}) >> 0$ and $\langle \text{sign}(\mathbf{x}) \rangle < 1$.

do not engage with peers having far opinions, we relax this hypothesis by introducing ϕ_i, which can be seen as user i's desired probability to contact a distant peer. The idea is to capture the observed desire of debate. Indeed, as reported in [54], despite the general tendency for social media networks to form homogeneous communities, networks formed through reply-to messages reveal a users' stance heterophily, with individuals using replies more often to express divergent opinions. We define the utility (i.e. satisfaction) of each user as $U_i(t) = -(f_i(t) - \phi_i)^2$, where $f_i(t)$ is the fraction of distinctiveness experienced by user i, which is compared to the desired one. Clearly, $f_i(t)$ depends on the SMR of user i (i.e. $\boldsymbol{\rho}_i(t) = \{\rho_i^{(\gamma)}(t)\}_{\gamma=1}^{\Gamma}$), since user's exposure to distinctiveness and assimilation depends on the connections formed on each platform, which in turn depend on his allocation probability. We assume therefore that user i dynamically updates $\boldsymbol{\rho}_i(t)$ in order to maximize his $U_i(t)$ (see Sec. 2.4.3 and Appendix 2.6.1 for additional details).

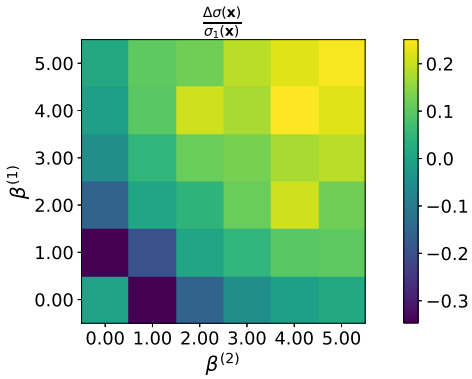
In the context of platforms' battle for users' attention, social media are interested in maximizing users' satisfaction (which translates in higher activity, thus revenue). We will now show how a market populated by two platforms achieves higher average satisfaction without the undesired drawback of increasing polarization. This is surprising, as one would expect that the additional degree of freedom increases social fragmentation, thus increasing polarization.

Suppose that each platform can vary its characteristics by tuning its $\beta^{(\gamma)}$[2]. First, we consider a single platform (i.e. $\Gamma = 1$) and consider $K = c = 3$, $r = 2$ and $\phi_i = 0.2\ \forall i$. In this setting, the highest single-platform average utility $\langle U \rangle_1$ is reached in $\beta^* = 3$, i.e. $\langle U \rangle_{\Gamma=1}(\beta^*) = \langle U \rangle_1$. On the other hand, a market populated by $\Gamma = 2$ platforms can result in an higher satisfaction. Figure 2.4 summarizes our results by showing the variation of average utility and standard deviation in passing from $\Gamma = 1$ to $\Gamma = 2$ for different values of $\beta^{(1)}$ and $\beta^{(2)}$. In particular, we define $\Delta\langle U \rangle(\beta^{(1)}, \beta^{(2)}) = \langle U \rangle_{\Gamma=2}(\beta^{(1)}, \beta^{(2)}) - \langle U \rangle_1$ and $\Delta\sigma(\mathbf{x}; \beta^{(1)}, \beta^{(2)}) = \sigma_{\Gamma=2}(\mathbf{x}; \beta^{(1)}, \beta^{(2)}) - \sigma_1(\mathbf{x})$, where $\sigma_1(\mathbf{x}) = \sigma_{\Gamma=1}(\mathbf{x}; \beta^*)$. Even if in the figure we reported, for the sake of completeness, the values corresponding to $\beta^{(1)} = 0$ or $\beta^{(2)} = 0$, in the following discussion we are going to neglect those points. The reason is that they correspond to a radicalized regime, which is highly undesired in

[2]We assume that $K^{(\gamma)} = K$ for all γ, implying that they all have the same political focus.



(a)



(b)

Figure 2.4: Relative increment passing from $\Gamma = 1$ to $\Gamma = 2$ platforms of (a) satisfaction and (b) opinions' standard deviation. Neglecting the first row and the last column ($\beta^{(1)} = 0$ or $\beta^{(2)} = 0$), which correspond to a radicalized regime, for "moderate" values of homophilic algorithm (e.g. $\beta^{(1)} = 1$ and $\beta^{(2)} = 2$), an increased satisfaction ($\approx 40\%$) is clear. Moreover, in such points, the standard deviation relative increse in negligible ($\approx 0\%$). For very high values of homophilic recommendations (e.g. $\beta^{(1)} = 5$ and $\beta^{(2)} = 5$), the satisfaction drops significantly while the standard deviation increases, making this the worst possible scenario.

a democracy, which lies on dialogue and debate. Thus, restricting the attention to$(\beta^{(1)}, \beta^{(2)}) \in [1, 5]^2$, the average utility increases for many couples of values with respect to the *best* possible single-platform average utility. The reason is that users can dynamically change their SMR, thus allocating their time among platforms to satisfy their desired differentiation ϕ. Moreover, focusing our attention on the point ($\beta^{(1)} = 2$, $\beta^{(2)} = 1$) (or, by symmetry, ($\beta^{(1)} = 1$, $\beta^{(2)} = 2$)), which is arguably very close to the two-platform optimum, we can see that the corresponding variation of po-

larization is almost zero, i.e. $\Delta\sigma(\mathbf{x}; 2, 1) \approx 0$. As we show in Fig. 2.6, the explanation for higher satisfaction without increasing polarization is the emerging quasi-segregation between moderate and extreme opinions. The moderate individuals, in absence of the extreme ones who populate another platform, get radicalized less, effectively reducing polarization. This is in qualitative agreement with findings on media habits and opinion stance, where individuals at both the left and right ends of the spectrum tend to be clustered around a single media source[3] [55]. So, by focusing only on users' satisfaction, the competition of platforms brings notable benefits. Finally, the points for which satisfaction decreases with respect to $\Gamma = 1$ are those where both $\beta^{(1)}$ and $\beta^{(2)}$ are large enough. In this regime, users connect only to very close peers, and can not find an alternative, cross-cutting platform. Thus, they cannot satisfy their desire for differentiation encapsulated in $\phi = 0.2$. In Appendix 2.6.4 Fig. 2.7 we show an alternative situation, where $\phi = 0.1$. In that case, users' desire for differentiation is too low and maximizing the utility leads to an increased polarization.

2.3 Discussion

Online social networks have become crucial for political news consumption [24, 25]. Existing studies [29, 30, 31, 32, 33, 7, 34, 35, 36, 37] explore the impact of recommendation algorithms on opinion dynamics, with some findings on Facebook's limited influence [29, 7]. However, news consumption spans various platforms [39], and individual preferences [40] play a role. It is thus crucial to consider the interplay between news repertoires and political communication processes to understand political opinions fully.

Here we show that when examining two platforms — one with a low political focus and a neutral recommendation algorithm, and another more politically oriented with a homophily-based algorithm — even if users spend the majority of their time on the neutral platform, opinion polarization can persist. This result casts some doubts on the generality of conclusions drawn from the recent study on the impact of Facebook recommendation algorithm on political opinions [7], which shows that people's political attitudes did not significantly change when they were exposed to more diverse content.

[3]Note however that there are some important differences between liberals and conservatives that the model cannot capture

Indeed, the consumption of political news on Facebook is incidental (low political focus), and the polarization might originate from (little) time spent on other news sources.

When we allow users to dynamically optimize their satisfaction by adjusting their SMR, the further segregation of individuals brought by multiple platforms leads to an increase of global satisfaction without, unexpectedly, an increase of polarization, with respect to the single platform case. In fact, more active (i.e. more extreme) users separate from users with "low" activity (i.e. more moderate), thus the latter are no longer influenced by the former, resulting in a reduced (or, at least, not increased) polarization.

While our multiplatform opinion dynamics model offers a comprehensive framework for understanding interactions across diverse platforms, it is imperative to acknowledge the inherent constraints and nuances that underpin its design. The following considerations delve into these aspects, shedding light on both the model's robustness and areas that warrant further refinement. In the first place, while our model describes social media platforms, there exist traditional media not captured by it, which influence users' opinions. Broadcasting platforms such as TVs and radios are an example, contributing to opinion formation in a way which is structurally and temporally different from social media. With reference to our model, they would act as hub nodes in the network of interactions, each diffusing a certain opinion to other nodes. However such nodes, unlike "normal" ones, do not change (or, rather, change very slowly) their opinions over time. Thus, effectively, traditional platforms can be seen as idiosyncratic relaxation terms in Eqn. 2.1, as opposed to having all opinions relax to zero $x_i(\infty) = x^* = 0$ when $K^{(\gamma)} = 0$ for all γ. The interplay between these two sources of information is left for future studies.

Regarding how users are affected by peers, while we assume that they converge (i.e. their opinions approach each other when interacting), other models [36, 56, 57, 58] consider also the presence of "polarizing" nodes, whose opinions move away if exposed to those of opposite sign. Since this characterization does not qualitatively change our results, we neglect it. In fact, as we have shown, both polarization and consensus emerge without the introduction of this additional feature.

Our last assumption is the conservation of activity over time, i.e. $a_i(t) = a_i \; \forall i, t$. In modeling users' satisfaction across platforms, it would be

reasonable to require that users' activity (which corresponds to users' engagement on social platforms) could decrease if they are not able to reach a certain degree of satisfaction (i.e. a certain value of their utility). However, modeling such behavior would require a clear definition of what a decreasing activity really mean — do users move their attention to traditional media? Do they allocate their SMR on a not-considered social platform? Or do they literally stop consuming news? — We believe that it is reasonable to assume the considered system (which, we note, can be made arbitrary large by increasing N and/or Γ) as an *isolated* social system, in which social energy (i.e. users' activity) is conserved[4]. In fact, it is possible to think that our model is valid in a span of time T in which activities remain constant, meaning that they change on a time scale much greater than T.

Finally, we want to stress how $\beta^{(\gamma)}$ encapsulates the role of the recommendation engine of platform γ. Typically, recommendation systems aim to infer users' needs, tastes and preferences, on the basis of their behavior on the platform, in order to suggest them the "best" product [60]. In terms of social media platforms, recommendation systems try to connect people who are "similar" according to some metric. Obviously, different platforms collect different kind of users' data, deploy different recommendation algorithms and have different purposes. Such dissimilarities translate into a non-homogeneous degree of recommendations' homophily. We imagine that the distances between users, evaluated by the recommendation algorithms by considering the huge amount of data social media collect, can be projected into 1-dimensional distances between political opinions, further distorted by the degree of homophily $\beta^{(\gamma)}$. Such assumption is based on the well-known phenomenon of issue alignment [61, 62, 46], i.e. individuals are much more likely to have a certain combination of opinions than others. On this note, an interesting extension of the model would be to consider a multi-topic opinion dynamics on a multiplex network; the idea would be that on a given layer/platform it might be more likely to talk about a topic than another.

[4]Although note that daily time spent on social media by internet users worldwide has been steadily increasing from 2012 to 2023[59].

2.4 Model

We consider a system of N agents, where each agent i has a one-dimensional continuous opinion variable $x_i(t) \in (-\infty, +\infty)$. The sign of x_i describes the agent's stance (e.g. being pro or against abortion). The absolute value of x_i quantifies the strength of this opinion: the larger $|x_i|$, the more extreme the stance of agent i. Moreover, we also consider that agents can interact on Γ different platforms. Each agent allocates his "time" in these Γ platforms. In the following sections we detail how opinions and interaction networks evolve and how each agent divides his attention among the platforms (see Appendix 2.6.1 for a detailed outline of the model simulation).

2.4.1 Opinions update

The opinion dynamics is driven by the interactions among agents, captured by a system of N coupled ordinary differential equations,

$$\dot{x_i} = -x_i + \sum_{\gamma=1}^{\Gamma} \Big(K^{(\gamma)} \sum_{j=1}^{N} A_{ij}^{(\gamma)}(t) \tanh\left(cx_j\right) \Big). \quad i \in \{1, ..N\} \tag{2.1}$$

$K^{(\gamma)} > 0$ represents the social interaction strength among agents on platform γ. The $\tanh(cx)$, with $c > 0$, embodies the fact that an agent i influences others in the direction of his own opinion, but such influence is "bounded". The term $A_{ij}^{(\gamma)}(t)$ is the entry of the $N \times N$ temporal Adjacency matrix $A^{(\gamma)}(t)$ corresponding to platform γ.

2.4.2 Network update

At every time step, user i can actively engage with the social media on platform γ with probability $a_i \rho_i^{(\gamma)}$, and/or passively engage on platform γ' with probability $p_i \rho_i^{(\gamma')}$ [5]. Active users contact users that are passive on the same platform, meaning that the opinions of the former affect the opinions of the latter. For example, if j is active on γ and contacts i, who is passive on γ, then $A_{ij}^{(\gamma)} = 1$. The term a_i (resp. p_i) is called *activity* (resp. *passivity*). Moreover, $\rho_i^{(\gamma)}$ represents the probability that user i, conditional on being active/passive, chooses platform γ. Of course, $\sum_{\gamma=1}^{\Gamma} \rho_i^{(\gamma)} = 1$. We assume

[5] Note that we can have $\gamma = \gamma'$.

that $a_i \in [\epsilon, 1]$ for all i, and that the activities are distributed according to a power law $F(a) \sim a^{-\eta}$. Moreover, we assume for simplicity that $p_i = 1\ \forall i$.

The temporal adjacency matrices $A_{ij}^{(\gamma)}(t)$ are assumed to evolve according to an activity-driven (AD) temporal network [63]. At each time step, each γ-active user contacts m γ-passive users. It is further assumed that these links are reciprocated with probability r. The probability $q_{ij}^{(\gamma)}$ that agent i contacts agent j on platform γ is given by the following expression:

$$q_{ij}^{(\gamma)} = \frac{|x_i - x_j|^{-\beta^{(\gamma)}}}{\sum_{k \in \mathcal{P}^{(\gamma)}} |x_i - x_k|^{-\beta^{(\gamma)}}} \tag{2.2}$$

where $\mathcal{P}^{(\gamma)}$ is the set of γ-passive users and $\beta^{(\gamma)} \geq 0$ captures the degree of homophily of γ's recommendation algorithm.

2.4.3 SMR update

While for a part of our results we considered $\rho_i^{(\gamma)} = \rho^{(\gamma)}$ homogeneous and constant in time, we also developed a model for which it changes on the basis of the observations of each user. In particular, we suppose that users allocate their "social energy" among platforms depending on their perceived quality. Grounded on the well-known psychological theory of optimal distinctiveness [52], individuals desire a balance of between assimilation (homophily) and differentiation (debate)(see [64] for an example of discrete opinion dynamics with optimal distinctiveness preferences). Borrowing from bounded confidence theory [65], an agent with opinion x considers those with opinion in $[x-r, x+r]$ contributing to assimilation, while the others to differentiation. Formally, if $\alpha_i^{(\gamma)}(t)$ (resp. $\delta_i^{(\gamma)}(t)$) is the number of in-degree connections contributing to assimilation (resp. differentiation) on γ at time t for user i[6], we define his utility as:

$$\begin{aligned} U_i(t_n) &= -(f_i(t_n) - \phi_i)^2 \\ &= -\left(\frac{\sum_\gamma \sum_{m=n-L+1}^{n} \delta_i^{(\gamma)}(t_m)}{\sum_\gamma \sum_{m=n-L+1}^{n} \delta_i^{(\gamma)}(t_m) + \sum_\gamma \sum_{m=n-L+1}^{n} \alpha_i^{(\gamma)}(t_m)} - \phi_i \right)^2 . \end{aligned} \tag{2.3}$$

[6] $\alpha_i^\gamma(t) + \delta_i^\gamma(t)$ is always equal to the total in-degree of node i.

Here $f_i(t)$ is the experienced distinctiveness, while ϕ_i the desired one. The parameter L encapsulates the time window over which users perceive the distinctiveness (which, in time units, is $\tau = Ldt$), thus representing their "memory". For example, consider $\Gamma = 2$, $L = 1$ and $\phi = 0.5$. If user i is connected to user j on platform 1 with $|x_j - x_i| < r$, and to user k on platform 2 with $|x_k - x_i| > r$ at time t_n, then it follows that $\delta_i^{(1)}(t_n) = \alpha_i^{(2)}(t_n) = 0$ and $\delta_i^{(2)}(t_n) = \alpha_i^{(1)}(t_n) = 1$. Thus, $f_i(t_n) = \phi_i = 0.5$ and $U_i(t_n) = 0$ is maximized.

It remains to specify how users maximize their utility. They can only control the fraction of time spent on each platform γ, proportional to $\rho_i^{(\gamma)}(t_n)$. We assume that users update their platform allocation every L steps, i.e. their preferences stay constant during the time interval over which assimilation and differentiation are experienced[7]. For this reason, each user can estimate the quality of platforms *given* her current preferences. Such estimates are in turn used to update the platform allocation, on the basis of an anticipated utility. It is reasonable to assume that each user acts as if the assimilation and differentiation experienced on each platform are proportional to the time spent on it. For this reason, it is possible to write:

$$\begin{aligned} \sum_{m=n-L+1}^{n} \delta_i^{(\gamma)}(t_m) &= L\omega_{\delta_i}^{(\gamma)}(t_n)\rho_i^{(\gamma)}(t_n), \\ \sum_{m=n-L+1}^{n} \alpha_i^{(\gamma)}(t_m) &= L\omega_{\alpha_i}^{(\gamma)}(t_n)\rho_i^{(\gamma)}(t_n). \end{aligned} \tag{2.4}$$

Here, $\omega_{\delta_i}^{(\gamma)}(t_n)$ and $\omega_{\alpha_i}^{(\gamma)}(t_n)$ are the differentiation and the assimilation slopes estimated by the user i. Formally, defining $\boldsymbol{\rho} = (\rho^{(1)}, \ldots, \rho^{(\Gamma)})$, each user aims to maximize:

$$\hat{U}_i(\boldsymbol{\rho}, t_n) = -\left(\frac{\boldsymbol{\omega}_{\delta_i}(t_n) \cdot \boldsymbol{\rho}}{\boldsymbol{\omega}_{\delta_i}(t_n) \cdot \boldsymbol{\rho} + \boldsymbol{\omega}_{\alpha_i}(t_n) \cdot \boldsymbol{\rho}} - \phi_i \right)^2, \tag{2.5}$$

where $\boldsymbol{\omega}_{\delta_i}(t_n) = (\omega_{\delta_i}^{(1)}(t_n), \ldots, \omega_{\delta_i}^{(\Gamma)}(t_n))$ is the vector of differentiation slopes estimated using user's previous interactions (an analogous definition holds for the assimilation slopes $\boldsymbol{\omega}_{\alpha_i}(t_n)$). $\hat{U}_i(\boldsymbol{\rho}, t_n)$ is the utility *estimated* by user

[7] In other words, users gather experience before changing their mind about a given social media.

i at time t_n. He assumes that it represents his future satisfaction (i.e. for $t > t_n$) depending on how he reallocates his $\boldsymbol{\rho}_i(t_{n+1})$. His assumption lies on hypothesizing that the slopes $\boldsymbol{\omega}_{\alpha_i}$ and $\boldsymbol{\omega}_{\delta_i}$ are roughly constant in time (which, instead, can vary due to the reallocation of all the other agents). User i updates then according to:

$$\boldsymbol{\rho}_i(t_{n+1}) = \begin{cases} \boldsymbol{\rho}_i(t_n) & \text{if } n/L \notin \mathbb{N} \\ \text{argmax}_{\boldsymbol{\rho}} \hat{U}_i(\boldsymbol{\rho}, t_n) & \text{otherwise.} \end{cases} \tag{2.6}$$

Clearly, the utility of user i depends not only on his $\boldsymbol{\rho}_i$, but also on how the other users have allocated their time on social media. Let us stress that in our model, users can only decide whether to be on a particular platform, but the connections are decided entirely by the recommendation algorithm. A justification for this comes from the work of [66], which shows that on Facebook the number of new links per day increased abruptly after the introduction of a "who to follow" recommendation algorithm. In other words, the individual agency in choosing connections is negligible with respect to the volume of content suggested by the platform itself.

2.4.4 Combined dynamics

We focus on a regime in which the three processes described above have different time scales. As already mentioned, we consider the network dynamics being much faster than the opinion dynamics. This is especially true in online social media context. In particular, for each network update we integrate Eq (2.1) for $dt = 0.001$. Moreover, the SMR dynamics lies between the two, representing the fact that the choice of the allocation among platforms is faster than the opinion dynamics, but requires a significant number of observations and interactions. Specifically, each user updates his preferences $\boldsymbol{\rho}_i(t_n)$ every $L = 100$ time steps. In short, every $L = 100$ network updates each user modifies her allocation preferences, and every 1000 network updates the opinions evolve of a unit time.

2.5 Conclusions

Our study examines the impact of social media competition for users' engagement on opinion dynamics. First, we show that opinion polarization

can persist as long as users spend a fraction of their time on a homophilic platform, highlighting the importance of multi-sources news diets. Second, we show that individual users' preferences interact in a non-trivial way with the recommendation algorithms in the presence of multiple platforms. The model indeed predicts the observed relationship between news outlet preference and political ideology. Interestingly, a multi-platform setup may be used to curb polarization while keeping user engagement intact. To this end, it is paramount to experimentally investigate users preferences for diversity (estimates for ϕ), either via surveys or controlled experiments. This avenue may help to shed light on healthy synergies between different social media platforms. From the revenue point of view, synergies are already well-known in this environment (think about users spending time on Whatsapp, Instagram and Facebook without ever going out from the Meta universe). Future research and efforts should thus gather cross-platform data, via survey [67, 51] or by experiments, in order to fully comprehend the subtle mechanism of opinion formation in online environments.

2.6 Appendix

2.6.1 Pseudo code

In each time step of the numerical algorithm the opinions, the temporal matrix and the SMR are update according to the following steps:

1. Each user i is only active with probability $a_i(1-p_i)$, only passive with probability $(1-a_i)p_i$, active and passive with probability $a_i p_i$ and inert with probability $(1-a_i)(1-p_i)$.

2. If active (resp. passive), user i chooses the platform γ (resp γ') on which he actively (resp. passively) engages with probability $\rho_i^{(\gamma)}(t)$ (resp. $\rho_i^{(\gamma')}(t)$) Note that it could be $\gamma = \gamma'$.

3. If active on platform γ, agent i influences m distinct agents $j \in \mathcal{P}^{(\gamma)}$ — where $\mathcal{P}^{(\gamma)}$ is the set of users passively engaged on platform γ — chosen according to Eqn. (2.2). This influence is expressed by updating the temporal adjacency matrix $A_{ji}^{(\gamma)}(t_n) = 1$.

4. With probability r the directed link is reciprocated, so that agent i receives influence from j, i.e. $A_{ij}^{(\gamma)}(t_n) = 1$.

5. Opinions x_i are updated by numerically integrating Eq. (2.1) using the total adjacency matrix elements $A_{ij}(t_n) = \sum_{\gamma=1}^{\Gamma} A_{ij}^{(\gamma)}(t_n)$.

6. Each user i collects the experienced assimilation and differentiation on each platform at the time step t_n as:

$$\alpha_i^{(\gamma)}(t_n) = \sum_{j \text{ s.t. } |x_i - x_j| < r} A_{ij}^{(\gamma)}(t_n)$$

$$\delta_i^{(\gamma)}(t_n) = \sum_{j \text{ s.t. } |x_i - x_j| > r} A_{ij}^{(\gamma)}(t_n).$$

7. If $\mod(t_n, L) \neq 0$, then the SMR remains constant for all users $\boldsymbol{\rho}_i(t_{n+1}) = \boldsymbol{\rho}_i(t_n)\ \forall i$. If $\mod(t_n, L) = 0$, each user updates his SMR according to the following steps:

   (a) estimate differentiation and assimilation slopes $\boldsymbol{\omega}_{\delta_i}(t_n)$ and $\boldsymbol{\omega}_{\alpha_i}(t_n)$ according to Eqn. (2.4)

(b) updates SMR according to $\boldsymbol{\rho}_i(t_{n+1}) = \text{argmax}_{\boldsymbol{\rho}} \hat{U}_i(\boldsymbol{\rho}, t_n)$, where $\hat{U}_i(\boldsymbol{\rho}, t_n)$ is defined in Eq. (2.5). To perform the utility maximization, we use a gradient descent algorithm on the Γ-dimensional simplex, consisting of n_{GD} iterations of learning rate Δ_{GD}. In particular, the following equation is iterated n_{GD} times:

$$\boldsymbol{\rho}_i(k+1) = \mathbf{P}_\Gamma\big(\boldsymbol{\rho}_i(k) - \Delta_{GD}\boldsymbol{\nabla}\hat{U}_i(\boldsymbol{\rho}_i(k), t_n)\big),$$

where $\mathbf{P}_\Gamma$ is the projection on the Γ-dimensional simplex, k runs from $k = 0$ to $k = n_{GD} - 1$, $\boldsymbol{\rho}_i(0) = \boldsymbol{\rho}_i(t_n)$ and $\boldsymbol{\rho}_i(n_{GD}) = \boldsymbol{\rho}_i(t_{n+1})$.

8. After each time step the temporal networks $A_{ij}^{(\gamma)}(t_n)$ are deleted.

Of course, when we considered a homogeneous and stationary (HS) SMR, steps 6. and 7. are ignored, i.e. $\boldsymbol{\rho}_i(t_n) = \boldsymbol{\rho}_i(t_0)\ \forall i, t$. As done in [37], we integrate Eq. (2.1) using an explicit fourth-order Runge-Kutta method with a time step of $dt = 0.01$ in the case of HS SMR, and $dt = 0.001$ in the case of evolving SMR. In the latter, we also consider $L = 100$. This leads to a timescale separation between the network dynamics, the SMR update and the opinion evolution mentioned in Sec. 2.4.4.

We independently sampled activities $\{a_i\}_{i=1}^N$ from the power law $F(a) = (1-\eta)a^{-\eta}/(1-\epsilon^{1-\eta})$, with parameters $\eta = 2.1$ and $\epsilon = 0.01$ [37]. We also set $p_i = 1\ \forall i$. Moreover, we consider the following parameters' values: $N = 800$, $r = 0.5$ and $c = 3$.

Throughout all of our simulations, we start with initial opinions $\{x_i(0)\}_{i=1}^N$ uniformly distributed in $[-1, 1]$. In the case of dynamical SMR, we let the opinions evolve for $n_{boot} = 2000$ steps before allowing users to reallocate themselves (i.e. we ignore point 7. for the n_{boot} steps before t_0). The reason is that we want to provide as input to our SMR update model an opinions' spectrum which is at (metastable, in the case of polarization) equilibrium . We initialize $\boldsymbol{\rho}_i(t_n) = 1/\Gamma\ \forall n \in \{-n_{boot}+1, \ldots, 0\}$. Then, at t_0, we "turn on" the SMR allocation described in point 7. of the pseudo-code.

2.6.2 Robustness of the phase diagram

Here we show the phase-diagram reported in the main text with $\beta^{(2)} = 2$, i.e. assuming the polarizing platform provides more cross-cutting content. By

comparing Fig. 2.3 and Fig. 2.5, we can see that the green region shrinks as $\beta^{(2)}$ decreases, in accordance with the meta-stability analysis of polarization reported in [37].

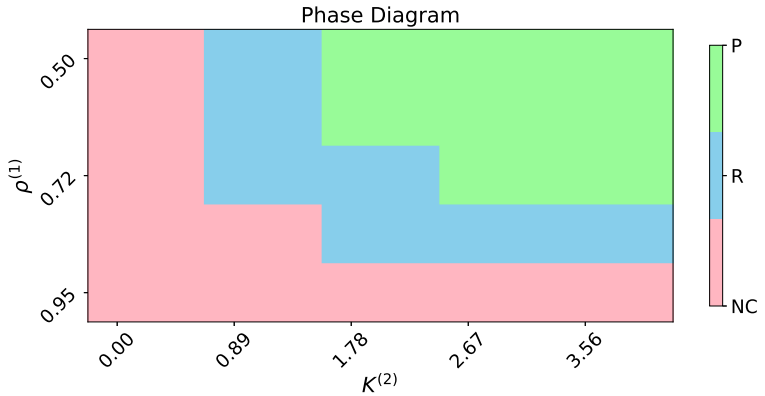


Figure 2.5: Phase diagram in the case $\beta^{(1)} = 0, \beta^{(2)} = 2$. The other parameters are as in the corresponding Fig. 2.3 reported in the main text.

2.6.3 Multi-platform segregation

To understand the benefits of the multi-platform reported in the main text, we define the following quantities.

$$\begin{aligned} f_\gamma^>(t) &= \frac{\sum_{i=1}^N \rho_i^{(\gamma)}(t)\mathcal{H}(a_i - \theta)}{\sum_{i=1}^N \mathcal{H}(a_i - \theta)} \\ f_\gamma^<(t) &= \frac{\sum_{i=1}^N \rho_i^{(\gamma)}(t)\mathcal{H}(\theta - a_i)}{\sum_{i=1}^N \mathcal{H}(\theta - a_i)}, \end{aligned} \tag{2.7}$$

where $\mathcal{H}$ is the Heaviside function. In words, $f_\gamma^>(t)$ (resp. $f_\gamma^<(t)$) is the "effective" share of users on platform γ whose activity is above (resp. below) a threshold θ. The idea is to understand whether these two classes of users exhibit qualitatively different behavior, notwithstanding their a-priori equal preferences (i.e. $\phi_i = \phi$ and $r_i = r$ for all users). Let us consider the 2 platform case reported in the main text, where $\beta^{(1)} = 2, \beta^{(2)} = 1$. In Fig. 2.6, we plot $f_1^>(t)$ and $f_1^<(t)$ for $\theta = 0.1$, which roughly divide the activity profile in 90 % below the threshold and 10% above. Highly active users deterministically prefer the more homophilic platform ($f_1^> \approx 1$), while less active users tend to explore both, though they prevalently occupy the

platform with more cross-cutting content ($f_1^< \approx 0.3$)

This is consistent with what we observe in reality, i.e. people with extreme opinions tend to have a more restricted outlet of like-minded sources, which further polarizes their view. On the other hand, more moderate users typically exhibit a more diverse news diet.

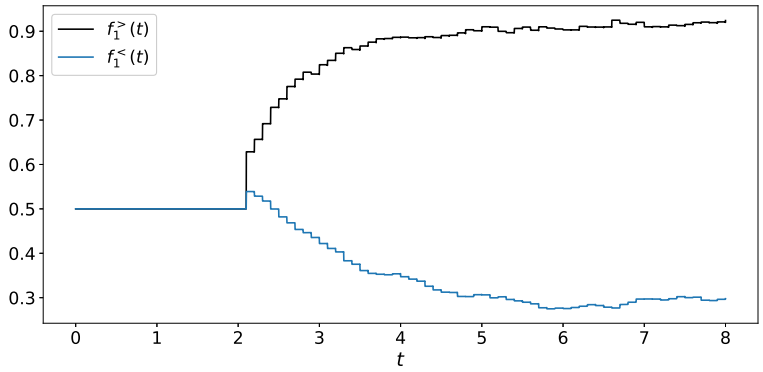
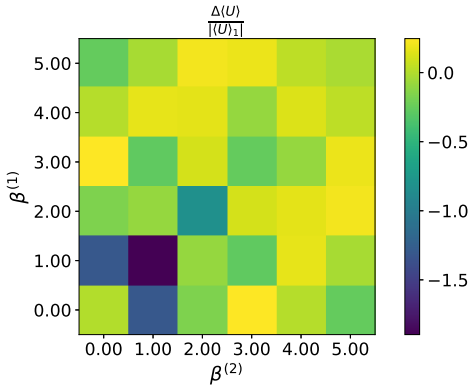


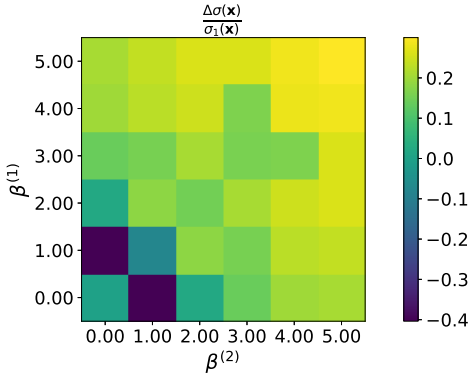
Figure 2.6: Effective share of users on platform 1 in the case where $\beta^{(1)} = 2, \beta^{(2)} = 1$. $\theta = 0.1$, so that approximately 90% of individuals have $a_i < \theta$.

2.6.4 Robustness with respect to desired distinctiveness

Here, we want to show how satisfaction and opinions standard deviation change with respect to the single-platform case for $\phi = 0.1$. The main physical difference with respect to the case presented in the main text (where $\phi = 0.2$) is that here users have an "halved" desire for differentiation, meaning that their maximal satisfaction requires a much high number of interactions contributing to assimilation, with respect to those contributing to differentiation. Figure 2.7 shows indeed how users' utility is maximized in those regimes for which they are more exposed to like-minded peers, i.e. for high values of $\beta^{(1)}$ and $\beta^{(2)}$. However, in this case, maximizing users' satisfaction leads to an increase in polarization, meaning that if users desire high confirmation opinions tend to extremize.



(a)



(b)

Figure 2.7: Relative increment of (a) satisfaction and (b) opinions' standard deviation, when passing from $\Gamma = 1$ to $\Gamma = 2$ platforms, for $\phi = 0.1$. In this regime, high values of both $\beta^{(1)}$ and $\beta^{(2)}$ lead to an increased users' satisfaction but also in an increased polarization (i.e. standard deviation).

Chapter 3

Learn your entropy from informative data: an axiom ensuring the consistent identification of generalized entropies

This chapter is based on: A. Somazzi and D. Garlaschelli. *Learn your entropy from informative data: an axiom ensuring the consistent identification of generalized entropies*. Available at `https://arxiv.org/abs/2301.05660`.

Shannon entropy, a cornerstone of information theory, statistical physics and inference methods, is uniquely identified by the Shannon-Khinchin or Shore-Johnson axioms. Generalizations of Shannon entropy, motivated by the study of non-extensive or non-ergodic systems, relax some of these axioms and lead to entropy families indexed by certain 'entropic' parameters. In general, the selection of these parameters requires pre-knowledge of the system or encounters inconsistencies. Here we introduce a simple axiom for any entropy family: namely, that no entropic parameter can be inferred from a completely uninformative (uniform) probability distribution. When applied to the Uffink-Jizba-Korbel and Hanel-Thurner entropy families, the axiom selects only Rényi entropy as viable. It also extends consistency with the Maximum Likelihood principle, which can then be generalized to estimate the entropic parameter purely from data, as we confirm numerically. Remarkably, in a generalized maximum entropy framework the axiom implies that the maximized log-likelihood always equals minus Shannon entropy, even if the inferred probability distribution maximizes a generalized entropy and not Shannon's, solving a series of problems encountered in previous approaches.

3.1 Introduction

The concept of entropy was introduced by Clausius in the thermodynamic framework [68] and later adopted in statistical physics by Boltzmann and Gibbs as a tool to describe macroscopic systems in terms of their probabilities of occupancy of microscopic states [69, 70]. Within information theory, Shannon axiomatically (re)derived the entropy as a quantification of the uncertainty encoded in a probability distribution, applicable to (among other things) the compressibility of sequences of symbols generated by ergodic probabilistic sources [12]. This allowed Jaynes to subsequently propose that the distribution that maximizes Shannon entropy, under the constraints implied by the empirical information available about a real system and realized via suitable Lagrange multipliers, provides the least biased (maximally noncommittal) inferential description of the unknown microscopic details of that system [8]. For systems whose physical entropy coincides with the Gibbs-Shannon one, this maximum entropy construction can be used to entirely reinterpret statistical physics from an information-theoretic viewpoint. In modern research, statistical inference and model identification based on entropy maximization are perfectly consistent with Maximum Likelihood estimation methods and are at the heart of several machine-learning techniques [71].

Several generalizations of Shannon entropy (the most popular of which were motivated by the statistical physics of non-extensive and/or non-ergodic systems) have been proposed in various contexts [72, 73, 74, 75, 76], resulting in extended families of entropy that depend, besides the usual 'structural' parameters (the Lagrange multipliers), on extra 'entropic' parameters that label the specific member of the entropy family. For a fixed choice of these parameters, one can still maximize the resulting entropy and generalize the inference procedure. This is possible when there is enough knowledge *a priori* about the system, so that the entropic parameter can be set 'by hand' to the correct value. However, the physical entropy and the information-theoretic one may in general no longer coincide for non-extensive or non-ergodic systems [77]. Moreover, it is generally not possible to maintain compatibility with the Maximum Likelihood principle and, crucially, to infer the values of the entropic parameters purely

from data without encountering inconsistencies, making the generalized methodology inapplicable without prior knowledge of the correct entropy.

In this chapter we discuss and alleviate those inconsistencies by introducing an axiom that restricts the form of parametric families of information-theoretic entropies. The axiom enforces a simple 'uninformativeness' requirement and allows for the consistent inference of the entropic parameters purely from the available data, as we show via analytical results and numerical examples. The chapter is organized as follows. In Sec. 3.2 we first review the theoretical background behind the axiomatic definitions of entropy, the Maximum Entropy Principle, and the Maximum Likelihood Principle. In Sec. 3.3 we then discuss the main contributions of the chapter, i.e. the introduction of the new axiom and its implications for the selection of information-theoretic entropies from certain popular families, the restoration of consistency with the Maximum Likelihood principle, and the generalization of the latter in order to infer the entropic parameter(s) purely from the data. Finally, in Sec. 3.4 we offer some concluding remarks.

3.2 Theoretical background

3.2.1 The Shannon-Khinchin axioms

Given a distribution (technically, a probability mass function) $P = (p(G_1), \ldots, p(G_\Omega))$, where $p(G_i)$ is the probability that the discrete random variable G takes the i-th outcome (or 'state') G_i, Ω is the total number of distinct outcomes, and clearly $\sum_{i=1}^{\Omega} p(G_i) = 1$, Shannon entropy $S[P]$ is axiomatically defined through the following four Shannon-Khinchin (SK) axioms [13]:

- *SK1 (continuity):* $S[P]$ is continuous in the entries of P.
- *SK2 (maximality):* $S[P]$ is maximal when P is the uniform distribution $P_u \equiv (\Omega^{-1}, \ldots, \Omega^{-1})$.
- *SK3 (expansibility):* $S[P]$ is expansible, i.e. it does not change if for the variable G an $(\Omega+1)$-th outcome with zero probability ($p(G_{\Omega+1}) = 0$)

is added:

$$S[(p(G_1),\ldots,p(G_\Omega))] = S[(p(G_1),\ldots,p(G_\Omega),0)].$$

- *SK4 (separability):* the entropy of the joint distribution $R = \big(r(G_1^{(1)},G_1^{(2)}),\ldots,r(G_{\Omega^{(1)}}^{(1)},G_{\Omega^{(2)}}^{(2)})\big)$ of two variables $G^{(1)}$ and $G^{(2)}$ with marginal distributions $P = (p(G_1^{(1)}),\ldots,p(G_{\Omega^{(1)}}^{(1)}))$ and $Q = \big(q(G_1^{(2)}),\ldots,q(G_{\Omega^{(2)}}^{(2)})\big)$ respectively, where $p(G_i^{(1)}) = \sum_{j=1}^{\Omega^{(2)}} r(G_i^{(1)},G_j^{(2)})$ and $q(G_j^{(2)}) = \sum_{i=1}^{\Omega^{(1)}} r(G_i^{(1)},G_j^{(2)})$, separates as

  $$S[R] = S[P] + S[Q|P].$$

  Here $S[Q|P]$ is the conditional entropy of Q on P, defined as $S[Q|P] = \sum_{k=1}^{\Omega^{(1)}} p(G_k^{(1)})S[Q_{|k}]$ with $Q_{|k} = \big(r(G_k^{(1)},G_1^{(2)})/p(G_k^{(1)}),\ldots,r(G_k^{(1)},G_{\Omega^{(2)}}^{(2)})/p(G_k^{(1)})\big)$ denoting the conditional distribution of the events in Q on the k-th event in P. Note that in particular, if the two events are independent ($Q = Q_{|k}$ for all k), then $S[R] = S[P]+S[Q]$, in which case separability becomes *additivity.*

It is possible to show that the only functional form of $S[P]$ respecting the four *SK* axioms is Shannon entropy:

$$S_1[P] = -\sum_{i=1}^{\Omega} p(G_i)\ln p(G_i), \tag{3.1}$$

where the subscript 1 will be justified later. The above expression is unique up to a positive overall multiplicative factor k, which is inessential from the information-theoretic point of view but is important for the identification (when appropriate) with the physical entropy, in which case k carries physical units and coincides with Boltzmann constant. As required by *SK2*, the maximum value of $S_1[P]$ is attained by the uniform distribution P_u, leading to Boltzmann entropy:

$$S_1[P_u] = \ln\Omega. \tag{3.2}$$

No distribution P can be such that $S_1[P] > S_1[P_u]$.

3.2.2 The Maximum Entropy Principle

The informational entropy $S_1[P]$ in Eq. (3.1) coincides (up to Boltzmann's constant) with the physical entropy derived by Gibbs [70], which in turn generalizes Boltzmann entropy in Eq. (3.2) [69]. For ergodic and short-range interacting systems, this equivalence is not coincidental and is rooted in statistical inference, as Jaynes showed with the introduction of the Maximum Entropy Principle (MEP) [8]. The MEP states that, given only a set I of pieces of empirical information about a system (in the physical situation, this typically means the knowledge of a few, macroscopic conserved quantities such as the total energy and/or the total number of particles), one should assign the possible microscopic states a probability distribution P that maximizes the entropy. In other words, entropy can be used as an inference functional whose maximization minimizes bias and prevents arbitrariness.

In particular, consider a system with a set of Ω potential microstates $\{G_i\}_{i=1}^{\Omega}$ and assume that the available information I is encoded in the empirical value $C^* = C(G^*)$ of a certain (scalar or vector) function C of the microstate of the system, where G^* is the 'true' (unobservable) microstate. For the moment, let us assume that C^* is the only observation available (later, we will consider multiple independent observations of the same variable). Since G^* is unknown, the microstate is treated as a random variable G. The MEP applied to $S_1[P]$ identifies the maximum entropy distribution for G, which we denote as $P_0 = (p_0(G_1), \ldots, p_0(G_\Omega))$ or $P_1 = (p_1(G_1), \ldots, p_1(G_\Omega))$, depending on whether C^* is treated as a 'hard' or 'soft' constraint, respectively.

In the case of hard constraints (*microcanonical ensemble*), only a restricted number $\Omega_{C^*} < \Omega$ of microstates i for which $C(G_i)$ matches C^* exactly are assigned a non-zero probability, which (due to *SK2* and *SK3*) has to be uniform over the restricted support, i.e. $p_0(G_i) = \Omega_{C^*}^{-1}$ if $C(G_i) = C^*$ and $p_0(G_i) = 0$ otherwise. The resulting entropy is

$$S_1[P_0] = \ln \Omega_{C^*} < S_1[P_u]. \tag{3.3}$$

Unfortunately, calculating Ω_{C^*} is generally a hard combinatorial problem,

which makes the microcanonical ensemble not amenable to analytical calculations. For this reason, soft constraints are considered more often in the literature, as we also do in this chapter.

In the case of soft constraints (*canonical ensemble*), only the expected value $\langle C \rangle$ of the observable is constrained to match C^*, i.e.

$$\langle C \rangle \equiv \sum_{i=1}^{\Omega} p(G_i)\, C(G_i) = C^*, \tag{3.4}$$

thus allowing for the full set of Ω microstates, however with a non-uniform probability $p(G_i)$ yet to be determined. To find the specific probability $p_1(G_i)$ maximizing S_1 under the soft constraint above, one can introduce the Lagrange multiplier θ (which has the same dimensionality as C), plus an additional scalar multiplier α enforcing the normalization of P, and look for the specific values (denoted as P_1, θ_1, α_1) for which all the derivatives of the Lagrangian function

$$\mathcal{L}_1[P] \equiv S_1[P] - \alpha \left[\sum_{i=1}^{\Omega} p(G_i) - 1 \right] - \theta \cdot [\langle C \rangle - C^*] \tag{3.5}$$

vanish (the notation $\theta \cdot C$ indicates the scalar product). Setting $\partial \mathcal{L}[P]/\partial P|_{P_1} = 0$, i.e. $\partial \mathcal{L}[P]/\partial p(G_i)|_{p_1(G_i)} = 0 \ \forall i$, leads to the functional form of P_1, which turns out to be the well-known Boltzmann-Gibbs distribution with entries

$$p_1(G_i, \theta) = \frac{e^{-\theta \cdot C(G_i)}}{Z_1(\theta)}, \quad Z_1(\theta) = \sum_{j=1}^{\Omega} e^{-\theta \cdot C(G_j)}, \tag{3.6}$$

where $Z_1(\theta)$ is the *partition function*, resulting from the normalization constraint

$$\left. \frac{\partial \mathcal{L}[P_1]}{\partial \alpha} \right|_{\alpha_1} = 0 \quad \Rightarrow \quad \sum_{i=1}^{\Omega} p_1(G_i, \theta) = 1 \tag{3.7}$$

which leads to

$$\alpha_1 = -1 + \ln Z_1(\theta) \tag{3.8}$$

independently of the value of θ.

Importantly, P_1 is not identified entirely, until the parameter θ is also determined. This is attained by enforcing the vanishing of the remaining derivatives, identifying the value θ_1 realizing Eq. (3.4):

$$\left.\frac{\partial \mathcal{L}[P_1]}{\partial \theta}\right|_{\theta_1} = 0 \quad \Rightarrow \quad \sum_{i=1}^{\Omega} p_1(G_i, \theta_1)\, C(G_i) = C^*, \tag{3.9}$$

where, if θ is a vector, the notation means again that all the derivatives of $\mathcal{L}[P]$ with respect to the components of θ vanish separately. The final solution to the MEP problem is therefore given by inserting θ_1 into Eq. (3.6), and we will denote it as $P_1(\theta_1) = (p_1(G_1, \theta_1), \ldots, p_1(G_\Omega, \theta_1))$. The MEP with soft constraints, which are appropriate when the observables are expected to fluctuate, has been used successfully for inference and model selection in many fields beyond physics, including network theory, neuroscience, economics and biology [78, 11].

3.2.3 The Maximum Likelihood Principle

It is very important to realize that the MEP procedure outlined above has deep connections and desirable consistencies with the Maximum Likelihood (ML) principle, which applies to more general (not necessarily maximum-entropy) parametric probability distributions and states that the optimal parameter value θ^* is the one maximizing the log-likelihood on the data G^*. Our parametric model is the exponential family from Eq. (3.6) and the ML principle would select the value

$$\theta_1^* = \underset{\theta}{\operatorname{argmax}}\, \ell_1(\theta), \quad \ell_1(\theta) \equiv \ln p_1(G^*, \theta). \tag{3.10}$$

As a first result, it is easy to show that the value θ_1^* defined by Eq. (3.10) coincides with the value θ_1 defined by Eq. (3.9) [79, 80], i.e. $\theta_1^* \equiv \theta_1$ (in our notation, the asterisk next to a parameter will always denote the ML value of that parameter), i.e.

$$\left.\frac{\partial \ell_1(\theta)}{\partial \theta}\right|_{\theta_1^*} = 0 \quad \Rightarrow \quad \sum_{i=1}^{\Omega} p_1(G_i, \theta_1^*)\, C(G_i) = C^* \tag{3.11}$$

in analogy with Eq. (3.9). This means that the ML principle can be seen as equivalent to the part of the Lagrangian optimization relative to θ.

Moreover, it is straightforward to show that the maximized log-likelihood equals minus the entropy:

$$S_1[P_1(\theta_1^*)] = -\ell_1(\theta_1^*), \tag{3.12}$$

which is the counterpart of Eq. (3.3) in the case of soft constraints. This relationship is very important, because the maximized likelihood is at the basis of model selection criteria [81, 82]: if alternative models (i.e. alternative parametric probability distributions) are compared against the same empirical data, the model to be preferred (assuming all models have the same complexity, e.g. the same number of parameters) is the one with highest maximized likelihood. Then, Eq. (3.12) ensures that the ranking of models based on ML is the same as the ranking based on minus their entropy: the least uncertain (i.e. most informative) model has to be preferred. For models with different numbers of parameters and/or functional forms, the ranking based on likelihood/entropy has to be revised by adding a term controlling for the variable model complexity, leading to criteria such as AIC, BIC, the Minimum Description Length, etc. [81, 82] (for simplicity, we will not consider this situation here). Also note that, when the maximum entropy distribution $P_1(\theta_1^*)$ is inserted into Eq. (3.5), we get

$$\mathcal{L}_1[P_1(\theta_1^*)] = S_1[P_1(\theta_1^*)] = -\ell_1(\theta_1^*), \tag{3.13}$$

so that the Lagrangian, evaluated at $P_1(\theta_1^*)$, coincides with minus the maximized log-likelihood and can therefore be used to rank alternative models as well. All the above results indicate that the MEP can be used as a model selection criterion, exactly as the ML principle, by ranking models based on their realized entropy.

It is also important to consider the case when there are M independent observations $\{C_m^*\}_{m=1}^M$ about the system, which technically means that there are M independent and identically distributed (i.i.d.) realizations $\{G_m^*\}_{m=1}^M$ of the microstate G (recall that G is treated as a random variable), on each of

which the quantity $C^*_m = C(G^*_m)$ $(m = 1, M)$ is observed. Clearly, since the system being observed multiple times is one, the probability distribution characterizing it must still be specified by a single value of the Lagrange multiplier θ coupled to the quantity C. It should at this point be noted that the principle that identifies how to optimally combine the M observations $\{C^*_m\}_{m=1}^M$ in order to estimate θ is not the MEP, but the ML one. Indeed, the ML principle applied to the joint log-likelihood $\sum_{m=1}^M \ln p_1(G^*_m, \theta)$, or equivalently to the average log-likelihood $\overline{\ell}_1(\theta) \equiv \sum_{m=1}^M \ln p_1(G^*_m, \theta)/M$, can be formulated by replacing Eq. (3.10) with

$$\theta^*_1 = \underset{\theta}{\operatorname{argmax}}\, \overline{\ell}_1(\theta), \quad \overline{\ell}_1(\theta) \equiv \frac{\sum_{m=1}^M \ln p_1(G^*_m, \theta)}{M}. \tag{3.14}$$

It is easy to show that the condition $\partial\overline{\ell}_1(\theta)/\partial\theta|_{\theta^*_1} = 0$ identifying θ^*_1 leads to the well-known result

$$\langle C \rangle = \frac{1}{M} \sum_{m=1}^{M} C^*_m \tag{3.15}$$

where the (arithmetic) sample average of the M observations has emerged. So, in order to find the ML parameter value θ^*_1, one should replace Eq. (3.4) with Eq. (3.15), or equivalently redefine C^* in Eq. (3.4) as the sample average of $\{C^*_m\}_{m=1}^M$. In plain words, the sample average is 'produced' by the ML principle. On the contrary, within the MEP construction, there is no way of 'telling' Eqs. (3.5) and (3.9) what, in case of M observations, the meaning and definition of C^* should be. So in this case the ML principle is more informative than the MEP; this is another reason why one wants the entropy to be fully consistent with what the ML principle leads to. In particular, it is easy to show that, due to the independence of the M samples, the maximized average log-likelihood $\overline{\ell}_1(\theta^*_1)$ is still equal to minus the entropy:

$$S_1[P_1(\theta^*_1)] = -\overline{\ell}_1(\theta^*_1), \tag{3.16}$$

generalizing Eq. (3.12). Note that there is no microcanonical counterpart of Eq. (3.16), since Eq. (3.3) cannot be generalized to the case $M > 1$, unless all the M values $\{C^*_m\}_{m=1}^M$ are identical. Indeed the microcanonical ensemble cannot be constructed, because by definition it cannot account for different realizations of the values of the constraints: in case of different observations of the same constraints, only the canonical ensemble is feasible.

The above discussion clarifies that it is important that the entropy is consistent with the maximized log-likelihood, because the ML principle is needed both for model selection and for the determination of how multiple observations of the same system should be combined in order to optimally estimate the parameters.

3.2.4 The Shore-Johnson axioms

An alternative axiomatic definition of an entropy functional, whose maximization in presence of a set I of pieces of information should lead to a probability distribution $P = \circ I$ with certain properties, was proposed by Shore and Johnson (SJ) through the following axioms [9]:

- *SJ1 (uniqueness):* given I, $P = \circ I$ is unique.
- *SJ2 (invariance):* if $\Gamma[\cdot]$ is a coordinate transformation (change of variables), then $\Gamma[\circ I] = \circ(\Gamma[I])$.
- *SJ3 (system independence):* given two independent systems A and B, it should not matter whether one accounts for distinct pieces of information about them separately (in terms of marginal probabilities) or jointly (in terms of a joint probability). This means $\circ(I_A \wedge I_B) = (\circ I_A)(\circ I_B)$, where $I_A \wedge I_B$ denotes the union of the available pieces of information I_A and I_B about A and B respectively.
- *SJ4 (subset independence):* it should not matter whether one treats an independent subset of system states in terms of a separate conditional density or in terms of the full system density. Consider a partition of the system's states into disjoint subsets $\{\Lambda_k\}_k$ such that $\bigcup_k \Lambda_k = \Omega$, for each k of which there is a piece of information I_k available. Then $(\circ I)_{\Lambda_k} = \circ I_k \ \forall k$, where $I = \bigwedge_k I_k$ is the total information, and $P_{\Lambda_k} = (p_{\Lambda_k}(G_1), \ldots, p_{\Lambda_k}(G_\Omega))$, where $p_{\Lambda_k}(G_i) = p(G_i | G_i \in \Lambda_k)$ denotes the conditional distribution relative to the subset Λ_k.
- *SJ5 (maximality)*[1]*:* with no information available ($I = \varnothing$), $P = \circ I$ is

[1] Actually, Shore and Johnson defined the maximality axiom only implicitly. Indeed, starting from the principle of minimum cross-entropy, they introduced the MEP as its equivalent in the case where the prior distribution is uniform. For this reason, even if not explicitly axiomatized, they considered the posterior P to be equal to the uniform distribution (i.e. the same as the prior) when no information is available.

the uniform distribution P_u.

Shore and Johnson claimed that Shannon entropy is the only inference functional compatible with their axioms, a statement suggesting the equivalence of the *SK* and the *SJ* axioms. However, it was later clarified [10, 14] that Shore and Johnson's conclusion was due to an additional hidden assumption they made inadvertently when formally using *SJ3* in their reasoning. Specifically, they considered a situation where distinct pieces of information I_A and I_B are known about two systems A and B, and implied that the resulting joint probability factorises as $\circ(I_A \wedge I_B) = (\circ I_A)(\circ I_B)$, thereby applying *SJ3* even if the independence of the two systems is not guaranteed (having only disjoint pieces of information about two systems does not guarantee that the two systems are independent) [10, 14]. The presence of this additional assumption implies that Shannon entropy is in fact the desired functional only when systems are independent: if this is not the case, then the resulting maximum entropy distribution is no longer 'maximally non committal with respect to missing information', as Jaynes' MEP demands it to be [8], because there is actually no 'information' available about the (in)dependence of the systems.

3.2.5 Generalized entropies

Uffink [10] showed that, if Shore and Johnson's proof is correctly revisited without the extra unjustified assumption, the entropy resulting from the *SJ* axioms is not uniquely determined and is actually an entire generalized family $S_q^{(f)}[P]$, given by any increasing function f of a certain functional $U_q[P]$ that we will call the Uffink functional, i.e.

$$S_q^{(f)}[P] = f(U_q[P]), \quad U_q[P] = \left(\sum_{i=1}^{\Omega} p^q(G_i) \right)^{\frac{1}{1-q}} \tag{3.17}$$

for some parameter $q > 0$. For a given f, each entropy in the family is identified by the parameter q, which we will therefore call the 'entropic parameter'. Note that an entropic parameter plays a different role with respect to other structural parameters entering the entropy, such as θ in the Shannon case discussed above. Clearly, Shannon entropy is one of the possible members of this family. Indeed, taking $f(x) = \ln x$, one can show

that

$$\begin{aligned}\lim_{q\to 1} S_q^{(\ln)}[P] &= \lim_{q\to 1} \ln U_q[P] \\ &= -\sum_{i=1}^{\Omega} p(G_i) \ln p(G_i) \\ &= S_1[P]. \end{aligned} \tag{3.18}$$

In other words, Shannon entropy formally corresponds to $q = 1$, justifying the subscript adopted in Eq. (3.1) (note instead that the subscript in the uniform distribution P_0 used in Sec. 3.2.2 to describe the microcanonical distribution under hard constraints has nothing to do with the case $q = 0$, which is inadmissible). Notably, Jizba and Korbel [14, 83] showed that an entropy of the type $f(U_q[P])$ can also be obtained from the *SK* axioms, provided that *SK4* is relaxed to a generalized separability condition where the sum is replaced by the so-called Kolmogorov-Nagumo sum[2], previously introduced in the context of generalized arithmetics [84, 85]. This shows that the *SJ* axioms are actually equivalent to a specific generalization of the *SK* ones. The generalization of *SK4* has been a matter of discussion in the statistical physics literature for decades, as it relates to the subject of non-extensive (or rather non-additive) thermodynamics [72]. We will call any entropy of the form $f(U_q[P])$ an Uffink-Jizba-Korbel (*UJK*) entropy.

Several other generalized families of entropy resulting from relaxations of the *SK* or *SJ* axioms have been proposed [74, 75]. A notable example is the so-called (c, d)-entropies $S_{c,d}[P]$ introduced by Hanel and Thurner [86, 76] by replacing *SK4* with the assumption of *trace-form* (or more in *general composable*) entropies, i.e. entropies that can be written as (functions of) a sum over the states $\{G_i\}_{i=1}^{\Omega}$ of the system. In particular, an entropy $S(P)$ is trace-form if it can be written as a sum $\sum_{i=1}^{\Omega} g\left(p(G_i)\right)$ for some function

[2]Considering a bijection $f^{-1} : M \mapsto N \subset \mathbb{R}$, the generalized arithmetics is defined as follows:

$$\begin{aligned} x \oplus y &= f(f^{-1}(x) + f^{-1}(y)), \\ x \ominus y &= f(f^{-1}(x) - f^{-1}(y)), \\ x \otimes y &= f(f^{-1}(x) f^{-1}(y)), \\ x \oslash y &= f(f^{-1}(x) / f^{-1}(y)). \end{aligned}$$

g. Note that Shannon entropy is in this class, with $g(x) = -x \ln x$. More generally, a composable entropy can be written as a function h of such a sum, i.e.

$$S_{c,d}^{(h,g)}[P] = h\left(\sum_{i=1}^{\Omega} g\left(p(G_i)\right)\right), \tag{3.19}$$

where the entropic parameters (c, d) are determined by how the entropy scales with the number Ω of accessible configurations [73, 86]. In particular, one considers the transformations $\Omega \to \lambda\Omega$, $\Omega \to \Omega^{1+a}$ and identifies c and d from the following limiting ratios:

$$\lim_{\Omega\to\infty} \frac{S_{c,d}^{(h,g)}[(p(G_1), ..., p(G_{\lambda\Omega}))]}{S_{c,d}^{(h,g)}[(p(G_1), ..., p(G_{\Omega})]} = \lambda^{1-c}, \tag{3.20}$$

$$\lim_{\Omega\to\infty} \frac{S_{c,d}^{(h,g)}[(p(G_1), ..., p(G_{\Omega^{1+a}})]}{S_{c,d}^{(h,g)}[(p(G_1), ..., p(G_{\Omega}))]} \Omega^{a(c-1)} = (1+a)^d. \tag{3.21}$$

Different choices of h and g may result in the same values of the entropic parameters, in which case the corresponding entropies are considered asymptotically equivalent [73]. Therefore in this case the entropic parameters identify equivalence classes of entropies with the same asymptotic properties. We will call the entropies that respect *SK1-SK3*, plus Eq. (3.19), the Hanel-Thurner (HT) entropies.

3.2.6 How to identify the correct entropy?

On *UJK* entropies, *HT* entropies and in principle any generalized entropy family, it is important to ensure that the MEP can be reformulated consistently as a tool to construct probability distributions starting from observations of the system. This procedure is sometimes called the Generalized Maximum Entropy Principle (GMEP). However, a number of serious conceptual and practical problems are currently open.

First, while it is still possible, for a fixed value of the entropic parameter(s), to identify the functional form of the probability distribution maximizing the generalized entropy under certain 'soft' constraints, it is no longer guaranteed in general that the enforcement of these constraints remains consistent with the application of the ML principle to the Lagrange

multipliers and that the entropy retains a role for model selection as in Eq. (3.12). Only for certain generalized entropies this consistency is retrieved, but not for all of them, as we show later with some notable examples. Since the ML principle is agnostic with regard to the form of the probability distribution, and even more so to the type of entropy the latter maximizes, this inconsistency raises suspicion. Unfortunately, its possible origin is poorly discussed in the literature.

Second, fundamental problems arise when considering the determination of the entropic parameters themselves, or in other words of the 'correct' entropy in a parametric family. In particular, two main approaches have been proposed. One approach requires some *a priori* knowledge of the system (e.g. how certain properties of the entropy or of the system change with the number of accessible configurations [86, 73, 87]) as in Eqs. (3.20) and (3.21), implying that, in absence of such knowledge, the entropic parameters cannot be consistently derived purely from data as the other parameters (in this approach, the knowledge of the entropic parameter is viewed as a different type of information, besides the information obtained by the empirical measurement). Another approach does allow for the entropic parameters to be inferred from data, again invoking some form of maximization of the generalized entropy [88, 89]. However, as we show below, this requirement conflicts with the ML principle, if the latter is extended to the estimation of the entropic parameters themselves. Finally, several analyses estimate the entropic parameter(s) by assuming a certain form of the entropy and fitting the resulting maximum-entropy probability on empirical distributions [90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103]. As we will show, this approach does not guarantee consistency between the maximized likelihood and the original entropy. Moreover, *maximum-entropy distributions, even when optimally fitted to the data, do not uniquely identify the entropy they maximize,* because they also maximize *any other* entropy functional that is a monotonic function of that entropy. This point becomes particularly critical when such monotonic function depends on the entropic parameter(s) themselves, as we will show below.

3.3 One axiom to rule them all

The above limitations make the GMEP either inapplicable in practice without prior knowledge of the correct entropic parameter(s), or inconsistent with the ML principle and the information-theoretic consequences of *SJ3* under independence. In the rest of this section, which contains our main results, we show that a possible solution to this problem can be achieved starting from a seemingly different viewpoint, i.e. by imposing an additional axiom that somehow 'aligns' all entropies in a given family and therefore allows to select the most likely member of the family purely from data (if the latter contain information) and without prior knowledge of the system's properties. Remarkably, the introduction of this simple requirement solves all the inconsistencies discussed in Sec. 3.2.6.

3.3.1 The uninformativeness axiom

We now introduce the axiom. Unlike the *SK* or *SJ* ones, this axiom applies not to an individual entropy in a generalized parametric family, but rather to the entire family. Indeed the axiom does not represent yet another generalization of the *SK* or *SJ* ones, but rather an 'auxiliary' requirement to be added precisely when any such generalization is made, to restrict the form of the resulting entropic family.

- *Uninformativeness Axiom*: In a parametric family of information-theoretic entropies, the value of the entropy attained by the uniform distribution P_u should not depend on the value of the entropic parameter(s).

Clearly, if the axiom is applied to families that include Shannon entropy S_1 as a particular case, it implies that all members of the family attain the same value $S_1[P_u] = \ln \Omega$ when applied to P_u. This requirement equips generalized entropies with a universal scale and meaning. As we show below, our axiom provides certain guarantees when the inference procedure is extended to the identification of the entropic parameters themselves. On one hand, the axiom ensures that no entropic parameter can be inferred from a completely uninformative (i.e. uniform) distribution, irrespective of how the parameter estimation procedure is conceived. On the other hand, when informative (non-uniform) data are available, the axiom ensures

consistency with a generalized ML principle and model selection approach where all parameters, including the entropic one, can be identified from empirical observations, without prior knowledge of the system. Note that, in general, the physical entropy characterizing the real system may be different from the information-theoretic one identified by our axiom; nonetheless, our axiom ensures that the maximum entropy distribution that best describes the physical observations can be identified consistently from the information-theoretic entropy, without prior knowledge of the physical entropy itself.

Note that, as required by *SK2* and *SJ5*, for a given value of q the Uffink functional in Eq. (3.17) is maximized by P_u. This requirement comes from a 'horizontal' perspective, in the sense that it holds for each q-entropy in the family. Our axiom, on the other hand, provides a 'vertical' perspective: among all the q-entropies, none of them has to be preferred when applied to P_u. In other words, the axiom ensures the uninformativeness role of the uniform distribution not only for a specific entropy in the family, but across all of them. Since *SK2* and *SJ5* ensure that no entropy can exceed the value it attains on P_u, the axiom establishes a sort of common reference frame or universal scale, which allows to compare different entropies in a parametric family consistently. In particular, it ensures that all entropies in a parametric family that respects *SK2* or *SJ5* and includes Shannon entropy as a particular case attain values in the same interval $[0, \ln \Omega]$, irrespective of the value of the entropic parameter(s). We will show that this guarantee ensures that the entropic parameter(s) can be estimated via a model selection approach purely from the input data, if the latter are informative (non-uniformly distributed).

3.3.2 Application to important entropy families

We now discuss some consequences of imposing the uninformativeness axiom to popular entropy families.

We start with the *UJK* entropies $S_q^{(f)}[P]$ under the requirement that the family should include Shannon entropy as a particular case. The entropy $S_q^{(f)}[P] = f(U_q[P])$, when evaluated on the uniform probability distribution

$P_u = (\Omega^{-1}, \ldots, \Omega^{-1})$, returns the value

$$S_q^{(f)}[P_u] = f\left(U_q[P_u]\right) = f(\Omega) \qquad \text{for} \quad q \neq 1. \tag{3.22}$$

Our axiom requires that $S_q^{(f)}[P_u]$ is independent of q, which implies that f should be independent of q. For $q = 1$, technically $S_q^{(f)}[P]$ is only defined as the limit

$$\lim_{q\to 1} S_q^{(f)}[P] = f\left(\lim_{q\to 1} U_q[P]\right), \tag{3.23}$$

where we have used the q-independence of f. If we require that, when $P = P_u$, this limit coincides with what Shannon entropy returns on P_u, i.e. $S_1[P_u] = \ln\Omega$, then we need a function f such that

$$\lim_{q\to 1} S_q^{(f)}[P_u] = f\left(\lim_{q\to 1} U_q[P_u]\right) = \ln\Omega, \tag{3.24}$$

i.e. $f(x) = \ln x$. Therefore, combining Eqs. (3.22) and (3.24) we obtain $f(x) = \ln x$ for all q, i.e. the only viable *UJK* entropy is Rényi entropy [15]

$$S_q[P] \equiv S_q^{(\ln)}[P] = \ln U_q[P] = \frac{1}{1-q}\ln\sum_{i=1}^{\Omega} p^q(G_i), \tag{3.25}$$

where, since the entropy above is the only 'surviving one' in the family $S_q^{(f)}[P]$, we have removed the superscript from the resulting $S_q^{(\ln)}[P]$. From Eq. (3.18) we can confirm that this entropy reduces to Shannon entropy in the limit $q \to 1$, a well-known result for Rényi entropy. This entropy is such that, on the uniform distribution P_u,

$$S_q[P_u] = \ln\Omega, \tag{3.26}$$

which does not depend on q, as demanded by our axiom. Therefore *the only viable UJK entropy is Rényi entropy.* In general, other *UJK* entropies do not respect our axiom.

An important counterexample is Tsallis entropy [16], defined as

$$S_q^{\text{Tsallis}}[P] \equiv S_q^{(\ln_q)}[P] = \frac{1}{1-q}\left(\sum_{i=1}^{\Omega} p^q(G_i) - 1\right) \tag{3.27}$$

and obtained from the so-called 'q-logarithm' $f(x) = \ln_q(x) \equiv (x^{1-q} - 1)/(1-q)$ (not to be confused with the ordinary logarithm of x to base q): indeed, when evaluated on P_u, this entropy takes the q-dependent value

$$S_q^{\text{Tsallis}}[P_u] = \frac{\Omega^{1-q} - 1}{1-q} = \ln_q(\Omega). \tag{3.28}$$

From the point of view of our axiom, such q-dependence is a contradiction: different values of q should not artificially attach different degrees of informativeness to an intrinsically uninformative distribution. Seen from another point of view, this contradiction arises from the q-dependence of the function f defining Tsallis entropy from the Uffink functional $U_q[P]$: such q-dependence is not admitted by our axiom because $f(U_q[P_u])$ should not depend on q. Note that the q-independence of the function f defining the *UJK* entropy $f(U_q[P_u])$ is a nontrivial consequence of our axiom, as it arises as necessary only when comparing entropies obtained for different values of q (if only a single value of q were considered, nothing would prevent f from being specified by that value of q). In particular, our axiom would demand $q = 1$ in order to have $S_q^{\text{Tsallis}}[P_u] = S_1[P_u]$, i.e. *the only viable Tsallis entropy is Shannon entropy.* We should stress at this point that the inadmissibility of Tsallis entropy under our axiom is not in contradiction with the many successful empirical applications of the so-called *q-exponential* or *Tsallis distribution* maximizing Tsallis entropy for fixed q [72, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103], because such distribution (that we explicitly consider later in this chapter) is exactly the same as the one maximizing Rényi entropy or any other monotonic function of the Uffink functional, as we also discuss below. However, when that distribution is 'put back' into the entropy, only Rényi entropy gives consistent results in terms of the absolute quantification of the uncertainty and the associated ML estimation and model selection procedures. Indeed, we will show that a ranking of models (or values of q) based on Tsallis entropy would 'mess up' the ranking based on ML, while the use of Rényi entropy restores and extends the consistency with the ML principle.

As another example, we apply the uninformativeness axiom to the *HT* family of composable (c, d)-entropies that can be written as in Eq. (3.19). If we require $S_{c,d}^{(h,g)}[P_u] = S_1[P_u] = \ln \Omega$ in analogy with Eq. (3.26), then the

axiom translates Eqs. (3.20) and (3.21) to:

$$\lim_{\Omega\to\infty} \frac{\ln \lambda\Omega}{\ln \Omega} = \lambda^{1-c} \tag{3.29}$$

$$\lim_{\Omega\to\infty} \frac{\ln \Omega^{1+a}}{\ln \Omega} = (1+a)^{d} \tag{3.30}$$

and implies $(c, d) = (1, 1)$. This parameter choice identifies the equivalence class of entropies that are additive for independent events. Both Shannon and Rényi entropies belong to this class. In particular, in the case $h(x) = x$ (trace-form entropy) and $g(x) = -x \ln x$ (Shannon entropy), one gets $(c, d) = (1, 1)$ [73], i.e. $S_{1,1}^{(x,-x\ln x)}[P] = S_1[P]$. Therefore *Shannon entropy is a viable trace-form HT entropy* under our axiom. Similarly, in the case $h(x) = \ln(x)/(1-q)$ and $g(x) = x^q$ (Rényi entropy) one again gets $(c, d) = (1, 1)$ [73], i.e. $S_{1,1}^{(\ln(x)/(1-q),x^q)}[P] = S_q[P]$. Therefore *Rényi entropy is a viable composable HT entropy.* By contrast, the case $h(x) = x$ and $g(x) = (x^q - \Omega^{-1})/(1-q)$ (Tsallis entropy) leads to $(c, d) = (q, 0)$ [73], confirming that Tsallis entropy (which is another trace-form entropy) does not respect our axiom.

The fact that, for both the *UJK* and *HT* families, only Rényi entropy (or an asymptotically equivalent one) 'survives' our axiom does not disagree with the possibility of non-extensivity of the entropy, which has led to the introduction of many variants of entropy over the last decades [72, 73]. Indeed, while our axiom selects entropy additivity for independent systems (as both Shannon and Rényi do), it does not have direct implications when independence is not present or even not known. In particular, it should be stressed that non-extensivity is a property not of the entropy itself, but of how the number Ω of configurations scales with the physical size of the system (i.e. the number n of units or particles) [73]. Even Shannon entropy can be non-additive if applied to a system where Ω (or Ω_{C^*}, when in presence of a constraint C^*) is not exponential in n, as clear from Eq. (3.2) or (3.3) (note that Eq. (3.3) applies in the microcanonical case, but a similar non-extensive scaling of the entropy would be exhibited in the canonical case as well). An important example in this respect is provided by random graphs: the number of all binary graphs on n vertices is $\Omega = 2^{\binom{n}{2}}$, so it is super-exponential [73, 104]. Even when subject to various types of constraints C^*,

the number Ω_{C^*} remains super-exponential [11]. At the opposite extreme, even for systems where Ω does increase exponentially in n, the system may still be subject to certain constraints such that Ω_{C^*} is sub-exponential in n, so that the resulting entropy is sub-extensive. An example is the class of State Space Reducing processes [86]. Later in the chapter, we will show that Shannon entropy can grow non-linearly in n even for a simple example of n independent observations. Therefore one first general result implied by the uninformativeness axiom is that non-extensivity or non-ergodicity (when present) should be completely encoded in the scaling of Ω_{C^*} with n, thus ultimately in the identification of the proper (effective) constraint C^*, and not in the expression of the entropy itself.

3.3.3 The generalized MEP

In a GMEP context, a direct consequence of the fact that our axiom restricts the viable expressions for the generalized entropies is, of course, a corresponding restriction on the probability distributions maximizing such generalized entropies under soft constraints (note that, under hard constraints, all maximum entropy distributions reduce to the microcanonical uniform distribution P_0 described in Sec. 3.2.2). This restriction can have two (related) effects: one on the functional form of the maximum entropy distribution and one on the way the distribution connects to the entropy itself and possibly other quantities. The *HT* and *UJK* entropies serve as good examples for both effects, as we now show.

For instance, while the general form for the probability distribution that maximizes the *HT* entropy $S_{c,d}^{(h,g)}[P]$ in trace form ($h(x) = x$) is the exponential of the so-called Lambert-W function[3] $\mathcal{W}(x)$ [86, 73], the only admissible form according to our axiom is the one corresponding to the choice $(c, d) = (1, 1)$. With this parameter choice, the $\mathcal{W}(x)$ function reduces to a linear function, so that the maximum entropy probability reduces to the Boltzmann-Gibbs distribution in Eq. (3.6) [73], consistently with the fact that the only admissible trace-form *HT* entropy according to our axiom is Shannon entropy, as we have shown above. To obtain a truly generalized maximum entropy probability, one should therefore consider

[3]The Lambert-W function $\mathcal{W}(x)$, which cannot be written in close form, is the solution to the equation $x = \mathcal{W}(x)e^{\mathcal{W}(x)}$. The real solutions are those that are relevant here.

non-trace-form entropies.

In particular, considering the Rényi entropy $S_q[P]$ which our axiom selects from both the *UJK* and the *HT* families, the GMEP can be formulated as the following well-known generalization of the MEP described in Sec. 3.2.2. Given an empirically observed value C^* of a (scalar or vector) function $C(G)$ of the unknown microstate G of a system, the least biased inference about G is provided by the distribution P_q that maximizes $S_q[P]$ under the (soft) constraint

$$\langle C \rangle_q \equiv \frac{\sum_{i=1}^{\Omega} p^q(G_i)\, C(G_i)}{\sum_{i=1}^{\Omega} p^q(G_i)} = C^*, \tag{3.31}$$

which generalizes the usual Shannonian constraint in Eq. (3.4) (note that $\langle C \rangle_1 = \langle C \rangle$). The quantity $\langle C \rangle_q$ is sometimes called *(normalized) q-mean* and can be regarded as a mean with respect to the so-called *escort* (or *zooming*) probability distribution $\tilde{p}(G_i) = p^q(G_i)/\sum_{j=1}^{\Omega} p^q(G_j)$ [105, 72]. This q-mean has been introduced to extend important properties and relations from the classical (i.e. Shannonian) statistical mechanics to the non-extensive one, including the Legendre structure of thermodynamics, the H-theorem and the Ehrenfest theorem [72]. However, from the point of view of statistical inference, whether $\langle C \rangle_q$ is a proper choice for a constraint is a debated issue in the literature, since this quantity might appear to lack a direct interpretation in relation to the available data. We will provide reassurance towards this concern: the use of $\langle C \rangle_q$ leads to a well-defined combination of the available data and, conveniently, regularizes the inference procedure in cases when $\langle C \rangle$ would be unstable. Indeed, $\langle C \rangle_q$ is always finite as soon as the distribution of C is normalizable, even when the ordinary mean $\langle C \rangle$ diverges and the ordinary inference process becomes inapplicable. Ensuring a finite value is crucial in order to estimate the Lagrange multiplier(s) from repeated observations and is especially important in our setting (described later in more detail) where we want to be able to determine q purely from data, without prior knowledge of its value and therefore without knowing beforehand whether the ordinary mean would diverge. A second concern may have arisen in the careful readers who noticed that the *UJK* entropies are usually derived from the *SJ* axioms while constraining the ordinary mean value $\langle C \rangle$, not $\langle C \rangle_q$. However,

later in this chapter we will show that the application of the ML principle to all parameters of the distribution (including q) leads exactly to the same numerical values of the GMEP probability distribution, irrespective of whether $\langle C\rangle$ or $\langle C\rangle_q$ is used (provided both quantities are finite). Finally and profoundly, the use of the q-mean restores a complete consistency between the ML principle and the Lagrangian optimization and yields a direct relationship between maximized likelihood and entropy, while the use of the ordinary mean would fail to do so.

To carry out the constrained maximization of $S_q[P]$, we look for the vanishing derivatives of the q-Lagrangian

$$\mathcal{L}_q[P] \equiv S_q[P] - \alpha\left[\sum_{i=1}^{\Omega} p(G_i) - 1\right] - \theta\cdot[\langle C\rangle_q - C^*] \tag{3.32}$$

with respect to P, α and θ, and assume $q \neq 1$ from now on. The resulting values are denoted as P_q, α_q, θ_q. In particular, setting $\partial\mathcal{L}_q[P]/\partial P|_{P_q} = 0$ we get

$$\begin{aligned} 0 &= \left.\frac{\partial\mathcal{L}_q[P]}{\partial p(G_i)}\right|_{p_q(G_i)} \\ &= \frac{q}{1-q}\frac{p_q^{q-1}(G_i)}{\sum_j p_q^q(G_j)} - \alpha - q\,p_q^{q-1}(G_i)\frac{\theta\cdot(C(G_i) - \langle C\rangle_q)}{\sum_j p_q^q(G_j)} \end{aligned} \tag{3.33}$$

for all i from 1 to Ω, from which it is clear that $p_q(G_i)$ depends on θ, as in the case $q = 1$, and additionally on q. The derivative of $\mathcal{L}_q[P]$ with respect to α leads to a condition identical to Eq. (3.7):

$$\left.\frac{\partial\mathcal{L}_q[P_q]}{\partial\alpha}\right|_{\alpha_q} = 0 \quad\Rightarrow\quad \sum_{i=1}^{\Omega} p_q(G_i,\theta) = 1, \tag{3.34}$$

which can be used to determine α_q by multiplying both sides of Eq. (3.33) and then summing over i. We then get

$$\alpha_q = \frac{q}{1-q} \qquad (q \neq 1), \tag{3.35}$$

which is the counterpart of Eq. (3.8). Substituting α_q in (3.33) and singling

out $p_q(G_i)$ yields

$$p_q(G_i,\theta) = \frac{[1-(1-q)\,\theta\cdot(C(G_i)-\langle C\rangle_q)]_+^{1/(1-q)}}{\left[\sum_{j=1}^{\Omega} p_q^q(G_j,\theta)\right]^{1/(1-q)}} \tag{3.36}$$

where we have used the notation $[x]_+^a \equiv 0$ if $x<0$, while $[x]_+^a \equiv x^a$ otherwise [72]. Note that the denominator of Eq. (3.36) equals the Uffink functional $U_q[P_q(\theta)]$ and must also equal the *generalized partition function*

$$W_q(\theta) \equiv \sum_{i=1}^{\Omega}[1-(1-q)\,\theta\cdot(C(G_i)-\langle C\rangle_q)]_+^{1/(1-q)} \tag{3.37}$$

since $p_q(G_i,\theta)$ is already normalized via the condition in Eq. (3.35). In other words,

$$W_q(\theta) = \left[\sum_{i=1}^{\Omega} p_q^q(G_i,\theta)\right]^{1/(1-q)} = U_q[P_q(\theta)]. \tag{3.38}$$

Finally, the maximum entropy probability equals

$$p_q(G_i,\theta) = \frac{[1-(1-q)\,\theta\cdot(C(G_i)-\langle C\rangle_q)]_+^{1/(1-q)}}{W_q(\theta)} \tag{3.39}$$

which has the form of a so-called q-exponential or Tsallis [72] distribution. Note that Eqs. (3.32) and (3.39) generalize Eqs. (3.5) and (3.6), respectively. Moreover note that, if we formally introduce a pseudostate $\tilde{G}$ such that $C(\tilde{G}) = \langle C\rangle_q$, it follows from Eq. (3.39) that $p_q(\tilde{G},\theta) = 1/W_q(\theta) = 1/U_q[P_q(\theta)]$. Then, from Eq. (3.38), one can see that:

$$p_q^{q-1}(\tilde{G},\theta) = \sum_{i=1}^{\Omega} p_q^q(G_i,\theta) = U_q^{1-q}[P(\theta)]. \tag{3.40}$$

We will discuss the relationship between $C(\tilde{G})$ and $C(G^*)$ later.

When $q \to 1$, $p_q(G_i,\theta) \to Z_1^{-1}(\theta)\exp(-\theta\cdot C(G_i))$, retrieving the Boltzmann-Gibbs distribution in Eq. (3.6). When $q \neq 1$, the q-exponential has nothing to do with the ordinary exponential and actually has power-law tails proportional to $C(G_i)^{1/(1-q)}$ for large values of $C(G_i)$. The presence of these heavy tails, which are widespread in several real-world complex

systems [72, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103], is one of the reasons why q-exponentials have attracted interest, their derivation from the maximization of a suitable entropy appearing convenient and parsimonious [73, 72]. In the literature, there is some confusion around the fact that q-exponentials derive from the maximization of Tsallis entropy given by Eq. (3.27). While this is certainly true, it is also true that they derive from *any* of the *UJK* entropies in Eq. (3.17): the distribution maximizing $U_q[P]$ necessarily maximizes $f(U_q[P])$ as well, for any monotonic f (indeed, our derivation above started from Rényi entropy). Therefore, the robust empirical support for the q-exponential distribution that has been highlighted in several analyses of empirical data in e.g. quantum chemistry [90], high energy physics [91, 92, 93, 94, 95], cosmology [96], finance [97], acoustics [100, 101], seismology [98], biology [99] and medicine [102, 103] *cannot be used as support for a specific member f of the entropy family $f(U_q[P])$*. Indeed, there is no direct *empirical* evidence which selects a specific entropy in the family, and the main arguments adopted in the literature towards the use of e.g. Rényi versus Tsallis entropy remain of theoretical or mathematical nature [72, 106], such as invoking consistency with some formal framework. In this respect, *our approach here can be regarded as an additional theoretical consistency argument* to select entropies within families (e.g. Shannon entropy within the Tsallis family) or restricted entropy families within 'super-families' (e.g. Rényi entropy within the *UJK* and *HT* families). Specifically, the differences among the members of the *UJK* entropy family arise when the maximum entropy q-exponential is put back into the entropy itself. When this happens, the uninformativeness axiom has the important role of selecting Rényi entropy as the member of the family that solves all the inconsistencies discussed in Sec. 3.2.6, as we show later in the chapter.

What remains to be done is the determination of the parameter θ. It is useful at this point to introduce the reparameterization

$$\psi(\theta) \equiv \frac{\theta}{1 + (1-q)\,\theta \cdot \langle C \rangle_q}, \tag{3.41}$$

through which it is possible to (formally) remove $\langle C \rangle_q$ from the expression

for $p_q(G_i, \theta)$ and get

$$p_q(G_i, \psi) = \frac{[1 - (1-q)\,\psi \cdot C(G_i)]_+^{1/(1-q)}}{Z_q(\psi)} \tag{3.42}$$

where, denoting the inverse of $\psi(\theta)$ as $\theta(\psi)$,

$$\begin{aligned} Z_q(\psi) &\equiv \sum_{i=1}^{\Omega} [1 - (1-q)\,\psi \cdot C(G_i)]_+^{1/(1-q)} && (3.43) \\ &= \frac{W_q\left(\theta(\psi)\right)}{[1 + (1-q)\,\theta(\psi) \cdot \langle C \rangle_q]^{1/(1-q)}} && (3.44) \end{aligned}$$

is the reparametrized partition function. Note that $Z_q(\psi) \neq W_q\left(\theta(\psi)\right)$ unless $q \to 1$, in which case $\psi \to \theta$ and $W_1(\theta) \to Z_1(\theta)$. The optimal value ψ_q is determined by the condition

$$\left.\frac{\partial \mathcal{L}[P_q]}{\partial \psi}\right|_{\psi_q} = 0 \quad \Rightarrow \quad \frac{\sum_{i=1}^{\Omega} p_q^q(G_i, \psi_q)\, C(G_i)}{\sum_{i=1}^{\Omega} p_q^q(G_i, \psi_q)} = C^* \tag{3.45}$$

corresponding to the intended requirement in Eq. (3.31) and generalizing Eq. (3.9) to the case $q \neq 1$. One the value ψ_q is determined via the condition above, it can be inserted into Eq. (3.42) to obtain the final maximum entropy probability distribution $P_q(\psi_q)$.

3.3.4 Link with the ML principle and model selection

We now show that the entropy selected by the uninformativeness axiom restores consistency with the ML principle and retains an interpretation for model selection, exactly as in the Shannon case. Both properties are not guaranteed for other entropies. At the same time, we show how to account for multiple independent observations about the same system.

In analogy with Sec. 3.2.3, we start with the case $M = 1$ and define the ML estimation procedure for the parameter ψ_q as follows:

$$\psi_q^* = \underset{\psi}{\operatorname{argmax}}\, \ell_q(\psi), \quad \ell_q(\psi) \equiv \ln p_q(G^*, \psi). \tag{3.46}$$

Requiring $\partial \ell_q(\psi)/\partial\psi|_{\psi_q^*} = 0$, one gets

$$\sum_{i=1}^{\Omega} C(G_i)\, p_q^q(G_i, \psi_q^*) = C(G^*)\, p_q^{q-1}(G^*, \psi_q^*) \tag{3.47}$$

and, dividing both terms by $\sum_G p_q^q(G^*, \psi_q^*)$,

$$\langle C \rangle_q = \frac{C(G^*)\, p_q^{q-1}(G^*, \psi_q^*)}{\sum_G p_q^q(G^*, \psi_q^*)}. \tag{3.48}$$

One might think that the right hand side of the above equation is different from the 'desired' value $C^* = C(G^*)$, however this is not the case. Indeed, considering again a pseudostate $\tilde{G}$ such that $C(\tilde{G}) = \langle C \rangle_q$ and using Eq. (3.40), we can rewrite Eq. (3.48) as

$$\frac{C(\tilde{G})}{C(G^*)} = \frac{1 - (1-q)\, \psi_q^* \cdot C(\tilde{G})}{1 - (1-q)\, \psi_q^* \cdot C(G^*)}, \tag{3.49}$$

which leads to $C(\tilde{G}) = C(G^*)$. In other words, the value ψ_q^* defined by Eq. (3.46) coincides with the value ψ_q defined by Eq. (3.45), i.e. $\psi_q^* \equiv \psi_q$, i.e.

$$\left.\frac{\partial \ell_q(\psi)}{\partial \psi}\right|_{\psi_q^*} = 0 \quad \Rightarrow \quad \frac{\sum_{i=1}^{\Omega} p_q^q(G_i, \psi_q^*)\, C(G_i)}{\sum_{i=1}^{\Omega} p_q^q(G_i, \psi_q^*)} = C^* \tag{3.50}$$

in analogy with Eq. (3.45). This means that the ML principle can still be seen as equivalent to the part of the Lagrangian optimization relative to ψ. Moreover, the application of the logarithm to both sides of Eq. (3.40) leads to

$$\ell_q(\psi_q^*) = -S_q[P_q(\psi_q^*)], \tag{3.51}$$

showing that, for $M = 1$, the log-likelihood of the observation coincides with minus the Rényi entropy. This extends Eq. (3.12) to the case $q \neq 1$, with the following consequence. If we put ourselves in a model selection framework where we interpret each $P_q(\psi)$ for different values of q as a different model for the same data C^* (for $M = 1$) and where each model has a maximized likelihood equal to $P_q(\psi_q^*)$, then the optimal model (if unique) corresponds to the one with the highest value of $P_q(\psi_q^*)$. Thanks to Eq. (3.51), we can equivalently interpret this ranking of different models according to their maximized likelihood as a ranking based on their realized Rényi entropy: the

model with highest maximized likelihood is the one with minimum Rényi entropy. Notably, other entropies of the *UJK* family, including Tsallis entropy, do not manifest this property. Also the relationship in Eq. (3.13) generalizes as follows:

$$\mathcal{L}_q[P_q(\psi_q^*)] = S_q[P_q(\psi_q^*)] = -\ell_q(\psi_q^*), \tag{3.52}$$

relating the value of the Lagrangian attained by $P_q(\psi_q^*)$ to the maximized log-likelihood. Therefore, up to this point, it seems that Rényi entropy retains all the desirable properties of Shannon entropy.

We now consider the case of $M > 1$ i.i.d. realizations $\{G_m^*\}_{m=1}^M$ of the system, leading to M independent observations $\{C_m^*\}_{m=1}^M$ of the constraint, where $C_m^* \equiv C(G_m^*)$ for all m. We have already seen in Sec. 3.2.3 that in this case it is the ML principle, not the MEP, that identifies how to combine the M observed values. Introducing again the average log-likelihood $\overline{\ell}_q(\psi)$, the ML condition for ψ becomes a straightforward generalization of Eq. (3.14):

$$\psi_q^* = \underset{\psi}{\operatorname{argmax}}\, \overline{\ell}_q(\psi), \quad \overline{\ell}_q(\psi) \equiv \frac{\sum_{m=1}^M \ln p_q(G_m^*, \psi)}{M}. \tag{3.53}$$

It is not difficult to show that requiring $\partial\overline{\ell}_q(\psi)/\partial\psi|_{\psi_q^*} = 0$ translates into:

$$\sum_{i=1}^{\Omega} C(G_i)\, p_q^q(G_i, \psi_q^*) = \frac{1}{M} \sum_{m=1}^{M} C(G_m^*)\, p_q^{q-1}(G_m^*, \psi_q^*) \tag{3.54}$$

or equivalently

$$\langle C \rangle_q = \frac{\sum_{m=1}^M C(G_m^*)\, p_q^{q-1}(G_m^*, \psi_q^*)}{M \sum_{i=1}^{\Omega} p_q^q(G_i, \psi_q^*)} \tag{3.55}$$

which extends the classical ($q = 1$) result in Eq. (3.15) to the general, non-Shannon case. We therefore learn that the arithmetic average is no longer the optimal way of combining the M available observations in order to determine the parameter ψ. Indeed, dismissing the arithmetic average makes sense if we recall that, *a priori*, we do not even know whether the first moment of the distribution generating the M values $\{C_m^*\}_{m=1}^M$ is finite. Indeed, the q-exponential distributions that are solution to the GMEP exhibit a power-law behavior for $q \neq 1$. As a consequence, in principle

all their moments could diverge, depending on the value of q. Assuming that q is not known beforehand and is rather determined by the inference procedure itself (as we assume later on), it would make no sense at all to use the arithmetic average to constrain the q-mean in case of multiple observations, since that average might become infinite in the $M \to \infty$ limit when $q > 3/2$, while the q-mean is by construction finite whenever the distribution is normalizable. The same problem might in principle apply to any higher moment $\langle C^n \rangle$ with $n > 1$, while any q-generalized moment $\langle C^n \rangle_q$ evaluated with respect to Eq. (3.42) converges if q is such that the distribution is normalizable (which is a basic requirement for this procedure to be consistent [72]). The ML estimator determined by Eq. (3.55) identifies the distribution's parameters, irrespective of the converge of any moment.

An important consequence of the fact that $\langle C \rangle_q$ is no longer equal to the arithmetic mean of the M observations is that in general, for $q \neq 1$ and $M > 1$,

$$S_q[P_q(\psi_q^*)] \neq -\overline{\ell}_q(\psi_q^*), \tag{3.56}$$

thus failing to generalize Eq. (3.16) to the case $q \neq 1$ and Eq. (3.51) to the case $M > 1$. Similarly, Eqs. (3.13) and (3.52) do not generalize here. Rather, a relationship that is still valid is

$$S_q[P_q(\psi_q^*)] = -\tilde{\ell}_q(\psi_q^*), \tag{3.57}$$

where $\tilde{\ell}_q(\psi) \equiv \ln p_q(\tilde{G}, \psi)$ is a sort of 'pseudolikelihood' involving the pseudostate $\tilde{G}$ such that $C(\tilde{G}) = \langle C \rangle_q$ introduced above. Unfortunately, $\tilde{\ell}_q(\psi)$ is no longer equal to the actual log-likelihood $\overline{\ell}_q(\psi)$ based on the M observations. Does this mean that, in presence of multiple i.i.d. observations of the same quantity about a system, the correspondence between log-likelihood and entropy is lost? The answer to this question emerges when looking at a seemingly unrelated problem, i.e. the selection of the optimal value of the entropic parameter q, and is provided in Sec. 3.3.5.

3.3.5 Inference of the entropic parameter

We now come to the last, and in many ways most crucial, benefit implied by the uninformativeness axiom, namely the possibility of consistently iden-

tifying the entropic parameter(s) purely from the data, without postulating *a priori* knowledge about the system — such as scaling laws of the type exemplified by Eqs. (3.20) and (3.21) [86, 73, 87].

To this end, starting directly with the general case $M \geq 1$, we invoke again the ML principle and, building on its restored consistency with the estimation of the other parameters of the maximum entropy distribution proven in Eq. (3.50), extend it to the identification of the entropic parameter(s) themselves. This means that we now turn the model selection procedure we discussed after deriving Eq. (3.51) (where we compared different models $\{P_q(\psi)\}_q$ in terms of their maximized likelihoods $\{P_q(\psi_q^*)\}_q$) into a single ML parameter estimation procedure, applied directly to the two-parameter distribution $P_q(\psi)$. Indeed, the ML principle treats any parameter agnostically, without specific interpretations, and is therefore 'unaware' of the fact that q and the other structural parameters play different roles in an information-theoretic setting. Considering again Reńyi entropy as the only viable entropy from the *UJK* and *HT* families, the ML principle applied to the entropic parameter q is formally stated as follows:

$$q^* = \underset{q}{\operatorname{argmax}}\, \overline{\ell}_q(\psi), \quad \overline{\ell}_q(\psi) \equiv \frac{\sum_{m=1}^{M} \ln p_q(G_m^*, \psi)}{M}. \tag{3.58}$$

On the other hand, combining the above expression with Eq. (3.53), it is clear that the estimation of q is coupled to that of ψ, so that the actual formulation of the extended ML principle is

$$(\psi_{q^*}^*, q^*) = \underset{(\psi,q)}{\operatorname{argmax}}\, \overline{\ell}_q(\psi), \quad \overline{\ell}_q(\psi) \equiv \frac{\sum_{m=1}^{M} \ln p_q(G_m^*, \psi)}{M}. \tag{3.59}$$

This expression immediately tells us that, once the ML principle is extended to the determination of q, the results we have discussed in Sec. 3.3.4 represent only one side of the coin. Now, requiring jointly

$$\left.\frac{\partial \overline{\ell}_q(\psi)}{\partial \psi}\right|_{(\psi_{q^*}^*, q^*)} = 0, \quad \left.\frac{\partial \overline{\ell}_q(\psi)}{\partial q}\right|_{(\psi_{q^*}^*, q^*)} = 0, \tag{3.60}$$

we arrive again at Eq. (3.55) (with q replaced by q^*) plus the additional

condition

$$\sum_{i=1}^{\Omega} p_{q^*}(G_i, \psi^*_{q^*}) \ln\left[1 - (1-q^*)\,\psi^*_{q^*} \cdot C(G_i)\right] \tag{3.61}$$
$$= \frac{1}{M}\sum_{m=1}^{M} p_{q^*}(G^*_m, \psi^*_{q^*}) \ln\left[1 - (1-q^*)\,\psi^*_{q^*} \cdot C(G_m)\right].$$

Recalling from Eq. (3.42) that

$$1 - (1-q^*)\,\psi^*_{q^*} \cdot C(G_i) = [p_{q^*}(G_i, \psi^*_{q^*}) Z_{q^*}(\psi^*_{q^*})]^{1-q^*} \tag{3.62}$$

we obtain the condition

$$\sum_{i=1}^{\Omega} p_{q^*}(G_i, \psi^*_{q^*}) \ln p_{q^*}(G_i, \psi^*_{q^*}) = \frac{\sum_{m=1}^{M} \ln p_{q^*}(G^*_m, \psi^*_{q^*})}{M}. \tag{3.63}$$

In other words, the additional ML condition determining q^* requires that *the maximized log-likelihood equals minus Shannon entropy*, i.e.

$$S_1[P_{q^*}(\psi^*_{q^*})] = -\overline{\ell}_{q^*}(\psi^*_{q^*}), \tag{3.64}$$

restoring an analogy with Eq. (3.16) that appeared to be lost and replaced by Eq. (3.57) when considering $q \neq 1$. Remarkably, we now realize that, when the ML principle is extended to q, the correspondence with Eq. (3.16) is not replaced, but rather *accompanied* by Eq. (3.57). On one hand, Eq. (3.64) generalizes to the class of q-exponentials the relationship in Eq. (3.16) that is well-known for the exponential distribution. On the other hand, it does not hold for any value of q and independently on the data, but only for the pair of values $(\psi^*_{q^*}, q^*)$ that maximize the likelihood. In particular, the connection between Shannon entropy and log-likelihood at the specific parameter value $(\psi^*_{q^*}, q^*)$ remains a general result, even for $q \neq 1$ and $M > 1$. This might look quite surprising, because, for $q \neq 1$, the log-likelihood is based on the q-exponential distribution that maximizes Rényi, not Shannon, entropy.

Despite the surprise, the above result makes perfect sense because we have assumed M independent observations. Actually, it solves the final in-

consistency we pointed out in Sec. 3.2.6: assuming independent observations justifies Shore and Johnson's original restricted interpretation of axiom *SJ3* and leads to Shannon entropy as the quantifier of the uncertainty of the data. Indeed the inequality in Eq. (3.56) should be put in relation with our initial discussion of the axiom *SJ3* about system independence. Recall that assuming that the M values $\{C_m^*\}_{m=1}^M$ come from independent observations is equivalent to assuming that there are M identical and independent copies of the same system, each copy being observed exactly once. Under this assumption of independence, the original reasoning by Shore and Johnson becomes appropriate and one should therefore expect that Shannon entropy, rather than Rényi entropy, is the proper entropy describing the combined system of M copies. Therefore the breakdown of the correspondence between the average log-likelihood and Rényi entropy can be regarded as a symptom of the assumed independence of the M observations. When $M = 1$, we can use Eq. (3.51) and combine it with Eq. (3.64) to obtain

$$S_{q^*}[P_{q^*}(\psi_{q^*}^*)] = S_1[P_{q^*}(\psi_{q^*}^*)] \tag{3.65}$$

showing that in this particular case the maximum entropy probability distribution returns coinciding values of Shannon and Rényi entropy, even if it maximizes the latter but not the former. This result does not hold in general for $M > 1$. An important consequence of the combination of Eqs. (3.16) and (3.64), valid also for $M > 1$, is that the Shannon entropy of the distribution that maximizes Rényi entropy is not larger than that of the exponential distribution, if the parameters are set according to ML:

$$S_1[P_{q^*}(\psi_{q^*}^*)] \leq S_1[P_1(\psi_1^*)], \tag{3.66}$$

meaning that the optimized q^*-exponential achieves a 'better compression' of the data than the ordinary exponential. As we show below for a simple example, $S_1[P_{q^*}(\psi_{q^*}^*)]$ and $S_1[P_1(\psi_1^*)]$ might even scale differently with the number of observations (or size of the system), making the inequality (3.66) particularly relevant for data compression purposes.

The remarkable result in Eq. (3.64) has an important consequence for the estimation of q^*. In particular, in order to determine both q^* and $\psi_{q^*}^*$,

one can consider a range of values for q and, for each value in the range, compute ψ_q^* according to Eq. (3.55). This produces, for each value of q, a log-likelihood $\overline{\ell}_q(\psi_q^*)$ that is only partially maximized, in the sense that the maximization has been carried out only with respect to ψ_q and not yet with respect to q. Then, among all these partially maximized log-likelihoods, one can select the one with the highest value.

This will identify the value q^* and the associated value $\psi_{q^*}^*$, which ultimately correspond to the completely maximized log-likelihood $\overline{\ell}_{q^*}(\psi_{q^*}^*)$. Only for this parameter choice $(q^*, \psi_{q^*}^*)$, the log-likelihood equals minus Shannon entropy. So from the ML condition Shannon entropy emerges spontaneously: while the probability P_q maximizes Rényi entropy and not Shannon entropy, the latter is the correct entropy for model selection to take independence into account. We stress once again that the physical entropy characterizing the real system may be different from both Shannon and Rényi entropies. Nonetheless, the introduction of our axiom is consistent with an information-theoretic model selection criterion, based on the maximization of Rényi entropy to obtain the functional form of the probability distribution and the ML principle to estimate its parameters, including the entropic one. In order to illustrate the performance of the above approach, we now consider two simple numerical examples.

Our first example is a system described by an observable $C(G)$ taking only positive real values, i.e. $C(G) \in [0, +\infty)$. Moreover, we assume that $\Omega_C = 1$ for all C, meaning that for each value $C(G)$ of the observable there is only one state G that realizes it. Thus, the sums over system states simplify into integrals over the observable values: $\sum_{i=1}^{\Omega} \to \int_0^\infty dC$. The probability distribution resulting from the GMEP is then:

$$p_q(G_i, \psi) = (2-q)\,\psi\,[1-(1-q)\,\psi \cdot C(G_i)]_+^{\frac{1}{1-q}}\,, \tag{3.67}$$

where we have used $Z_q(\psi) = 1/(2-q)\psi$. For different values of ψ and q, we have drawn an i.i.d. sample of $M = 10^3$ realizations from the distribution above, with the aim of inferring the true value of those parameters purely from the data so generated. In particular, we have generated samples from an exponential distribution (i.e. $q_{\text{true}} = 1$), a q-exponential distribution with finite first moment $\langle C \rangle$ ($q_{\text{true}} = 1.3$) and a q-exponential distribution

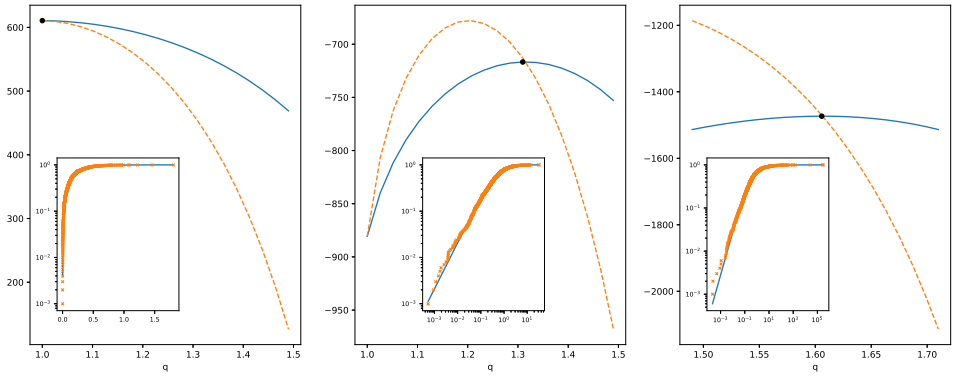


Figure 3.1: Comparison between the average partially maximized log-likelihood $\overline{\ell}_q(\psi_q^*)$ (solid line) and minus Shannon entropy $-S_1[P_q(\psi_q^*)]$ (dashed line) as a function of q, for three samples of $M = 10^3$ deviates generated from the probability distribution $P_q(\psi)$ in Eq. (3.67), and in particular: exponential distribution where $q_{\text{true}} = 1$ and $\psi_{\text{true}} = 5.0$ (left), q-exponential (power-law) distribution with finite first moment where $q_{\text{true}} = 1.3$ and $\psi_{\text{true}} = 3.0$ (center), and q-exponential (power-law) distribution with diverging first moment where $q_{\text{true}} = 1.6$ and $\psi_{\text{true}} = 7.0$ (right). The insets show the comparison between the empirical cumulative distributions of the M realized values (crosses) and the retrieved maximum entropy distribution using the inferred values $(q^*, \psi_{q^*}^*)$ (solid line).

with diverging first moment ($q_{\text{true}} = 1.6$). Figure 3.1 shows, for the three cases, $\overline{\ell}_q(\psi_q^*)$ (blue line) and $-S_1[P_q(\psi_q^*)]$ (orange line) as functions of q. The black dot indicates the intersection between the two curves, which identifies the estimated value q^* where Eq. (3.64) is realized. The true values of the parameters and their inferred ML estimates $(q^*, \psi_{q^*}^*)$ are presented in Table 3.1. Since the left plot corresponds to $q_{\text{true}} = 1$, it is a standard exponential distribution. In such a case, the two curves intersect only for $q = 1$. By contrast, the other two cases correspond to $q_{\text{true}} \neq 1$ and the two curves intersect in two points, namely $q = 1$ and $q = q_{\text{true}}$. In these cases, both intersections are solutions of Eq. (3.64), but the solution $q \neq 1$ is the one that corresponds to higher log-likelihood (and lower entropy). This example is very simple but explanatory: it shows directly how Shannon entropy plays a role in model selection even when the distribution taken into consideration comes from the GMEP and maximizes Rényi, not Shannon. We also stress once more that, in the last case, constraining the usual mean rather than the q-mean would have not been appropriate, since for $q > 1.5$ the usual mean diverges as $M \to \infty$; instead, by using the q-average, it becomes possible to consistently characterize the original infinite-mean

power-law distribution. Note that, in the real world, the physical entropy characterizing the system producing the data simulated here might be unspecified and, as such, could be different from the information-theoretic one we are using for the inference procedure. Indeed, as stated above, also Tsallis entropy and the HT entropies are naturally associated with power-law distributions. Moreover we would also like to notice that, since the maximized log-likelihood coincides with minus the Shannon entropy, the selected probability distribution could be more compressible than the one obtained setting $q = 1$. An example is shown in the middle panel of Figure 3.1: if the probability in Eq. (3.67) represented the probability distribution of a source generating i.i.d. symbols, then the one selected by our GMEP and ML would be, unsurprisingly, more compressible then the one obtained by setting $q = 1$. Moreover, the precise value $\overline{\ell}_{q^*}(\psi^*_{q^*})$, coincides with the lower bound of compression, in accordance with Eq. (3.66). We also note that, if the true process generating the data has finite mean, then the estimation of ψ^*_1 is a well-defined problem, otherwise it is not. Figure 3.2 shows that the Shannon entropy of both distributions is linear in M, with the one referred to the exponential distribution being larger. Moreover, since in this case the data are generated according to a finite-mean distribution, the fluctuations in both $\psi^*_{q^*}$ and ψ^*_1 are small. On the other hand, Figure 3.3 shows a situation in which the generating process has a diverging mean. While the estimated $\psi^*_{q^*}$ is robust, ψ^*_1 has huge fluctuations and results in a super-linear growth of the Shannon entropy. Since $\psi^*_1 = M/\sum_{j=1}^{M} C^*_j$, the large fluctuations arise from the fluctuating empirical average of M observations coming from an infinite mean process. Indeed in this particular case, where the observations come from the distribution in Eq. (3.67) with $q = 1.6$, i.e. $p(C) \sim C^{-\alpha-1}$ with $\alpha = 2/3$, we get that, for large M, the arithmetic average diverges as $M^{-1}\sum_{j=1}^{M} C^*_j \sim M^{1/\alpha-1} = M^{1/2}$. The use of the q-average fixes this problem, keeping the estimation of its associated Lagrange multiplier stable. Such divergence is evident even if one plugs the estimated probability densities back into the Shannon entropy. In fact, in general, $S_1[P_q(\psi)] = \frac{1}{2-q} - \log\big((2-q)\psi\big)$. This leads to a super-linear growth of the entropy if the standard approach is applied: $S_1[P_1(\psi^*_1)] = 1 - \log\psi^*_1 \sim \log M^{1/2}$. Instead, by applying the generalized GMEP, the entropy is $S_1[P_{q^*}(\psi^*_{q^*})] = \frac{1}{2-q^*} - \log\big((2-q^*)\psi^*_{q^*}\big) \sim$ const, since both q^* and $\psi^*_{q^*}$ do not change with M. Incidentally, this example

shows that Shannon entropy, even if additive for independent events, can grow nonlinearly in the number of independent observations because of an 'anomalous' scaling of the relevant Lagrange multiplier(s).

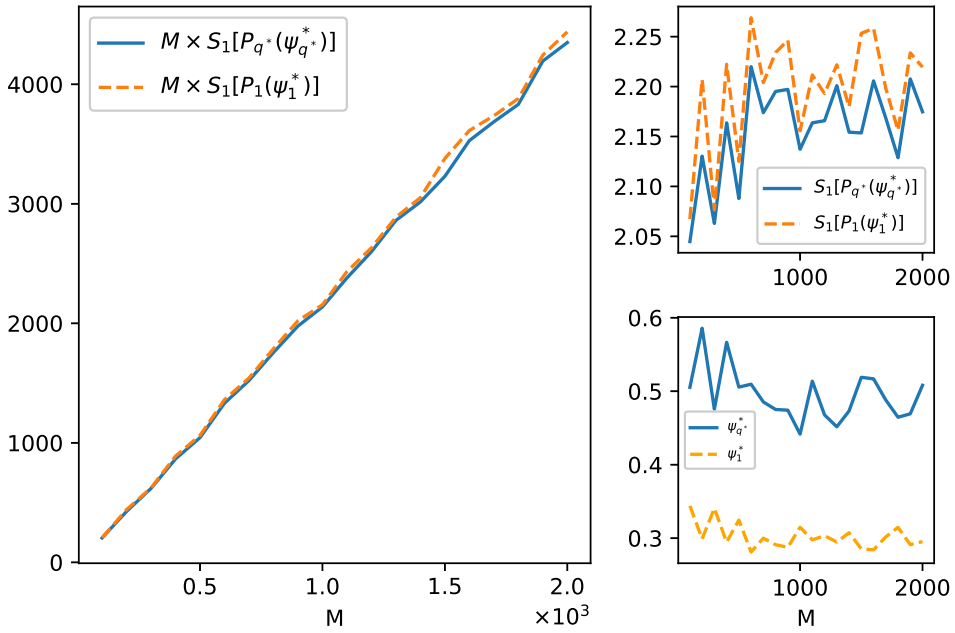


Figure 3.2: . Data generated according to Eq. (3.67) with $q_{\text{true}} = 1.2$ and $\psi_{\text{true}} = 0.5$. Left panel: Shannon entropy. Top right panel: Shannon entropy *per observation*. Bottom right panel: Estimated Lagrange multipliers.

Table 3.1: Comparison of true parameters' values with ML estimates.

q_{true}	ψ_{true}	q^*	$\psi^*_{q^*}$
1.0	5.0	1.0	5.0
1.3	3.0	1.3	2.9
1.6	7.0	1.6	7.3

Our second and last example is the simple case of a system characterized by a Bernoulli random variable $C(G)$ taking value $C(G) = 1$ with true underlying probability p_{true}, and value $C(G) = 0$ with probability $1 - p_{\text{true}}$. Constraining the q-average yields

$$p_q(G_i, \psi) = \frac{[1 - (1-q)\,\psi \cdot C(G_i)]_+^{1/(1-q)}}{1 + [1 - (1-q)\,\psi]^{1/(1-q)}}. \tag{3.68}$$

Let us now call $p_q(\psi)$ the probability $p_q(G, \psi)$ when $C(G) = 1$ and $1 - p_q(\psi)$

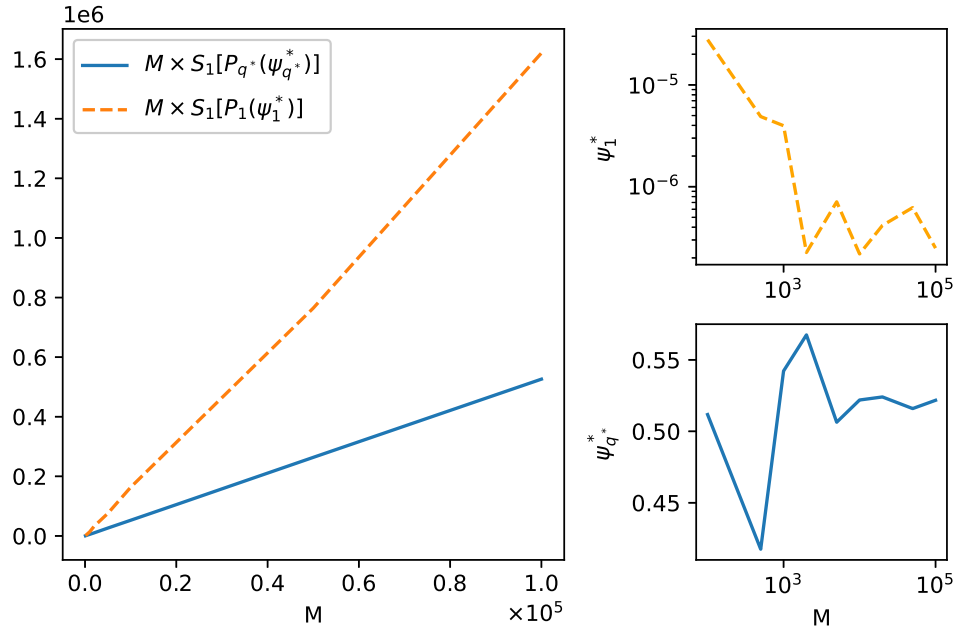


Figure 3.3: . Data generated according to Eq. (3.67) with $q_{\text{true}} = 1.7$ and $\psi_{\text{true}} = 0.5$. Left panel: Shannon entropy. Top right panel: Estimated ψ^*_1. Bottom right panel: Estimated $\psi^*_{q^*}$.

the probability $p_q(G, \psi)$ when $C(G) = 0$. It is easily verified that

$$\langle C \rangle = p_q(\psi) \tag{3.69}$$

and

$$\langle C \rangle_q = \frac{p_q^q(\psi)}{p_q^q(\psi) + [1 - p_q(\psi)]^q}. \tag{3.70}$$

If we now consider M i.i.d. realizations $\{C^*_m\}_{m=1}^M$ of C and apply Eq. (3.54), we get $p_{q^*}^{q^*}(\psi^*_{q^*}) = p_{q^*}^{q^*-1}(\psi^*_{q^*}) f_1$ where $f_1 = \sum_{m=1}^M C^*_m/M$ is the empirical frequency of the observed instances where $C^*_m = 1$. This relation trivially reduces to

$$f_1 = p_{q^*}(\psi^*_{q^*}). \tag{3.71}$$

Since there are infinite couples of $(\psi^*_{q^*}, q^*)$ that satisfy the ML condition and produce exactly the same maximized log-likelihood, none of them has to be preferred over the other. According to our approach, one finds a result which recalls the Shannonian case: for a Bernoulli random variable, the parameters of the maximum entropy distribution have to be set so that the

estimated probability matches the empirical frequency. This can be done for any value of q and is therefore a degenerate case where no specific value of q can be learned from the data, because the resulting maximum entropy distributions are all identical to each other. This is not unexpected: in fact, what we have done here in practice is trying to capture the properties of a one-parameter binary random variable with a distribution that depends on two parameters. Therefore this example illustrates a situation where the data are not informative enough to infer the entropic parameter.

3.3.6 Relation with ordinary average constraints

The *SJ* axioms are explicitly formulated for constraints that are *linear* in probability: they not only find a particular functional form for the entropy, but define the whole maximum entropy procedure including the estimation of the Lagrange multipliers. Therefore, one may question whether it makes sense to apply our GMEP procedure, which uses q-means, to the *UJK* family, which derives from the *SJ* axioms.

In [14], Jizba and Korbel evaluated explicitly the functional form of the probability distribution resulting from the maximization of $U_{q'}[P]$ with linear constraints. Here, we report their result using our notation. If one constrains the ordinary mean $\langle C \rangle$ and follows the same maximization procedure for $S_{q'}[P]$ as described in Sec. 3.3.3, a different maximum entropy distribution $\hat{P}_q(\theta)$ is obtained:

$$\hat{p}_{q'}(G_i) = \frac{[1 - (q'-1)\,\hat{\theta} \cdot (C(G_i) - \langle C \rangle)]_+^{\frac{1}{q'-1}}}{\hat{W}_{q'}(\hat{\theta})}. \tag{3.72}$$

Following the reparameterization previously introduced, it is also possible to write:

$$\hat{p}_{q'}(G_i, \hat{\psi}) = \frac{\left[1 - (q'-1)\,\hat{\psi} \cdot C(G_i)\right]_+^{1/(q'-1)}}{\hat{Z}_{q'}(\hat{\psi})}, \tag{3.73}$$

where

$$\hat{\psi}(\hat{\theta}) \equiv \frac{\hat{\theta}}{1 + (q'-1)\,\hat{\theta} \cdot \langle C \rangle}. \tag{3.74}$$

Note that the transformation $q' \to 2 - q$ formally links the two types of

constraint. In particular, one can see that

$$\hat{p}_{q'}(G_i, \hat{\psi}) = p_{2-q'}(G_i, \hat{\psi}) = p_q(G_i, \hat{\psi}). \tag{3.75}$$

The difference between our and *UJK* approaches lies in the estimation of the Lagrange parameters, and in the fact that *UJK* assume that q is known. They consider a set of observations coming from systems which they *do not know being independent of each other or not* $\{C_m^*\}_{m=1}^M$, and select the Lagrange multiplier in order to satisfy:

$$\langle C \rangle(\hat{\theta}^*) = C_{JK}^*, \tag{3.76}$$

where $C_{JK}^* = \frac{1}{M}\sum_{m=1}^M C_m^*$ is simply the sample average of the observations. So, they are using the method of moments as estimation technique. It follows that their probability distribution becomes:

$$\hat{p}_{q'}(G_i) = \frac{\left[1 - (q'-1)\,\hat{\theta}^* \cdot (C(G_i) - C_{JK}^*)\right]_+^{\frac{1}{q'-1}}}{\hat{W}_{q'}(\hat{\theta}^*)}. \tag{3.77}$$

With our approach we consider instead a different situation. We imagine that our set of observations comes from replicas of the same system. Thus, we are considering an ensamble of systems which we *know* are independent of each other, but could internally have correlations which are better captured by generalized entropies (and, consequentially, non-factorizable probabilities). In other words, we are addressing the *hierarchical* nature of entropies, which arises from the scale-dependence of the presence of correlations. Even if our GMEP procedure does not satisfy *SJ4* (subset independence) axiom due to the presence of non-linear constraints, it is consistent with the fact that the resulting probabilities should behave differently depending on the scale of the conditioning subset in *SJ4*. Moreover, it is possible to establish a relationship between the two approaches, showing that our GMEP-ML procedure returns a probability distribution which is equivalent to the one *UJK* would obtain by considering a deformed C_{JK}^*. The probability distribution resulting from our approach has the form:

$$p_q(G_i, \psi^*) \propto \left[1 - (1-q)\,\psi^* \cdot C(G_i)\right]_+^{1/(1-q)}, \tag{3.78}$$

where ψ^* is related to θ^* through Eq. (3.41) and is here a numerical value satisfying Eq. (3.54). It is possible to rewrite ψ^* as a function of the linear average of C with respect to P_q:

$$\psi^* = \frac{\hat{\theta}^*}{1 + (1-q)\hat{\theta}^* \cdot \langle C \rangle}. \tag{3.79}$$

Plugging it back in Eq. (3.78), we get:

$$p_q(G_i, \hat{\theta}^*(\psi^*)) \propto \left[1 - (1-q)\,\hat{\theta}^* \cdot (C(G_i) - \langle C \rangle)\right]_+^{1/(1-q)}, \tag{3.80}$$

which is equivalent to Eq. (3.77) under the transformation $q \to 2-q'$ and by setting $C^*_{JK} = \langle C \rangle \neq 1/M \sum_m C^*_m$. The important difference with respect to the JK approach is that $\langle C \rangle$ is not the sample mean of the observation, but the theoretical average calculated with respect to the P_q resulting from our approach, based on GMEP and ML estimation. This is neither surprising nor in contradiction with SJ axioms or JK approach: while the sample mean is a good estimator when one wants to be completely unbiased with respect to the possible correlations among different systems, it fails when one *knows* that at certain scales probability distributions should factorize, but still wants to allow local correlations.

In order to complete the discussion, in the following we describe how our approach behaves when the linear average is constrained. According to the GMEP, we maximize the Lagrangian:

$$\mathcal{L}'_{q'}[\hat{P}] = S_{q'}[\hat{P}] - \alpha \left[\sum_{i=1}^{\Omega} \hat{p}(G_i) - 1\right] - \hat{\theta} \cdot [\langle C \rangle - C^*]\,. \tag{3.81}$$

We obtain then the probability distribution in Eq. (3.72), which is equivalent to Eq. (3.73), and use the ML principle to estimate $\hat{\psi}^*$ and q'^*. It is easy to show that ML on $\hat{\psi}$ leads to a condition which, under the $q' \to 2-q$ transformation, is equivalent to Eq. (3.54):

$$\sum_{i=1}^{\Omega} C(G_i)\, \hat{p}_{q'}^{2-q'}(G_i, \hat{\psi}^*_{q'}) = \frac{1}{M} \sum_{m=1}^{M} C(G^*_m)\, \hat{p}_{q'}^{1-q'}(G^*_m, \hat{\psi}^*_{q'}). \tag{3.82}$$

The same holds for the estimation of the entropic parameter, which satisfy

Eq. (3.64):

$$S_1[\hat{P}_{q'^*}(\hat{\psi}^*_{q'^*})] = -\overline{\ell}_{q'^*}(\hat{\psi}^*_{q'^*}). \tag{3.83}$$

Imagine now to have a set of observations and to estimate the parameters of $P_q(\psi)$ and $\hat{P}_{q'}(\hat{\psi})$, obtained respectively constraining $\langle C\rangle_q$ and $\langle C\rangle$. We would obviously get that if $q = q^*$, then $q'^* = 2 - q^*$ and that $\psi^* = \hat{\psi}^*$. So, according to Eq. (3.75), we would obtain the same probability distribution in both cases. However, the consistency between the GMEP and the ML is maintained only when the q-average is constrained. In fact, considering a single observation C^* corresponding to the state G^*, in the first case we have:

$$\mathcal{L}_{q^*}[P_{q^*}(\theta^*)] = S_{q^*}[P_{q^*}(\theta^*)] = -\ln p_{q^*}(G^*, \theta^*).$$

In the second case, instead:

$$\begin{aligned} \mathcal{L}'_{2-q^*}[P_{q^*}(\hat{\theta}^*)] &= S_{2-q^*}[P_{q^*}(\hat{\theta}^*)] - \hat{\theta}^* \cdot [\langle C\rangle - C^*] \\ &\neq -\ln p_{q^*}(G^*, \hat{\theta}^*) \\ &\neq -\ln p_{2-q^*}(G^*, \hat{\theta}^*). \end{aligned}$$

Thus, the correspondence between (minus) the log-likelihood and the Lagrangian is valid only in the first case, as well the one with the considered Rényi entropy (i.e. the one we maximize in the GMEP).

3.4 Conclusions

A large body of literature has discussed the generalized axiomatic definition of entropy deriving from the relaxation (or unrestricted interpretation) of some of the *SK* and *SJ* axioms (in particular, *SK4* and *SJ3*). It is known that, when generalized in that way, the definition of entropy leads to parametric entropy families where a specific value of the entropic parameter(s) usually retrieves the ordinary Shannon functional. In a maximum entropy approach, each entropy family leads to a corresponding family of maximum entropy probability distributions, indexed again by the entropic parameter(s), that provide the least biased inference about a system for which only limited information is available, in the form of empirical observations of a quantity treated as a soft constraint. Unfortunately, when the estimated maximum entropy distribution is 'put back'

into its defining generalized entropy, a number of inconsistencies typically arise, including incompatibility with the ML principle, impossibility of determining the value of the entropic parameter(s) purely from empirical data, and disconnection from Shannon entropy when multiple independent observations of the same system are available.

In this chapter, based on the fact that every member of an entropy family is ultimately intended as a quantification of the uncertainty encoded in the input probability distribution, we have introduced an uninformativeness axiom demanding that the maximally uncertain (i.e. uniform) probability distribution should always return the same (maximal) value of the entropy, irrespective of the value of the entropic parameters. This simple axiom implies that all entropies take values within the same interval $[0, \ln \Omega]$, where Ω is the number of possible (unconstrained) microstates of the system, thereby equipping generalized entropies with a universal scale and meaning. The axiom considerably restricts the admissible members of entropy families. In particular, for both the *UJK* and *HT* entropies, the axiom selects only Rényi entropy as viable. A notable counterexample, dismissed by the axiom, is Tsallis entropy. From an inferential point of view, the axiom guarantees that completely uninformative data (or equivalently the complete absence of empirical information) cannot be used to learn the value of entropic parameters. At the same time we have showed that, when informative data are available, a straightforward extension of the ML principle leads to the optimal estimation of the entropic parameter(s), purely from empirical observations and without making any assumptions.

The resulting generalized ML approach couples the determination of the entropic parameters with that of the other structural parameters (Lagrange multipliers) of the maximum entropy distribution. In particular, while the ML condition for the Lagrange multipliers indicates which specific combination of M independent observations should be put equal to the generalized mean value of the constraint, the one for the entropic parameters coincides with the requirement that the log-likelihood of the data equals minus Shannon entropy. This remarkable result shows that the connection between Shannon entropy and log-likelihood holds true also for generalized entropies (for the appropriate ML value of the entropic parameter) and is

consistent with the assumed independence of the M observations. When $M = 1$, the maximum entropy probability returns coinciding values of Rényi and Shannon entropies, even if it maximizes the former but not the latter. For multiple independent observations ($M > 1$), the connection between log-likelihood and Shannon entropy remains, while the connection with Rényi entropy disappears, as a result of independence. Therefore the log-likelihood, when maximized also over the entropic parameters, automatically finds the correct entropy to be used for model fitting and selection.

We believe that the introduction of the uninformative axiom has beneficial effects for statistical inference and its many applications, offering a way of constructing generalized entropies that have still controllable and consistent properties.

Chapter 4

On nonlinear compression costs: when Shannon meets Rényi

This chapter is based on: A. Somazzi, P. Ferragina, and D. Garlaschelli. *On nonlinear compression costs: when Shannon meets Rényi.* Available at `https://arxiv.org/abs/2310.18419`.

Shannon entropy is the shortest average codeword length a lossless compressor can achieve by encoding i.i.d. symbols. However, there are cases in which the objective is to minimize the *exponential* average codeword length, i.e. when the cost of encoding/decoding scales exponentially with the length of codewords. The optimum is reached by all strategies that map each symbol x_i generated with probability p_i into a codeword of length $\ell_D^{(q)}(i) = -\log_D \frac{p_i^q}{\sum_{j=1}^N p_j^q}$. This leads to the minimum exponential average codeword length, which equals the Rényi, rather than Shannon, entropy of the source distribution. We generalize the established Arithmetic Coding (AC) compressor to this framework. We analytically show that our generalized algorithm provides an exponential average length which is arbitrarily close to the Rényi entropy, if the symbols to encode are i.i.d.. We then apply our algorithm to both simulated (i.i.d. generated) and real (a piece of Wikipedia text) datasets. While, as expected, we find that the application to i.i.d. data confirms our analytical results, we also find that, when applied to the real dataset (composed by highly correlated symbols), our algorithm is still able to significantly reduce the exponential average codeword length with respect to the classical 'Shannonian' one. Moreover, we provide another justification of the use of the exponential average: namely, we show that by minimizing the exponential average length it is possible to minimize the probability that codewords exceed a certain threshold length. This relation relies on the connection between the exponential average and the cumulant generating function of the source distribution, which is in turn related to the probability of large deviations. We test and confirm our results again on both simulated and real datasets.

4.1 Introduction

In the realm of (lossless) data compression, the main goal is to efficiently represent data in a manner that requires reduced space without compromising its integrity. At the heart of this challenge lies the encoding strategy, which determines how individual symbols or sequences of symbols are transformed into compressed representations. Traditionally, these strategies aim to minimize the average length of the encoded symbols. By achieving a shorter average encoded symbol length, one can ensure a more compact representation of the entire input data, thereby achieving the central objective of many data compression problems.

Consider a stationary source generating symbols from an alphabet $\Sigma = \{x_1, \dots, x_N\}$ of size $|\Sigma| = N$, with probability $p = \{p_1, \dots, p_N\}$. Then, the problem consists in finding the encoding strategy which maps each symbol $x_i \in \Sigma$ into a D-ary codeword of length $\ell_D(i)$ such that

$$L(0) = \sum_{i=1}^{N} p_i \ell_D(i) \tag{4.1}$$

is minimized. $L(0)$ is the codewords' average length, and the use of such notation will be clarified later.

In his pioneering work [12], Shannon proved that for a source generating i.i.d. symbols, Eqn. (4.1) is minimized by all encoding strategies such that $\ell_D(i) = -\log_D p_i$, for all $i = 1, 2, \dots, N$. However, in most cases, strategies that guarantee such equality for each symbol do not exist but only get 'close' to it. This leads to the notorious relation

$$L(0) \geq H_1[p], \tag{4.2}$$

where $H_1[p] = -\sum_{i=1}^{N} p_i \log_D p_i$ is the Shannon entropy of the source, which can be understood as the codewords' minimum average length. The use of the subscript in H_1 will also be clarified later.

We would also like to mention that Eqn. (4.1) can be seen as a *cost function* C, because minimizing Eqn. (4.1) is equivalent to minimizing the cost of encoding/decoding $C(0) \propto L(0)$ under the assumption that such cost is linear in the codewords' length.

Beyond the conventional focus on the linear average of codeword lengths,

it's essential to acknowledge that this is not the only viable metric to target for minimization. For example, there could be a *nonlinear relation* between the cost of encoding/decoding symbols and their codewords' length. Delving deeper into the theoretical underpinnings of averages, we encounter the Kolmogorov-Nagumo (KN) averages [84, 85]: a more general family of averages that offers a richer landscape for exploration. One might be driven to consider minimizing these KN averages, recognizing the possibility of uncovering novel compression strategies and further refining data representation techniques that are suitable in different scenarios. Following the introduced notation, the codewords' KN average length is defined as

$$\langle \ell_D \rangle_\varphi = \varphi^{-1}\bigg(\sum_{i=1}^{N} p_i \, \varphi(\ell_D(i)) \bigg), \tag{4.3}$$

where φ is a continuous injective function. Note that for $\varphi(x) = x$ the usual average length (4.1) is recovered. While, in general, KN averages depend on φ, there is a natural requirement that an average length measure should satisfy, that restricts the space of admissible functions [107, 108]. Namely, it should be *additive* for independent symbols. In particular, consider two independent sets of symbols $\Sigma^{(1)} = \{x_1, \ldots, x_N\}$ and $\Sigma^{(2)} = \{y_1, \ldots, y_M\}$, respectively. The associated probabilities are $p = \{p_1, \ldots, p_N\}$ and $q = \{q_1, \ldots, q_M\}$, and each symbol is encoded in a codeword of length $\{\ell_D^{(1)}(i)\}_{i=1}^N$ and $\{\ell_D^{(2)}(j)\}_{j=1}^M$. Then, the additivity requirement is formulated as follows:

$$\begin{aligned} &\varphi^{-1}\bigg(\sum_{i=1}^{N}\sum_{j=1}^{M} p_i q_j \, \varphi\big(\ell_D^{(1)}(i) + \ell_D^{(2)}(j)\big) \bigg) \\ &= \varphi^{-1}\bigg(\sum_{i=1}^{N} p_i \varphi(\ell_D^{(1)}(i)) \bigg) + \varphi^{-1}\bigg(\sum_{j=1}^{M} q_j \varphi(\ell_D^{(2)}(j)) \bigg). \end{aligned} \tag{4.4}$$

It is possible to prove that Eqn. (4.4) leads to the so-called exponential KN averages [108, 109, 110], that correspond to $\varphi(x) = \varphi_t(x) = \gamma D^{tx} + b$. Substituting φ_t into Eqn. (4.3), one gets that

$$\langle \ell_D \rangle_{\varphi_t} \equiv L(t) = \frac{1}{t} \log_D \bigg(\sum_{i=1}^{N} p_i \, D^{t\,\ell_D(i)} \bigg), \tag{4.5}$$

where $t > -1$ and $L(t)$ is then the *exponential average* of the codeword's

length. Notice that for t approaching 0, the exponential average converges to the linear average, i.e. $\lim_{t\to 0} L(t) = L(0)$, which clarifies the notation we have adopted before.

The utility of the exponential average in data compression can be understood from two distinct fronts. Firstly, when the costs associated with the encoding or decoding steps amplify ($t > 0$), they might grow in an exponential fashion with respect to the codewords' lengths. This leads to a nonlinear relation having the form $C(t) \propto \sum_i p_i D^{t\ell_D(i)}$. Minimizing $C(t)$ is then equivalent to minimizing Eqn. (4.5) if $t > 0$, since the latter is a monotonically increasing function of the former. A case falling in such scenario could be DNA coding [18, 19], where the apparatus involved in encoding and decoding procedures is very costly. Minimizing this exponential cost function then could become essential for effective and efficient data handling. Secondly, at a more theoretical level, the exponential average arises naturally when aiming at curtailing the risk of buffer overflow [20, 111] or bolstering the probability of transmitting a message in a short timeframe. This could be the case in aerospace communication scenarios, where it can happen that antennas are visible for fleeting moments, necessitating the rapid and reliable transmission of information [112]. In such scenarios, estimating the likelihood of large deviations (for the events to be avoided) involves the cumulant generating function of the probability distribution, which in turn leads to the exponential average. Finally, it has been shown that minimizing Eqn. (4.5) with $t < 0$ is a problem related to maximizing the chance of receiving a message in a single snapshot [113].

In his valuable paper [114], Campbell proved that the optimal encoding lengths that minimize the exponential cost of Eqn. (4.5) are

$$\ell_D^{(q)}(i) = -\log_D \frac{p_i^q}{\sum_{j=1}^N p_j^q}, \tag{4.6}$$

where $q = 1/(1+t)$. Moreover, he proved that the lower bound for the exponential cost is given by the Rényi entropy of order $q = 1/(1+t)$ of the source, defined as

$$H_q[p] = \frac{1}{1-q} \log_D \left(\sum_{i=1}^N p_i^q \right), \tag{4.7}$$

so that

$$L(t) \geq H_{\frac{1}{1+t}}[p] \tag{4.8}$$

where the equality holds iff Eqn. (4.6) is exactly satisfied. Note that $\lim_{q\to 1} H_q[p] = H_1[p]$, i.e. Shannon entropy is a particular case of Rényi entropy. It follows that for $t \to 0$, Eqn. (4.8) reduces to Eqn. (4.2).

The probability distribution $p^{(q)} = \left\{ \frac{p_1^q}{\sum_{j=1}^N p_j^q}, \ldots, \frac{p_N^q}{\sum_{j=1}^N p_j^q} \right\}$ which appears in Eqn. (4.6) is often referred as *escort* or *zooming* probability distribution of p [72, 105, 22]. The reason is that, depending on the value of q, it can amplify/suppress values in the tails of the original distribution p (and, since it is normalized, suppress/amplify the others). Escort distributions have been applied and have emerged in various fields, ranging from non-extensive statistical mechanics [72], chaotic systems [105] and statistical inference [22]. Another notable link among the Rényi entropy, the KN exponential average, and escort distributions comes from an axiomatic point of view. While Shannon entropy can be derived by the four Shannon-Khinchin axioms (SK1-SK4) [13], Rényi entropy is derived by relaxing SK4 (also called *additivity* axiom) to a more general version, which involves both the KN exponential average and the escort distributions [83].

Since Campbell, from the point of view of data compression problems, escort distributions are also the optimal distributions according to which one has to encode symbols in order to minimize the exponential average codeword length $L(t)$. However, although Campbell provided the existence of an optimal encoding length, he did not suggest any operational strategy to achieve it. Some specific algorithms have been later proposed [20, 111, 115, 116], and [117] noted that, since the optimal lengths defined in Eqn. (4.6) have the same form of the lengths which minimize the linear average length of Eqn. (4.1) if p is replaced by its escort $p^{(q)}$, then it is sufficient to feed a standard (i.e. 'Shannonian') encoder with $p^{(q)}$ instead of p in order to reach a cost $L(t)$ close to its minimum $H_{\frac{1}{1+t}}[p]$.

In this chapter, we provide a series of contributions. i) We lay the mathematical ground to the observations of the previous papers by applying the above conceptual framework to one of the most efficacious algorithms in the realm of data compression: i.e., Arithmetic Coding (AC) (Sec. 4.2). ii) We experimentally analyze the performance of the proposed escort distribution-

based compressor in the case of optimizing the exponential average codeword length, over both synthetic and real datasets. We confirm the theoretical results on the former (composed by i.i.d. generated symbols) and achieve surprising results on the latter (composed by correlated symbols). In particular, we show that on a sample of Wikipedia text the application of our compressor with escort probability leads to an improved compression ratio (when the considered metric is the exponential average codeword length) with respect to a standard Shannon compressor, even if the optimal value of q (i.e. the exponent leading to the escort distribution) is unknown to the encoder (Sec. 4.3). iii) Finally, we examine analytically and experimentally the practical case in which its crucial to not exceed a certain threshold in the codewords' lengths (such as in the context of bounded buffers), by showing that the exponential average naturally appears in the probability of large deviations thus further justifying the study performed in the present chapter. In particular, we will show that by using our approach it is possible to significantly reduce the probability that the length of the codeword assigned to a given sequence of symbols exceeds a certain threshold with respect to a classic Shannon compressor (Sec. 4.4).

It goes without saying that all our results and experimental achievements could benefit of the use of more recent statistical compressors (i.e., ANS [118]) in place of the arithmetic coder, whose simplicity is exploited in this chapter just for clarity of explanation.

4.2 Methods

In the ensuing section, we undertake an examination of the arithmetic coding compression scheme. We commence by providing a theoretical description of AC, delineating its operating principles. Following this, we weigh the pros and cons of AC, offering a balanced viewpoint on its utility and limitations in various application contexts. Finally, we advance the discourse by generalizing AC with an aim to achieve the theoretical limit as predicted by Campbell's theorem.

4.2.1 Arithmetic coding

Arithmetic coding is a lossless encoding scheme [119]. Compressor and decompressor both need the alphabet of symbols Σ, the associated probability

distribution p and the length of the stream of symbols to encode/decode. Consider a string $\vec{s} = (s_1, \ldots, s_M)$ of length M, where each $s_j = x_{i_j}$ is a symbol randomly generated by a source from alphabet Σ with associated probability p. Then, in order to encode $\vec{s}$ into a D-ary alphabet, the encoder performs the procedure illustrated in Algorithm 1.

Algorithm 1 Arithmetic Coding

Require: The input string $\vec{s} = x_{i_1} x_{i_2} \cdots x_{i_M}$, the probabilities $p = \{p_1, \ldots, p_N\}$ and the cumulative $f = \{f_1, \ldots, f_N\}$ of p.
Ensure: A subinterval $[a, a + \mathcal{S})$ of $[0, 1)$.

$\mathcal{S}_0 = 1$
$a_0 = 0$
$j = 1$
while $j \neq M$ **do**
  $\mathcal{S}_j = \mathcal{S}_{j-1} \cdot p_{i_j}$
  $a_j = a_{j-1} + \mathcal{S}_{j-1} \cdot f_{i_j}$
  $j = j + 1$
end while
return $\langle k \in [a_M, a_M + \mathcal{S}_M), M \rangle$

Essentially the encoder, starting from the interval $[0, 1)$, iteratively divides it proportionally to the probabilities in p and, at each iteration j, chooses the subinterval corresponding to the associated symbol $s_j = x_{i_j}$. After M iterations, the encoder emits a number k, contained in the final subinterval $[a_M, a_M + \mathcal{S}_M)$, with $\mathcal{S}_M = \prod_{j=1}^{M} p_{i_j}$, which is uniquely associated with the original string $\vec{s}$. Such number k is then converted into its D-ary representation and communicated to the decoder (together with the original string length M), which can reverse this procedure to get the original string. It follows that the encoded string's length $\ell_D(\vec{s})$ is equal to the number of symbols (bits if $D = 2$) necessary to encode k in the desired alphabet.

From now on, we will consider for simplicity a binary ($D = 2$) encoding alphabet. Nonetheless, while our focus is on the classic binary AC, the results we present are inherently generalizable to $D > 2$, ensuring that the core features and principles of AC we discuss remain applicable and valid to those other cases too.

It is possible to show that the length of the encoded number k (i.e. encoded string) depends only on the length of the final subinterval $\mathcal{S}_M$. In particular, by choosing $k = a_M + \mathcal{S}_M/2$ and by truncating its binary

representation to the first $\lceil \log_2 \frac{2}{\mathcal{S}_M} \rceil$ bits, the approximation error is so small that such truncation is guaranteed to fall into the interval $[a_M, a_M + \mathcal{S}_M)$. Considering that

$$\ell_2(\vec{s}) = \left\lceil \log_2 \frac{2}{\mathcal{S}_M} \right\rceil < 2 - \log_2 \mathcal{S}_M = 2 - \sum_{j=1}^{M} \log_2 p_{i_j}, \tag{4.9}$$

and that it is possible, for M large enough, to approximate each p_i with the fraction of occurrences of symbol x_i in $\vec{s}$, i.e $p_i \simeq n_i(\vec{s})/M$, one gets that:

$$\ell_2(\vec{s}) < 2 + M \cdot H_0[p]. \tag{4.10}$$

Eqn. (4.10) unveils the main strength of the AC scheme: the number of bits that are 'wasted' in encoding $\vec{s}$ is 2, thus resulting *intensive* with respect to the string length M (provided that we operate with infinite precision arithmetic [119]). As M increases, the number of wasted bits per character goes to 0, in fact

$$\frac{\ell_2(\vec{s})}{M} < \frac{2}{M} \cdot H_0[p]. \tag{4.11}$$

A primary limitation of arithmetic coding (AC) lies in its operational framework. Unlike certain encoding schemes that allocate distinct codewords to individual symbols, AC assigns a codeword to the entire string. This means that the decoding process cannot commence in tandem with encoding so the decoder must wait for the encoder's completion of encoding the entire string (see e.g. the variant Range Coding for relaxing this limitation [120]). As the efficiency of AC generally improves with an increase in M, this waiting period can be time-consuming, rendering AC unsuitable for some applications. Conversely, AC boasts superior performance compared to encoding mechanisms that designate codewords to each symbol particularly when probability distributions are highly skewed. Such encoders mandate a minimum of 1 bit per symbol. However, the optimal length — expressed as $-\log_2 p_i$ — can be significantly less than 1.

4.2.2 Generalized AC

We now propose a generalization of AC in order to optimally minimize the exponential cost $L(t)$ defined in Eqn. (4.5). In analogy with the classical case, we try to execute AC by dividing each segment according to the escort distribution $p^{(q)}$, where $p_i^{(q)} = \frac{p_i^q}{\sum_{j=1}^N p_j^q}$, in order to reach the optimal lengths defined in Eqn. (4.6). We will call this procedure AC_q. Moreover, we will call $\mathcal{S}_j^{(q)}$ the length of the segment generated by AC_q at iteration j. The logarithm of the length of the final segment $\mathcal{S}_M^{(q)}$ for a string $\vec{s}$ is:

$$\begin{aligned}\log_D \mathcal{S}_M^{(q)}(\vec{s}) &= \log_D \prod_{j=1}^{M} \frac{p_{i_j}^q}{\sum_{i=1}^N p_i^q} \\ &= \sum_{j=1}^{M} \left(\log_D p_{i_j}^q - \log_D \sum_{i=1}^{N} p_i^q \right) \\ &= q \sum_{i=1}^{N} n_i(\vec{s}) \log_D p_i - M \log_D \sum_{j=1}^{N} p_j^q\end{aligned} \tag{4.12}$$

where $n_i(\vec{s})$ counts how many times the symbol x_i appears in the string $\vec{s}$. From this result, it is possible to evaluate the number of bits emitted to encode a particular string in a binary alphabet ($D = 2$):

$$\ell_2^{(q)}(\vec{s}) = \left\lceil \log_2 \frac{2}{\mathcal{S}_M^{(q)}(\vec{s})} \right\rceil < 2 - \log_2 \mathcal{S}_M^{(q)}(\vec{s}). \tag{4.13}$$

Let's define now the exponential cost $L_M(t)$ of a string of length M composed by independent symbols:

$$L_M(t) = \frac{1}{t} \log_2 \sum_{\vec{s}} P(\vec{s})\, 2^{t\ell_2(\vec{s})}, \tag{4.14}$$

where $P(\vec{s}) = \prod_{i=1}^N p_i^{n_i(\vec{s})} = \mathcal{S}_M(\vec{s})$. Given Eqn. 4.5, it is $L_M(t) = M \cdot L(t)$. Substituting Eqn. (4.13) in the definition of $L_M(t)$, and considering the

optimal parameter value $q = 1/(1+t)$, we get that:

$$\begin{aligned} L_M(t) &= \frac{1}{t}\log_2 \sum_{\vec{s}} P(\vec{s})\, 2^{t\ell_2^{((t+1)^{-1})}(\vec{s})} \\ &< \frac{1}{t}\log_2 \sum_{\vec{s}} P(\vec{s})\, 2^{t(2-\log_2 \mathcal{S}_M^{((t+1)^{-1})}(\vec{s}))} \\ &= \frac{1}{t}\log_2 \left(2^{2t}\, 2^{tM\log_2 \sum_j p_{i_j}^{(t+1)^{-1}}} \right. \\ &\qquad \left. \cdot \sum_{\vec{s}} \left(P(\vec{s}) \prod_{i=1}^{N} \left(p_i^{n_i(\vec{s})}\right)^{-\frac{t}{t+1}} \right) \right) \\ &= 2 + \frac{Mt}{t+1} H_{\frac{1}{t+1}}[p] + \frac{1}{t}\log_2 \sum_{\vec{s}} P(\vec{s})^{1-\frac{t}{t+1}} \\ &= 2 + \frac{Mt}{t+1} H_{\frac{1}{t+1}}[p] + \frac{M}{t+1} H_{\frac{1}{t+1}}[p] \\ &= 2 + M H_{\frac{1}{t+1}}[p]. \end{aligned} \tag{4.15}$$

Which reads:

$$L_M(t) < 2 + M H_{\frac{1}{t+1}}, \tag{4.16}$$

where $H_q[p] = \frac{1}{1-q}\log_2 \sum_{i=1}^{N} p_i^q$ is the Rényi entropy of the source for a single symbol. Notice that for independent symbols the Rényi entropy is additive, i.e. for i.i.d. symbols $H_q[P] = M \cdot H_q[p]$ holds. So, the compressor AC_q leads to an average cost per symbol which is close to $H_{\frac{1}{1+t}}[p]$ as M increases:

$$L(t) = \frac{L_M(t)}{M} < \frac{2}{M} + H_{\frac{1}{1+t}}[p]. \tag{4.17}$$

It is possible to visualize this result by considering the cost $L_M(t,q)$, in which the parameters t and q are now decoupled: t is the exponent of the cost function, while q is used in the AC_q procedure. In particular, it reads:

$$L_M(t,q) = \frac{1}{t}\log_2 \sum_{\vec{s}} P(\vec{s})\, 2^{t\,\ell_2^{(q)}(\vec{s})}. \tag{4.18}$$

Here, $\ell_2^{(q)}$ represents the number of bits emitted by applying the AC_q procedure with the escort distribution of order q (notice that, if $q = 1$, then the

$\mathrm{AC}_{q=1}$ reduces to the classic compressor AC).

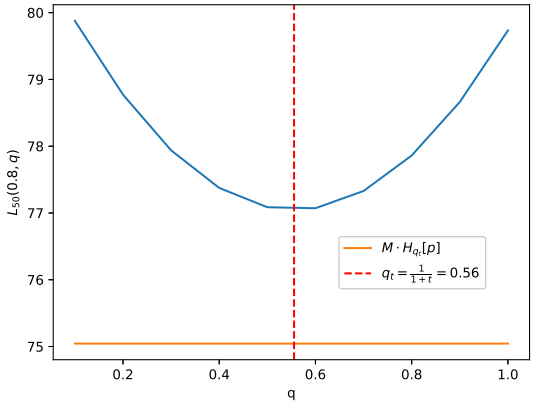


Figure 4.1: Exponential average codeword length of strings of length $M = 50$, composed by i.i.d. symbols sampled according to $p_i \propto i^{-1}$, $i \in [1, 3]$. Here $t = 0.8$. Its minimum is reached in the proximity of the red vertical dashed line, corresponding to the optimal q, i.e. q_t. For that value of q, the distance with respect to the Rényi entropy of the string (flat orange line) is approximately 2.

Figure 4.1 shows $L_M(t, q)$ for different values of q. The minimum is reached exactly in the value of q prdicted by Campbell, that we now call $q_t = \frac{1}{1+t}$. Moreover, the distance between the minimum of $L_M(t, q)$, i.e. $L_M(t, q_t)$, and the orange line (corresponding to $M \cdot H_{\frac{1}{1+t}}[p]$) is very close to 2, confirming the result of Eqn. (4.16).

4.2.3 A note on the semi-static approach

In this section, we discuss how the probability distribution p can be measured for different encoding schemes, focusing on the AC_q. An encoding scheme can be *static* or *semi-static*, depending on how the probability distribution of the source is computed or updated. In the first case, the probability distribution approximating the source's one is fixed and never changed while strings are generated. In the second case, instead, the probability distribution is evaluated each time a string needs to be encoded, and it is set equal to the frequency of symbols appearing in that string. In [118], the AMS coding is focused on the second case.

In this section, we want to elucidate how the use of a semi-static approach, instead of a static one, affects the exponential cost of Eqn. (4.18).

Let's assume that is possible to reach codewords' lengths as expressed in Eqn. (4.6), for any symbol and any value of q. Then, it is possible to write:

$$\begin{aligned}\ell_D^{(q)}(\vec{s}) &= \sum_{j=1}^{M}\left(-\log_D \frac{p_{i_j}^q}{\sum_{i=1}^N p_i^q}\right) \\ &= -\sum_{i=1}^{N} n_i(\vec{s}) \log_D \frac{p_i^q}{\sum_{i=1}^N p_i^q} = M \cdot H_1[f(\vec{s})||p^{(q)}],\end{aligned} \tag{4.19}$$

where $H_1[f\,||\,p] = -\sum_i f_i \log_D p_i$ is the *cross-entropy* between distributions f and p, and $f(\vec{s}) = \left(\frac{n_1(\vec{s})}{M}, \ldots, \frac{n_N(\vec{s})}{M}\right)$ is the empirical frequency of each symbol in the string $\vec{s}$. We also remind that $p^{(q)}$ is the escort distribution of order q of the distribution p. So, Eqn. (4.18) can be rewritten as:

$$L_M(t,q) = \frac{1}{t}\log_D\left(\sum_{\vec{s}} P(\vec{s})\, D^{t\,M\,H_1[f(\vec{s})\,||\,p^{(q)}]}\right). \tag{4.20}$$

While Campbell [114] showed that the best strategy (i.e., the best q) to minimize the exponential cost consists of taking $q = q_t = 1/(1+t)$, in the semi-static approach, the exponential cost of each string is minimized individually by taking $q = 1$. The reason is that the cost of encoding a single string is $D^{tMH_1[f(\vec{s})||p^{(q)}]}$. Since it is assumed that the probability of the source is equal to the empirical frequency appearing in the string to encode, i.e. $p = f(\vec{s})$, setting $q = 1$ provides the lowest cost (if $t > 0$) since

$$H_1[f^{(1)}(\vec{s})||f^{(1)}(\vec{s})] < H_1[f^{(1)}(\vec{s})||f^{(q)}(\vec{s})], \quad \forall q \geq 0.$$

In other words, if one assumes that $p = f(\vec{s})$ then all the observed strings are encoded as if they were members of the *typical* set of strings, thus they are better encoded by considering $q = 1$, i.e. the Shannon-like approach [121]. Notice that if more than one string $\vec{s_i}$ is to be encoded, by assuming $p = f(\vec{s_i})$ at each string, one has to take into account that the probability distribution of the source is non stationary since, in general, $f(\vec{s}_i) \neq f(\vec{s}_j)$. This violates Campbell's hypothesis, thus making our compression approach non applicable in this case.

On the other hand, the situation is much different if one considers the static approach. In this case, the strings $\vec{s}_i$ are considered to be generated by a stationary source according to a distribution p. So it becomes possible

to observe strings whose corresponding $f(\vec{s})$ is outside the typical set of p, meaning that they are very expensive to encode and thus making the use of our approach very advantageous in the case of an exponential cost.

4.3 Application to Wikipedia

Having delineated the theoretical side of our AC_q in the preceding sections, we now transition to a more empirical scenario. This section is dedicated to the application of our outlined procedure to real-world data.

In particular, we applied AC_q to Wikipedia data.[1] The dataset used for our analysis contains $W \approx 7 \cdot 10^8$ symbols from an alphabet Σ of size $|\Sigma| = N = 27$. In order to perform coding in a static approach as we mentioned earlier, we computed from the whole dataset the empirical frequency of the 27 distinct symbols, shown in Fig.4.2, and then used it to set the probability distribution $p = \{p_1, \ldots, p_{27}\}$.

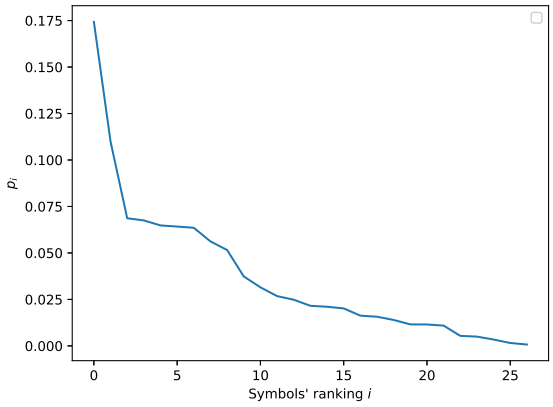


Figure 4.2: Probability distribution of the individual symbols (i.e., characters) in the Wikipedia dataset. Symbols have been ordered by decreasing frequency and assigned a rank. The probability distribution has been estimated in a frequentist approach as $p_i = n_i/W$, with n_i being the number of times that symbol x_i appears in Wikipedia.

Since the theoretical results presented so far are valid for i.i.d. symbols, we first discuss and apply our procedure in the case of i.i.d. symbols. After that, we will move to the real Wikipedia dataset.

[1] The dataset FIL9 can be downloaded from `https://fasttext.cc/docs/en/unsupervised-tutorial.html`.

To begin with, we generated $\eta = 3.500.000$ strings of length $M = 20$ composed by i.i.d. symbols sampled according to p. We then applied the AC_q for different and discretized values of q. For each string we evaluated the length of the corresponding codeword generated by AC_q algorithm, without actually generating it, as $\ell_2^{(q)}(\vec{s}) = \log_2\lceil \frac{2}{\mathcal{S}_M^{(q)}(\vec{s})} \rceil$. Such lengths have been stored in a matrix $\mathcal{L}$, whose entry $\mathcal{L}_{ij}$ is the length of the codeword of the j-th string, generated with AC_{q_i}, where q_i is the i-th value of q that we encode with. Algorithm 2 summarizes this procedure. By using the

Algorithm 2 Wikipedia data analysis

Require: data, $\Sigma = (x_1, \ldots, x_N)$, $p = (p_1, \ldots, p_N)$
Ensure: $\mathrm{len}(data) = M \cdot \eta$
Ensure: $q_{end} > 0$
$M \leftarrow 20$
$\eta \leftarrow \frac{\mathrm{len}(data)}{M}$
$q \leftarrow 0$
$q_{idx} \leftarrow 0$
while $q \leq q_{end}$ **do**
  $n \leftarrow 0$
  $p_i^{(q)} \leftarrow \frac{p_i^q}{\sum_j p_j^q} \quad \forall x_i \in \Sigma$
  while $n < \eta$ **do**
    string $\leftarrow$ data$[M \cdot n : M \cdot (n+1) - 1]$
    $\mathcal{S} \leftarrow 1$
    for c in string **do**
      $i^* \leftarrow i | c = x_i$
      $\mathcal{S} \leftarrow \mathcal{S} \cdot p_{i^*}^{(q)}$
    end for
    $\mathcal{L}[q_{idx}, n] \leftarrow \log_2\lceil \frac{2}{\mathcal{S}} \rceil$
    $n \leftarrow n + 1$
  end while
  $q \leftarrow q + 0.1$
  $q_{idx} \leftarrow q_{idx} + 1$
end while

matrix $\mathcal{L}$ generated by Algorithm 2, it is possible to evaluate the empirical exponential average length, for different values of t, as:

$$L_M^{emp}(t, q_i) = \frac{1}{t} \log_2 \left(\frac{1}{\eta} \sum_{j=1}^{\eta} 2^{t\mathcal{L}_{ij}} \right). \tag{4.21}$$

Figure 4.3 shows the empirical L_M^{emp} as a function of q for three different

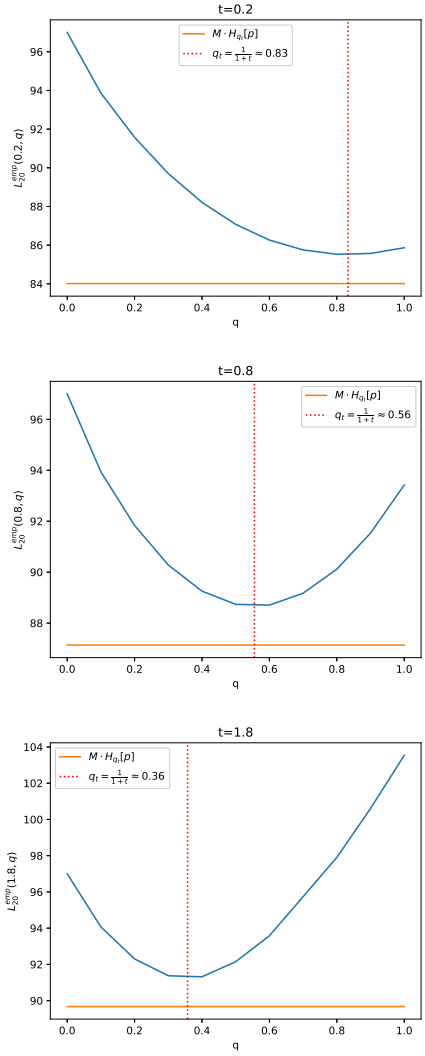


Figure 4.3: This Figure shows, for the synthetic i.i.d. generated symbols, the empirical exponential average $L_M^{emp}(t,q)$ for $M = 20$ (blue line), the Rényi entropy $M \cdot H_{q_t}[p]$ (horizontal orange line) with $q_t = 1/(1+t)$ (dotted vertical red line). The three panels show different values of $t \in \{0.2, 0.8, 1.8\}$. We can see that, in each case, the minimum of $L_M^{emp}(t,q)$ is reached at $q = q_t$, and that $L_M^{emp}(t,q_t) - M \cdot H_{q_t}[p] \approx 2$.

values of t, with the corresponding Rényi entropy $M \cdot H_{q_t}$ and the optimal $q_t = 1/(1+t)$. It is clear that the minimum of $L_M^{emp}(t,q)$ is reached at $q = q_t$ and that it is very close to the Rényi entropy $M \cdot H_{q_t}[p]$.

Let us now move to analyze the real Wikipedia dataset. We divided it in $\eta = 35.653.488$ strings of length $M = 20$. We applied again Algorithm 2, and proceeded in the same way as we did for the i.i.d. symbols

scenario. Figure 4.4 reports our results. In particular, we can see that $\text{argmin}_q(L_M^{emp}(t,q)) < q_t$ and that, for $t = 0.2$ (top panel), our AC_q can perform better than what Campbell predicted (in fact, $\min_q L_M^{emp}(t,q) < M \cdot H_{q_t}[p]$). The emergence of such discrepancies is not surprising, since real English text is not composed by i.i.d. symbols, and thus the hypothesis on which our theoretical description lies is not satisfied.

But what does it mean, 'physically', the fact that, in this case, the average empirical cost is minimized by considering a q smaller than q_t? Since we are using escort distributions $p^{(q)}$ of order q as encoding strategy in Eqn. (4.6), decreasing the value of q is equivalent to increasing the probability of the rare strings. This translates into assigning them shorter codewords, more than it would be done by using $q = q_t$. In other words, when the real optimal q is smaller than q_t, this means that 'rare' strings are actually more abundant in the dataset than they would be if they were generated by a probability distribution calculated as the product of the probability of i.i.d. symbols. Figure 4.5 shows, for different values of t, the real (empirical) optimal q overlapped to the theoretical q_t. For most values of the exponent t, the empirical best q is smaller than q_t.

Finally, we want to stress that even if English text does not satisfy the i.i.d. symbols hypothesis on which Campbell's theoretical description lies, the use of the AC_q still outperforms the standard AC if the average length is exponential, although the empirical optimal q is not the one predicted by Campbell. In fact, while the value of q that is actually optimal in the case real English text can not be known a priori, by using the one which is optimal for i.i.d. symbols (i.e. q_t) it is possible to significantly reduce the exponential average length, or the cost, with respect to the standard case $q = 1$. This is shown in Figure 4.6, where we can see that, even if the true optimal q is different from q_t, by encoding according to $q_t = \frac{1}{1+t}$ there is a notable exponential average length drop with respect to the usual $q = 1$ encoding strategy. Of course, if one would know the true optimal q the advantage would be even greater.

4.4 Discussion

In this section, we'll take a closer look at the core ideas and findings from our research. We'll first explore one of the reasons behind the use of the

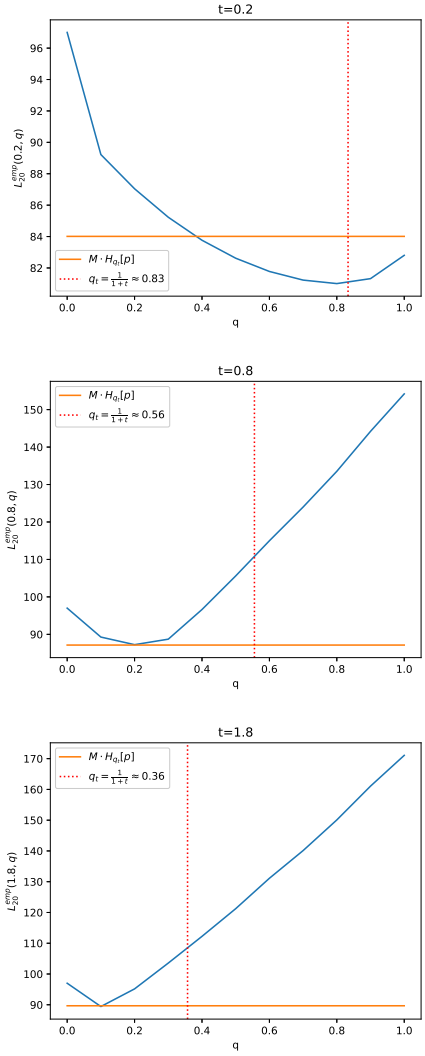


Figure 4.4: This Figure shows, for the real Wikipedia dataset,, the empirical exponential average $L_M^{emp}(t, q)$ for $M = 20$ (blue line), the Rényi entropy $M \cdot H_{q_t}[p]$ (horizontal orange line) with $q_t = 1/(1 + t)$ (dotted vertical red line). The three panels show different values of $t \in \{0.2, 0.8, 1.8\}$. In all these cases, it is instead evident that the minimum of $L_M^{emp}(t, q)$ is reached for $q < q_t$. Additionally, $\min_q L_M^{emp}(q, t)$ could be lower (first panel) or almost exactly reach (second and third panels) the value $M \cdot H_{q_t}[p]$. We can also see that encoding according to AC_{q_t} (i.e., see the intersection between the blue line and the dotted line) can lead to an exponential average length smaller than the Rényi entropy (first panel), or to an error which greater than 2, i.e. $\min_q L_M^{emp}(t, q) - M \cdot H_{q_t}[p] > 2$ (second and third panels).

exponential cost in our study, thus explaining why it's important and how it

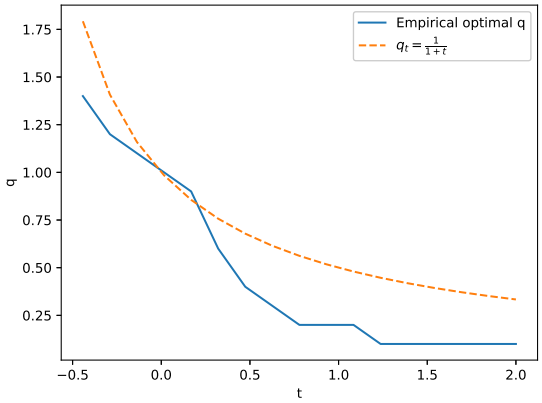


Figure 4.5: Empirical best q for the Wikipedia dataset (blue line solid line) and q_t (orange dashed line) for different values of t. For almost every value of t, the empirical optimal q is smaller than q_t, meaning that in the Wikipedia dataset there is an abundance of 'rare' strings.

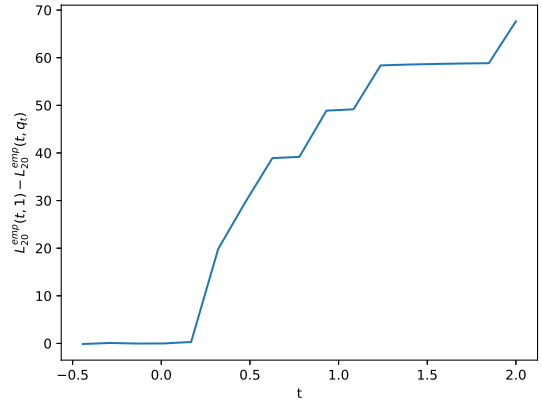


Figure 4.6: Difference $L_M^{emp}(t, q = 1) - L_M^{emp}(t, q = q_t)$ for $M = 20$ as a function of t, for the Wikipedia dataset. Since, for t big enough, this quantity is positive (and increasing), if the average to consider is the exponential one, there is an advantage (which increases with t) in encoding according to q_t instead of $q = 1$.

fits into the bigger picture of data compression. After that, we will provide further analysis of real data, which confirms our previous findings. Moreover, we will also mention what are the errors that can come up when our guesses about the true probability distribution of the source are not accurate, shedding light on some of the challenges we faced and how they might be addressed in future studies.

4.4.1 A justification to the exponential cost with Cramér's theorem

In this subsection, we provide a simple yet powerful idea about the usefulness of the exponential average and its minimization. Such idea relies on the linkage between the exponential average and the cumulant generating function of a distribution. As we anticipated in the introduction of this chapter, such application could be useful in scenarios in which it is imperative to minimize the probability that codewords' lengths exceed a certain threshold.

Suppose that we are interested in encoding strings of fixed length M, and that we do not want the corresponding codewords length's exceed the threshold $M \cdot a$. So, using the usual notation, the aforementioned problem translates into finding an encoding strategy $x_i \to \ell_D(i)$ that assigns to each symbol x_i of the alphabet Σ a length $\ell_D(i)$ to its codeword minimizing:

$$\text{Prob}\left[\frac{1}{M}\sum_{j=1}^{M}\ell_D(i_j) \geq a\right]. \tag{4.22}$$

Here, the sum runs over the whole string, and $\ell_D(i_j)$ is the length of the encoded symbol appearing at the j-th position of the string to be encoded. Moreover, since the threshold for the encoded strings is $M \cdot a$, we can see a as the threshold *per symbol*. According to Cramér's theorem, it is possible to write the following Chernoff bound:

$$\text{Prob}\left[\frac{1}{M}\sum_{j=1}^{M}\ell_D(i_j) \geq a\right] \leq e^{-M\left(ta-\mu(t)\right)} \quad \forall t > 0, \tag{4.23}$$

where $\mu(t) = \log \mathbb{E}_p[e^{t\ell_D(i)}]$ is the symbols' distribution's cumulant-generating function and log is the natural logarithm (i.e. with base e). Eqn. (4.23) gives us an important degree of control on the probability of exceeding the threshold, since, as we will show, it is possible to control its upper bound. Of course, we are interested in situations in which the exponent $-M\big(ta - \mu(t)\big)$ is negative, otherwise, we would get an upper bound of a probability distribution greater than 1, thus totally uninformative. It

is possible to rewrite the exponent of the upper bound as:

$$\begin{aligned} -M(ta-\mu(t)) &= -M\left(ta-\log\big(\sum_i p_i e^{t\ell_D(i)}\big)\right) \\ &= -M\left(ta-\frac{t\log_D(\sum_i p_i D^{t\ell_D(i)\log_D e})}{t\log_D e}\right) \\ &= -M\cdot t(a-L(t\log_D e)). \end{aligned} \tag{4.24}$$

Since M, t and a are positive by definition, we are interested in finding the strategy minimizing $L(t\log_D e)\ \forall t>0$. We know that, for a given value of $t'=t\log_D e$, the minimum of $L(t')$ is $H_{q_{t'}}[p]$, with $q_{t'}=\frac{1}{1+t'}$. Moreover, we know that such minimum is reached with the strategy $\ell_D^{(q)}(i)=-\log_D p_i^{(q)}$ (see Eqn. (4.6)). So, by writing Eqn. (4.24) as a function of $q_{t'}$ (and, for simplicity, by dropping the subscript 't''), we get that:

$$-M(ta-\mu(t)) = -M\frac{1-q}{q\log_D e}(a-H_q[p]). \tag{4.25}$$

So, it is possible to write Eqn. (4.23) as:

$$\text{Prob}\left[\frac{1}{M}\sum_{j=1}^{M}\ell_D(i_j)\geq a\right] \leq e^{-M\frac{1-q}{q\log_D e}(a-H_q[p])}\ \ \forall q\in(0,1]. \tag{4.26}$$

Having pointed out that the best strategy consists in setting the codewords lengths according to Eqn. (4.6), with $q=q_{t'}=1/(1+t')$, we have to determine which is the correct $t>0$ (and, in turn, q_t) to consider. We expect that the choice depends on the threshold a. In order to choose the best parameter q, we will minimize the right-hand side of Eqn. (4.25). Since we are assuming it to be negative, this guarantees that the upper bound in Eqn. (4.26) is minimized.

Before going into the analytical details of such minimization, we will consider two simple examples which will provide an intuition on how the encoding strategy is related to the threshold a. Recall that $H_q[p]$ is a decreasing function of q, i.e., $H_0[p]\geq\cdots\geq H_1[p]$.

Case $a > H_0[p] = \log_D|\Sigma|$. In the first case we consider, we assume that the threshold a is bigger than the Rényi entropy of order 0. Since

$H_0[p] = \log_D |\Sigma|$, we are assuming that the threshold exceeds the Shannon entropy of a distribution that shares the same support as the original p, but with entries replaced by $1/|\Sigma|$, i.e. a uniform distribution. In this scenario, the term $(a - H_q[p])$ in Eqn. (4.25) is positive and finite $\forall q \in (0, 1]$. The r.h.s. of Eqn. (4.25) is then maximized by letting $q \to 0$ (i.e. $t \to +\infty$). By writing $a = H_0[p] + \epsilon$, with $\epsilon > 0$, Eqn. (4.26) reads:

$$P\left[\frac{1}{M}\sum_{j=1}^{M}\ell_D(i_j) \geq H_0[p] + \epsilon\right] \leq \lim_{q\to 0} e^{-M\frac{1-q}{q\log_D e}\epsilon} = 0. \tag{4.27}$$

So, the probability of emitting a codeword longer than the threshold vanishes. This result is trivial: by setting $q \to 0$, the encoding strategy is equivalent to the Shannon encoding for symbols generated with a uniform probability distribution. In fact, $\ell_D^{(0)}(i) = -\log p_i^{(0)} = \log |\Sigma|$, and this holds for any probability distribution p. In other words, if it is imperative that the average codeword length does not exceed $H_0[p]$, just encode the sequence as if the symbols are uniformly distributed, irrespective of their actual probability distribution.

Case $a < H_1[p]$. In this second case, we are going to consider a threshold smaller than the Shannon entropy of the underlying probability distribution p. So, it follows that $(a - H_q[p])$ is negative $\forall q$ because $H_0[p] \geq \cdots \geq H_1[p] > a$. Then, by setting $a = H_1[p] - \epsilon$, with $\epsilon > 0$, Eqn. (4.26) reads:

$$P\left[\frac{1}{M}\sum_{j=1}^{M}\ell_D(i_j) \geq H_1[p] - \epsilon\right] \leq e^{-M\frac{1-q}{q\log_d e}\left(H_1[p]-\epsilon-H_q[p]\right)}. \tag{4.28}$$

The exponent $-M\frac{1-q}{q\log_d e}\big(H_1[p] - \epsilon - H_q[p]\big) > 0$ is positive $\forall M \in \mathbb{N}$, and so it does not satisfy our hypothesis of a negative exponent. As previously mentioned, this means that the above right-hand side term is greater than 1. For this reason, it gives no information on the probability of exceeding the threshold. We can however see that since $H_1[p]$ is the shortest achievable codewords' (linear) average length, the latter can be smaller than $H_1[p]$ only due to fluctuations in the observed symbols frequency, which are suppressed in the large M limit. The best strategy is then letting $q = 1$, but still, the threshold will be exceeded almost always if M is not unrealistically small.

Case $H_1[p] \leq a \leq H_0[p]$. Now that we have shown the two extreme cases $a > H_0[p]$ and $a < H_1[p]$, let's focus our attention on the most interesting case: i.e. $H_0[p] \leq a \leq H_1[p]$. As previously mentioned, we are interested in finding the value $q = q^*$ for which $-M\frac{1-q}{q\log_D e}(a - H_q[p])$ is minimized. Taking the derivative, one gets:

$$\begin{aligned}\frac{d}{dq}\Bigg(- M\frac{1-q}{q\log_D e}(a - H_q[p])\Bigg) = \\ = -\frac{M}{\log_D e}\Bigg(\frac{1}{q(1-q)}D_{KL}(p^{(q)}||p) - \frac{1}{q^2}(a - H_q[p])\Bigg),\end{aligned} \tag{4.29}$$

where $D_{KL}(p^{(q)}||p) = \sum_i p_i^{(q)} \log_D \frac{p_i^{(q)}}{p_i}$ is the Kullback-Leibler divergence between the escort of p and p itself. The minimum is then found by setting the derivative to zero, leading to the condition:

$$\begin{aligned}a - H_{q^*}[p] &= \frac{q^*}{1-q^*}D_{KL}(p^{(q^*)}||p) \\ &= \frac{q^*}{1-q^*}\sum_{i=1}^{|\Sigma|} p_i^{(q^*)} \log_D \frac{p_i^{(q^*)}}{p_i} \\ &= \frac{q^*}{1-q^*}\Bigg[\sum_{i=1}^{|\Sigma|}\Bigg(\frac{p_i^{q^*}}{\sum_{j=1}^{|\Sigma|} p_j^{q^*}} \log_D p_i^{q^*-1}\Bigg) \\ &\qquad - \sum_{i=1}^{|\Sigma|}\Bigg(\frac{p_i^{q^*}}{\sum_{j=1}^{|\Sigma|} p_j^{q^*}} \log_D \sum_{j=1}^{|\Sigma|} p_j^{q^*}\Bigg)\Bigg] \\ &= q^*(H_1[p^{(q^*)}||p] - H_{q^*}[p]).\end{aligned} \tag{4.30}$$

Moreover, it is useful to write the Shannon entropy of the escort $p^{(q)}$:

$$\begin{aligned}H_1[p^{(q)}] &= -\sum_{i=1}^{|\Sigma|}\frac{p_i^q}{\sum_{j=1}^{|\Sigma|} p_j^q}\log_D \frac{p_i^q}{\sum_{j=1}^{|\Sigma|} p_j^q} \\ &= qH_1[p^{(q)}||p] + (1-q)H_q[p]\end{aligned} \tag{4.31}$$

By plugging Eqn. (4.31) into Eqn. (4.30), one gets that the value q^* which sets the derivative to 0 (i.e. minimizes the upper bound in the r.h.s. of

Eqn. (4.26)) satisfies:

$$H_1[p^{(q^*)}] = a. \tag{4.32}$$

This equation relates the threshold a to the encoding strategy driven by p^{q^*}. In particular, such relation unveils that, if $a \in [H_1[p], H_0[p]]$, the optimal encoding strategy 'pretends' that the symbols are generated according to their distribution's escort instead of the original p. Then, since by encoding with AC_{q^*}, we are actually (almost) reaching the shortest linear average length if symbols were generated according to $p^{(q^*)}$, it is reasonable that the best $q = q^*$ is the one for which the threshold is such shortest linear average, i.e. $H_1[p^{(q^*)}]$.

Summarizing our contributions in this section, we note that we have justified the use of the exponential average by the necessity of not exceeding a certain threshold in the length of the encoded string. In particular, given the value of the threshold a as an input, the procedure has three steps:

1. Estimate the probability distribution p of the input symbols.
2. Find q^* by solving Eq. (4.32).
3. Encode the input data with AC_{q^*}.

Such procedure guarantees that, if $a > H_1[p]$, it is possible to reduce the number of codewords exceeding the threshold with the use of the described AC_q algorithm, which reaches the Rényi entropy bound with an error of at most 2 bits.

In the following paragraph, we will show a couple of examples over real and simulated data on how to infer the proper q^*, and how much this choice impacts the fraction of strings exceeding the threshold.

4.4.2 Example

Throughout this section, we will apply our procedure to both the usual Wikipedia dataset and simulated strings composed by i.i.d. symbols. We will generate the latter according to the probability $p = (p_1, \ldots, p_{27})$ extracted from the Wikipedia dataset (see Figure 4.2 for a visual reference). In order to understand which is the range of interest for the threshold a, we have evaluated that $H_0[p] \approx 4.75$ and $H_1[p] \approx 4.12$. For this reason, we will

consider a threshold $a \in [4.12, 4.75]$. Figure 4.7 shows both the value q^* for

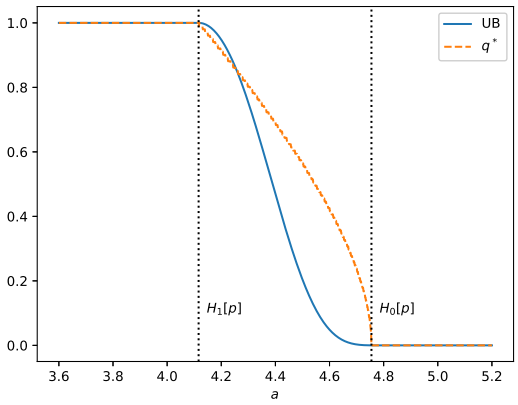


Figure 4.7: Dashed orange line: q^* as solution of Eqn. (4.32). Solid blue line: upper bound of the probability of exceeding the threshold. Black dotted vertical lines: $H_1[p]$ and $H_0[p]$. As the threshold a increases, the probability of exceeding it decreases until it reaches 0 for $a = H_0[p]$.

different values of a, evaluated as the solution of Eqn. (4.32), and the corresponding upper bound (UB) of the probability of exceeding the threshold, evaluated as:

$$\text{UB} = e^{\left(-M \frac{1-q^*}{q^* \log_D e}(a - H_{q^*}[p])\right)}, \tag{4.33}$$

where we set $M = 20$. For $a \leq H_1[p]$, the upper bound UB is equal to 1, thus it gives no information on the probability of exceeding the threshold. Instead, when a increases, UB gets smaller until, for $a \geq H_0[p]$, it reaches 0 (and so does q^*), meaning that if the threshold is bigger than $H_0[p]$, by encoding with escort distribution of order 0 it becomes impossible to exceed the threshold. This agrees with our previous analysis.

So, we expect that, by applying AC_{q^*} to both Wikipedia and simulated data, the fraction of strings that exceed the threshold $M \cdot a$ is smaller than the one obtained by using the classic arithmetic coder, i.e. AC_1. Figure 4.8 shows, as a function of a, the fraction of strings of length $M = 20$ exceeding the threshold $M \cdot a$ when AC_{q^*} and AC_1 are applied, over Wikipedia and simulated data (i.i.d. symbols). It can be noted that, by generalizing the encoding procedure, the number of codewords exceeding the threshold can be decreased significantly, especially for 'large' a. Such a drop is more pro-

nounced in the case of the Wikipedia data. The reason is that, since there is an abundance of 'rare' strings in the real data (as we already discussed), the encoding strategy with escort distribution, which penalizes frequent symbols in favor of rare ones, is more efficient than it is for truly i.i.d. symbols.

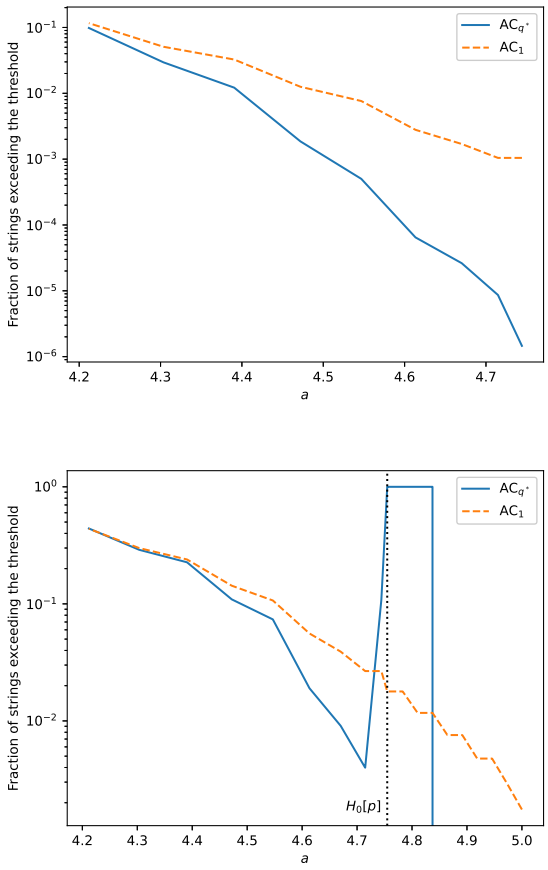


Figure 4.8: Fraction of strings exceeding the threshold for Wikipedia data (top panel) and for the simulated i.i.d. symbols (bottom panel). The orange dashed line is obtained with the classic arithmetic coder AC_1, while the solid blue line with AC_{q^*}. In the second panel, the spike before the plateau comes from the fact that an error $2/M = 2/20 = 0.1$ occurs in AC_q procedure. Indeed, the plateau's width is approximately 0.1.

Moreover, we also want to explain the presence of the spike, followed by a short plateau, in the fraction of strings exceeding the threshold shown in the bottom panel of Figure 4.8, occurring for $a \gtrsim H_0[p]$. It is caused by the intrinsic 2-bits error of AC_{q^*} procedure (see Eqn. (4.16)). In fact, if $a = H_0[p] \approx 4.75$, then $M \cdot a \approx 95$. If we could exactly reach the

desired symbols' length of Eqn. (4.6) with $q^* = 0$, we would never exceed the threshold. But AC_{q^*} carries an intrinsic error: the encoded strings' lengths are all 97 bits, in accordance with the predicted AC_{q^*} error. The fraction of strings exceeding the threshold is then 1 until a becomes such that $M \cdot a = 97$, i.e. $a = 4.85$. After such value, the exceeding fraction drops to 0. In other words, when the threshold is close to $H_0[p]$, even the very small error of the Arithmetic Coding procedure can lead to exceed it. Despite that, the ensemble of such cases is very small with respect to all the possibilities: for every $a \in (4.12, 4.75)$ the AC_{q^*} procedure performs better than the usual AC_1, both for real (correlated) symbols and simulated (independent) ones.

4.4.3 A note on the estimation of the source probability distribution

So far, we have considered that the probability p of the source generating i.i.d. symbols is known to the encoder. In reality, this could not be the case and a measure of error is needed if the probability $r = \{r_1, \ldots, r_N\}$ is used to encode symbols generated by the probability p. In the classical case, this is a well known problem. Assuming that it is possible to achieve the best encoding length which minimize the average length $L(0)$, i.e. $\ell_D(i) = -\log p_i$, then if the probability r is practically used to encode symbols generated according to p, the average codewords length is simply given by

$$H_1[p||r] = -\sum_{i=1}^{N} p_i \log_D r_i. \tag{4.34}$$

$H_1[p||r]$ is called cross-entropy. From this, it is possible to define the number of bits that are wasted by encoding according to r as the difference between the cross-entropy (i.e. the actual average length) and the Shannon entropy (i.e. the lowest possible average length), thus getting the Kullback-Leibler divergence:

$$D_{KL}[p||r] = H_1[p||r] - H_1[p] = \sum_{i=1}^{N} p_i \log \frac{p_i}{r_i}. \tag{4.35}$$

Following the same path, we would like to provide a measure in the case of an exponential average. While Rényi himself defined a generalized D_{KL} [15], further analyzed in [122] and [123], and different definitions of a generalized cross-entropy exist [124], we would like to define such quantities in the framework of data compression. In particular, the exponential average codeword length when r is used to perform the compression is given by:

$$\begin{aligned} H_q[p||r] &= \frac{1}{t}\log_D \sum_{i=1}^{N} p_i D^{-t\log_D(r_i^{(q)})} \\ &= \frac{q}{1-q}\log_D \sum_{i=1}^{N} p_i r_i^{q-1} + (1-q)H_q[r], \end{aligned} \tag{4.36}$$

where $r^{(q)}$ is the escort distribution of r, $q = 1/(1+t)$ and $H_q[r]$ is the Rényi entropy of the distribution r. From this definition, it is possible to write a function for the error of encoding with distribution r instead of the true p, as the difference between the actual exponential average length $H_q[p||r]$, and the lowest possible exponential average length $H_q[p]$, that would be obtained by the exact guessing of p, i.e. with $r = p$:

$$\begin{aligned} \mathrm{ER}_q[p||r] &= H_q[p||r] - H_q[p] \\ &= \frac{q}{1-q}\log_D \sum_{i=1}^{N} p_i r_i^{q-1} + (1-q)H_q[r] - H_q[p]. \end{aligned} \tag{4.37}$$

It is easy to see that $\mathrm{ER}_q[p||p] = 0\ \forall q > 0$ and that $\lim_{q\to 1}\mathrm{ER}_q[p||r] = D_{KL}[p||r]$. Figure 4.9 shows the error function for varying q, with given p and r such that $p_i \propto i^{-1}$ and $r_i \propto i^{-2}$.

To our knowledge, despite the different definitions of generalized divergences and cross-entropies in the literature, the quantity ER_q has not been defined. Yet, it has a direct interpretation and provides a measure of how a wrong estimate of the probability p propagates on the exponential average codeword length $L(t)$.

4.5 Conclusions

In this chapter, we have provided an operational scheme to encode sequences of symbols in order to minimize the exponential average codeword length.

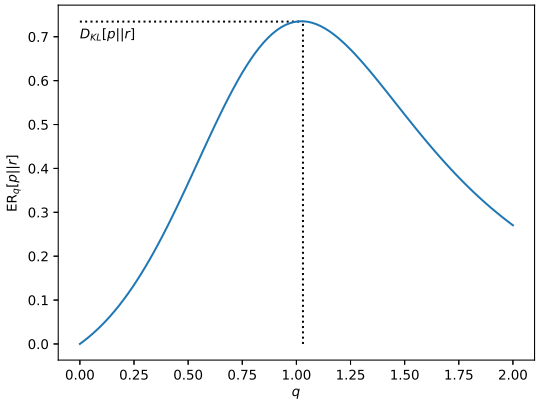


Figure 4.9: Error $\mathrm{ER}_q[p||r]$ for two instances of p and r: $p_i \propto i^{-1}$ and $r_i \propto i^{-2}$. The horizontal dashed black line corresponds to $D_{KL}[p||r]$, which corresponds to $\mathrm{ER}_1[p||r]$.

Our algorithm leads to an exponential average length per symbol that is arbitrarily close to the Rényi entropy of the source distribution. While our theoretical analysis relies on the symbols being i.i.d., we have shown that it provides advantageous results even in the case of correlated symbols, with respect to the usual $q = 1$ Shannonian compressor. Moreover, we have detailed a possible application of the exponential average, based on its connection with the cumulant generating function of the source's probability distribution. Namely, if the encoder's priority is to minimize the risk of exceeding a certain codewords' threshold length, minimizing the exponential average is a better solution than minimizing the linear average. Even if all our theoretical considerations are based on the hypothesis that the symbols are i.i.d. distributed and that the encoder knows the true source distribution p, we have both shown empirically that AC_q is advantageous also in the presence of correlations and provided a measure of the error when the encoder guesses the incorrect source distribution. However, a theoretical description explaining quantitatively how correlations lead to an optimal q different from q_t, and what is the expected error of guessing the source distribution given a certain ("small") training dataset still lacks and can be the object of future studies.

Chapter 5

Conclusions

Throughout this work, I have investigated some of the challenges posed by complex systems to the domain of data science. In particular, I have considered how the strong dependencies between the constituents of complex systems and the broad distributions that characterize their structural and dynamical properties require different perspectives in the analysis, inference and compression of data.

Specifically, within the realm of opinion dynamics, I presented a multiplex opinion dynamics model to investigate the behavior of users interacting across multiple social media platforms. My findings underscore that studying the algorithmic effects of recommendation systems on political polarization based solely on single-platform considerations can lead to skewed interpretations. In fact, by considering two platforms, I have shown that opinion polarization can be sustained by the presence of a polarizing platform even if users spend the majority of their time on a neutral one. Furthermore, introducing a dynamical social media repertoire update in the model revealed a notable segregation between extreme and moderate users. Interestingly, in this environment, the competition between platforms morphs into cooperation, promoting greater user satisfaction without amplifying opinion polarization.

Then, in the context of theoretical methods to handling partial information, I have introduced the Generalized Maximum Entropy Principle (GMEP). By maximizing generalized entropies that are parameter-dependent, it offers the possibility not to a priori neglect unobserved correlations among system components. Moreover, though the uninformativeness axiom, it is possible to pinpoint a specific generalized entropy from a given family. If applied to the Uffink-Jizba-Korbel or Hanel-Thurner entropies,

this axiom selects the Rényi entropy. Furthermore, it reconciles the GMEP with the Maximum Likelihood (ML) principle, leading to the natural emergence of the q-average. The latter can be now be interpreted as the natural average that maximizes the likelihood when there's only a single observation about the system. Furthermore, for i.i.d. observations, I have shown that the entropic parameter q can be determined via ML. Interestingly, its numerical value ensures that the negative log-likelihood equals the Shannon entropy, even if the observed distribution maximizes Rényi entropy rather than Shannon entropy.

Finally, I have extended the established Arithmetic Coding (AC) scheme to AC_q, which takes escort distributions as input probabilities. Notably, AC_q effectively approaches the lower bound of the exponential average codeword length—equivalent to the Rényi entropy of the source distribution—with vanishing error. When applied to both simulated i.i.d. strings and real English Wikipedia text, the robustness of AC_q becomes evident. Despite Campbell's theoretical underpinnings rely on the hypothesis of i.i.d. symbols, AC_q provides remarkable results in terms of exponential average length even when symbols are correlated. Additionally, my research has underscored that the exponential average is the quantity to minimize if the objective is to reduce the likelihood of exceeding a given codewords' length threshold. The application on both i.i.d. and real-world data confirm my analytical findings. In fact, by applying AC_q, there is a noticeable reduction in the count of strings that exceed a given threshold, compared to traditional Shannonian encoding scheme.

To conclude, my research has underscored the pitfalls of drawing conclusions from partial information—often leading to misinterpretations or superficial insights. The main tool for navigating such challenges, and ensuring unbiasedness, is the maximum entropy principle. I've explored this realm, particularly focusing on reconciling the GMEP with the ML principle. Further, I've investigated the estimation of the parameter q, the indicator of system correlations in the context of GMEP. The emergence of Rényi entropy from the uninformativeness axiom, led me to probe its application in the field of data compression. The intricate nature of certain human-engineered systems, exemplified by DNA coding or applications where memory optimization is paramount, can lead to pose emphasis on the exponential average codeword length, rather than the linear one. The

introduction of AC_q, coupled with insights into its capacity to reduce the number of losses in memory-constrained environments, offers a promising avenue for enhancing such complex systems.

Bibliography

[1] James Ladyman, James Lambert, and Karoline Wiesner. What is a complex system? *European Journal for Philosophy of Science*, 3:33–67, 2013.

[2] Stefano Battiston, J Doyne Farmer, Andreas Flache, Diego Garlaschelli, Andrew G Haldane, Hans Heesterbeek, Cars Hommes, Carlo Jaeger, Robert May, and Marten Scheffer. Complexity theory and financial regulation. *Science*, 351(6275):818–819, 2016.

[3] Didier Sornette. Physics and financial economics (1776–2014): puzzles, ising and agent-based models. *Reports on progress in physics*, 77(6):062001, 2014.

[4] Simon Levin, Tasos Xepapadeas, Anne-Sophie Crépin, Jon Norberg, Aart De Zeeuw, Carl Folke, Terry Hughes, Kenneth Arrow, Scott Barrett, Gretchen Daily, et al. Social-ecological systems as complex adaptive systems: modeling and policy implications. *Environment and Development Economics*, 18(2):111–132, 2013.

[5] Wil Van Der Aalst and Wil van der Aalst. *Data science in action*. Springer, 2016.

[6] Tiziano Squartini, Guido Caldarelli, Giulio Cimini, Andrea Gabrielli, and Diego Garlaschelli. Reconstruction methods for networks: The case of economic and financial systems. *Physics reports*, 757:1–47, 2018.

[7] Brendan Nyhan, Jaime Settle, Emily Thorson, Magdalena Wojcieszak, Pablo Barberá, Annie Y Chen, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, et al. Like-minded sources on facebook are prevalent but not polarizing. *Nature*, pages 1–8, 2023.

[8] Edwin T Jaynes. *Physical review*, 106(4):620, 1957.

[9] John Shore and Rodney Johnson. *IEEE Transactions on information theory*, 26(1):26–37, 1980.

[10] Jos Uffink. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 26(3):223–261, 1995.

[11] Tiziano Squartini and Diego Garlaschelli. *Maximum-Entropy Networks: Pattern Detection, Network Reconstruction and Graph Combinatorics*. Springer, 2017.

[12] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[13] AI Khinchin. Mathematical foundations of information theory dover publications inc. *New York*, 1957.

[14] Petr Jizba and Jan Korbel. Maximum entropy principle in statistical inference: case for non-shannonian entropies. *Physical review letters*, 122(12):120601, 2019.

[15] Alfréd Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 547–561. University of California Press, 1961.

[16] Constantino Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics*, 52(1):479–487, 1988.

[17] Guido Bellomo, Gustavo M Bosyk, Federico Holik, and Steeve Zozor. Lossless quantum data compression with exponential penalization: an operational interpretation of the quantum rényi entropy. *Scientific Reports*, 7(1):14765, 2017.

[18] Linda C Meiser, Bichlien H Nguyen, Yuan-Jyue Chen, Jeff Nivala, Karin Strauss, Luis Ceze, and Robert N Grass. Synthetic dna applications in information technology. *Nature communications*, 13(1):352, 2022.

[19] Pooja Mishra, Chiranjeev Bhaya, Arup Kumar Pal, and Abhay Kumar Singh. Compressed dna coding using minimum variance huffman tree. *IEEE Communications Letters*, 24(8):1602–1606, 2020.

[20] Frederick Jelinek. Buffer overflow in variable length coding of fixed rate sources. *IEEE Transactions on Information Theory*, 14(3):490–501, 1968.

[21] Andrea Somazzi, Giuseppe Maria Ferro, Diego Garlaschelli, and Simon Asher Levin. Social media battle for attention: opinion dynamics on competing networks. *arXiv preprint arXiv:2310.18309*, 2023.

[22] Andrea Somazzi and Diego Garlaschelli. Learn your entropy from informative data: an axiom ensuring the consistent identification of generalized entropies. *arXiv preprint arXiv:2301.05660*, 2023.

[23] Andrea Somazzi, Paolo Ferragina, and Diego Garlaschelli. On nonlinear compression costs: when shannon meets rényi. *arXiv preprint arXiv:2310.18419*, 2023.

[24] Robert M Bond, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295–298, 2012.

[25] Elisa Shearer and Elizabeth Grieco. Americans are wary of the role social media sites play in delivering the news. *Pew Research Center*, 2, 2019.

[26] Pablo Barberá, John T Jost, Jonathan Nagler, Joshua A Tucker, and Richard Bonneau. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*, 26(10):1531–1542, 2015.

[27] Cass R Sunstein. Is social media good or bad for democracy. *SUR-Int'l J. on Hum Rts.*, 15:83, 2018.

[28] Andrew Guess, Brendan Nyhan, Benjamin Lyons, and Jason Reifler. Avoiding the echo chamber about echo chambers. *Knight Foundation*, 2(1):1–25, 2018.

[29] Eytan Bakshy, Solomon Messing, and Lada A Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.

[30] Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. The spreading of misinformation online. *Proceedings of the national academy of Sciences*, 113(3):554–559, 2016.

[31] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 world wide web conference*, pages 913–922, 2018.

[32] Ferenc Huszár, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. Algorithmic amplification of politics on twitter. *Proceedings of the National Academy of Sciences*, 119(1):e2025334119, 2022.

[33] Ro'ee Levy. Social media, news consumption, and polarization: Evidence from a field experiment. *American economic review*, 111(3):831–870, 2021.

[34] Petter Holme and Mark EJ Newman. Nonequilibrium phase transition in the coevolution of networks and opinions. *Physical Review E*, 74(5):056108, 2006.

[35] Xin Wang, Antonio D Sirianni, Shaoting Tang, Zhiming Zheng, and Feng Fu. Public discourse and social network echo chambers driven by socio-cognitive biases. *Physical Review X*, 10(4):041042, 2020.

[36] Fernando P Santos, Yphtach Lelkes, and Simon A Levin. Link recommendation algorithms and dynamics of polarization in online social networks. *Proceedings of the National Academy of Sciences*, 118(50):e2102141118, 2021.

[37] Fabian Baumann, Philipp Lorenz-Spreen, Igor M Sokolov, and Michele Starnini. Modeling echo chambers and polarization dynamics in social networks. *Physical Review Letters*, 124(4):048301, 2020.

[38] William Hart, Dolores Albarracín, Alice H Eagly, Inge Brechan, Matthew J Lindberg, and Lisa Merrill. Feeling validated versus being correct: a meta-analysis of selective exposure to information. *Psychological bulletin*, 135(4):555, 2009.

[39] Uwe Hasebrink and Jutta Popp. Media repertoires as a result of selective media use. a conceptual approach to the analysis of patterns of exposure. *Communications*, 31(3):369—387, 2006.

[40] Dolf Zillmann. Mood management in the context of selective exposure theory. *Annals of the International Communication Association*, 23(1):103–123, 2000.

[41] Alan M Rubin. The uses-and-gratifications perspective of media effects. In *Media effects*, pages 535–558. Routledge, 2002.

[42] Edson C Tandoc Jr, Chen Lou, and Velyn Lee Hui Min. Platform-swinging in a poly-social-media context: How and why users navigate multiple social media platforms. *Journal of Computer-Mediated Communication*, 24(1):21–35, 2019.

[43] R Lance Holbert and William L Benoit. A theory of political campaign media connectedness. *Communication Monographs*, 76(3):303–332, 2009.

[44] Elaine Yuan. News consumption across multiple media platforms: A repertoire approach. *Information, communication & society*, 14(7):998–1016, 2011.

[45] Uwe Hasebrink and Hanna Domeyer. Media repertoires as patterns of behaviour and as meaningful practices: A multimethod approach to media use in converging media environments. *Participations*, 9(2):757–779, 2012.

[46] Fabian Baumann, Philipp Lorenz-Spreen, Igor M. Sokolov, and Michele Starnini. Emergence of polarized ideological opinions in multidimensional topic spaces. *Physical Review X*, 11, 1 2021.

[47] Alexander J Stewart, Joshua B Plotkin, and Nolan McCarty. Inequality, identity, and partisanship: How redistribution can stem the tide of

mass polarization. *Proceedings of the National Academy of Sciences*, 118(50):e2102140118, 2021.

[48] Eva Jonas, Stefan Schulz-Hardt, Dieter Frey, and Norman Thelen. Confirmation bias in sequential information search after preliminary decisions: an expansion of dissonance theoretical research on selective exposure to information. *Journal of personality and social psychology*, 80(4):557, 2001.

[49] John V Duca and Jason L Saving. Income inequality and political polarization: Time series evidence over nine decades. *Review of Income and Wealth*, 62(3):445–466, 2016.

[50] Sam Bestvater, Sono Shah, Gonzalo River, and Aaron Smith. Politics on twitter: One-third of tweets from us adults are political. *Pew Research Center*, 2022.

[51] J Liedke and KE Matsa. Social media and news fact sheet. *Pew Research Center's Journalism Project*, 2022.

[52] Marilynn B Brewer. The social self: On being the same and different at the same time. *Personality and social psychology bulletin*, 17(5):475–482, 1991.

[53] Rainer Hegselmann and Ulrich Krause. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artifical Societies and Social Simulation (JASSS) vol*, 5(3), 2002.

[54] Mirko Lai, Marcella Tambuscio, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. Stance polarity in political debates: A diachronic perspective of network homophily and conversations on twitter. *Data & Knowledge Engineering*, 124:101738, 2019.

[55] Markus Prior. Media and political polarization. *Annual Review of Political Science*, 16:101–127, 2013.

[56] Christopher A Bail, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221, 2018.

[57] Yotam Shmargad and Samara Klar. How partisan online environments shape communication with political outgroups. *International Journal of Communication*, 13:27, 2019.

[58] Elizabeth Levy Paluck. Is it better not to talk? group polarization, extended contact, and perspective taking in eastern democratic republic of congo. *Personality and Social Psychology Bulletin*, 36(9):1170–1185, 2010.

[59] We Are Social, DataReportal, and Hootsuite. (january 26, 2023). *Daily time spent on social networking by internet users worldwide from 2012 to 2023 (in minutes) [Graph]. In Statista. Retrieved October*, 23, 2023.

[60] Robin Burke. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 12:331–370, 2002.

[61] Delia Baldassarri and Andrew Gelman. Partisans without constraint: Political polarization and trends in american public opinion. *American Journal of Sociology*, 114(2):408–446, 2008.

[62] Daniel DellaPosta, Yongren Shi, and Michael Macy. Why do liberals drink lattes? *American Journal of Sociology*, 120(5):1473–1511, 2015.

[63] Nicola Perra, Bruno Gonçalves, Romualdo Pastor-Satorras, and Alessandro Vespignani. Activity driven modeling of time varying networks. *Scientific reports*, 2(1):469, 2012.

[64] Olivia Jessica Chu. *Heterogeneity in Human Populations, from Structure to Personality—A Modeling and Data Approach*. Princeton University, 2021.

[65] Ulrich Krause et al. A discrete nonlinear and non-autonomous model of consensus formation. *Communications in difference equations*, 2000:227–236, 2000.

[66] Bimal Viswanath, Alan Mislove, Meeyoung Cha, and Krishna P Gummadi. On the evolution of user interaction in facebook. In *Proceedings of the 2nd ACM workshop on Online social networks*, pages 37–42, 2009.

[67] Trevor Diehl, Matthew Barnidge, and Homero Gil de Zuniga. Multi-platform news use and political participation across age groups: Toward a valid metric of platform diversity and its effects. *Journalism & Mass Communication Quarterly*, 96(2):428–451, 2019.

[68] Rudolf Clausius. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 12(79):241–265, 1856.

[69] Ludwig Boltzmann. *Über die Beziehung zwischen dem zweiten Hauptsatze des mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung, respective den Sätzen über das Wärmegleichgewicht*. Kk Hof-und Staatsdruckerei, 1877.

[70] J Willard Gibbs. *Elementary principles in statistical mechanics*. Courier Corporation, 2014.

[71] Kevin P Murphy. *Probabilistic machine learning: an introduction*. MIT press, 2022.

[72] Constantino Tsallis. *Introduction to nonextensive statistical mechanics: approaching a complex world*. Springer Science & Business Media, 2009.

[73] Stefan Thurner, Rudolf Hanel, and Peter Klimek. *Introduction to the theory of complex systems*. Oxford University Press, 2018.

[74] José M Amigó, Sámuel G Balogh, and Sergio Hernández. A brief review of generalized entropies. *Entropy*, 20(11):813, 2018.

[75] António M Lopes and José A Tenreiro Machado. A review of fractional order entropies. *Entropy*, 22(12):1374, 2020.

[76] Jan Korbel, Rudolf Hanel, and Stefan Thurner. Classification of complex systems by their sample-space scaling exponents. *New Journal of Physics*, 20(9):093007, 2018.

[77] Stefan Thurner, Bernat Corominas-Murtra, and Rudolf Hanel. Three faces of entropy for complex systems: Information, thermodynamics, and the maximum entropy principle. *Physical Review E*, 96(3):032124, 2017.

[78] Jawaharlal Karmeshu. *Entropy measures, maximum entropy principle and emerging applications*, volume 119. Springer Science & Business Media, 2003.

[79] Diego Garlaschelli and Maria I Loffredo. Maximum likelihood: extracting unbiased information from complex networks. *Physical Review E*, 78(1):015101, 2008.

[80] Diego Garlaschelli and Maria I Loffredo. Maximum likelihood: Extracting unbiased information from complex networks. *Physical Review E*, 78(1):015101, 2008.

[81] David R. Anderson Kenneth P. Burnham. *Model Selection and Multimodel Inference*. Springer New York, NY, 2002.

[82] Peter D Grünwald, In Jae Myung, and Mark A Pitt. *Advances in minimum description length: Theory and applications*. MIT press, 2005.

[83] Petr Jizba and Jan Korbel. When shannon and khinchin meet shore and johnson: Equivalence of information theory and statistical inference axiomatics. *Physical Review E*, 101(4):042126, 2020.

[84] Andreĭ Nikolaevich Kolmogorov and Guido Castelnuovo. *Sur la notion de la moyenne*. G. Bardi, tip. della R. Accad. dei Lincei, 1930.

[85] Mitio Nagumo. Über eine klasse der mittelwerte. In *Japanese journal of mathematics: transactions and abstracts*, volume 7, pages 71–79. The Mathematical Society of Japan, 1930.

[86] Rudolf Hanel and Stefan Thurner. A comprehensive classification of complex statistical systems and an axiomatic derivation of their entropy and distribution functions. *EPL (Europhysics Letters)*, 93(2):20006, 2011.

[87] Sámuel G Balogh, Gergely Palla, Péter Pollner, and Dániel Czégel. Generalized entropies, density of states, and non-extensivity. *Scientific reports*, 10(1):1–12, 2020.

[88] AR Plastino, HG Miller, and A Plastino. General thermostatistical formalisms based on parameterized entropic measures. *Continuum Mechanics and Thermodynamics*, 16(3):269–277, 2004.

[89] AG Bashkirov. Maximum rényi entropy principle for systems with power-law hamiltonians. *Physical review letters*, 93(13):130601, 2004.

[90] Robert Wild and Roland Wester. Rate of quantum-tunnelling reaction revealed. *Measurement*, 10:3, 2023.

[91] Cheuk-Yin Wong, Grzegorz Wilk, Leonardo JL Cirto, and Constantino Tsallis. Possible implication of a single nonextensive pt distribution for hadron production in high-energy pp collisions. In *EPJ Web of Conferences*, volume 90, page 04002. EDP Sciences, 2015.

[92] Cheuk-Yin Wong, Grzegorz Wilk, Leonardo JL Cirto, and Constantino Tsallis. From qcd-based hard-scattering to nonextensive statistical mechanical descriptions of transverse momentum spectra in high-energy p p and p p collisions. *Physical Review D*, 91(11):114027, 2015.

[93] D Brian Walton and Johann Rafelski. Equilibrium distribution of heavy quarks in fokker-planck dynamics. *Physical review letters*, 84(1):31, 2000.

[94] Airton Deppman, Eugenio Megías, and Debora P Menezes. Fractals, nonextensive statistics, and qcd. *Physical Review D*, 101(3):034019, 2020.

[95] Eugenio Megias, Airton Deppman, Roman Pasechnik, and Constantino Tsallis. Comparative study of the heavy-quark dynamics with the fokker-planck equation and the plastino-plastino equation. *arXiv preprint arXiv:2303.03819*, 2023.

[96] Petr Jizba and Gaetano Lambiase. Tsallis cosmology and its applications in dark matter physics with focus on icecube high-energy neutrino data. *The European Physical Journal C*, 82(12):1123, 2022.

[97] Josef Ludescher and Armin Bunde. Universal behavior of the interoccurrence times between losses in financial markets: Independence of the time resolution. *Phys. Rev. E*, 90:062809, Dec 2014.

[98] Chris G Antonopoulos, George Michas, Filippos Vallianatos, and Tassos Bountis. Evidence of q-exponential statistics in greek seismicity. *Physica A: Statistical Mechanics and its Applications*, 409:71–77, 2014.

[99] Mikhail I Bogachev, Airat R Kayumov, and Armin Bunde. Universal internucleotide statistics in full genomes: A footprint of the dna structure and packaging? *PLoS One*, 9(12):e112534, 2014.

[100] Annalisa Greco, Constantino Tsallis, Andrea Rapisarda, Alessandro Pluchino, Gabriele Fichera, and Loredana Contrafatto. Acoustic emissions in compression of building materials: Q-statistics enables the anticipation of the breakdown point. *The European Physical Journal Special Topics*, 229:841–849, 2020.

[101] Sergio C Vinciguerra, Annalisa Greco, Alessandro Pluchino, Andrea Rapisarda, and Constantino Tsallis. Acoustic emissions in rock deformation and failure: New insights from q-statistical analysis. *Entropy*, 25(4):701, 2023.

[102] Dimitri Marques Abramov, Constantino Tsallis, and Henrique Santos Lima. Neural complexity through a nonextensive statistical–mechanical approach of human electroencephalograms. *Scientific Reports*, 13(1):10318, 2023.

[103] Razi J Al-Azawi, Nadia MG Al-Saidi, Hamid A Jalab, Hasan Kahtan, and Rabha W Ibrahim. Efficient classification of covid-19 ct scans by using q-transform model for feature extraction. *PeerJ Computer Science*, 7:e553, 2021.

[104] Qi Zhang and Diego Garlaschelli. Strong ensemble nonequivalence in systems with local constraints. *New Journal of Physics*, 24(4):043011, 2022.

[105] Christian Beck and Friedrich Schögl. *Thermodynamics of chaotic systems: an introduction*. Number 4. Cambridge University Press, 1995.

[106] Sabir Umarov and Tsallis Constantino. *Mathematical Foundations of Nonextensive Statistical Mechanics*. World Scientific, 2022.

[107] Joseph Aczél and Zoltán Daróczy. *On measures of information and their characterizations.* Academic Press, New York, 1975.

[108] LL Campbell. Definition of entropy by means of a coding problem. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 6(2):113–118, 1966.

[109] Godfrey Harold Hardy, John Edensor Littlewood, and George Pólya. *Inequalities*. Cambridge university press, 1952.

[110] Pablo A Morales, Jan Korbel, and Fernando E Rosas. Thermodynamics of exponential kolmogorov–nagumo averages. *New Journal of Physics*, 25(7):073011, jul 2023.

[111] Pierre Humblet. Generalization of huffman coding to minimize the probability of buffer overflow (corresp.). *IEEE Transactions on Information Theory*, 27(2):230–232, 1981.

[112] Liam David and Anonto Zaman. Simulating iridium satellite coverage for cubesats in low earth orbit. *in Proc. 32nd Annu. AIAA/USU Conf. Small Satellites*, 2018.

[113] Michael B Baer. Optimal prefix codes for infinite alphabets with nonlinear costs. *IEEE Transactions on Information Theory*, 54(3):1273–1286, 2008.

[114] L Lorne Campbell. A coding theorem and rényi's entropy. *Information and control*, 8(4):423–429, 1965.

[115] Neri Merhav. Universal coding with minimum probability of codeword length overflow. *IEEE Transactions on Information Theory*, 37(3):556–563, 1991.

[116] Anselm C Blumer and Robert J McEliece. The rényi redundancy of generalized huffman codes. *IEEE Transactions on Information Theory*, 34(5):1242–1249, 1988.

[117] J-F Bercher. Source coding with escort distributions and rényi entropy bounds. *Physics Letters A*, 373(36):3235–3238, 2009.

[118] Alistair Moffat and Matthias Petri. Large-alphabet semi-static entropy coding via asymmetric numeral systems. *ACM Transactions on Information Systems (TOIS)*, 38(4):1–33, 2020.

[119] Ian H Witten, Radford M Neal, and John G Cleary. Arithmetic coding for data compression. *Communications of the ACM*, 30(6):520–540, 1987.

[120] Paolo Ferragina. *Pearls of Algorithm Engineering*. Cambridge University Press, 2023.

[121] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[122] Tim Van Erven and Peter Harremos. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

[123] Ziad Rached, Fady Alajaji, and L Lorne Campbell. Rényi's divergence and entropy rates for finite alphabet markov sources. *IEEE Transactions on Information theory*, 47(4):1553–1561, 2001.

[124] Ferenc Cole Thierrin, Fady Alajaji, and Tamás Linder. Rényi cross-entropy measures for common distributions and processes with memory. *Entropy*, 24(10):1417, 2022.

Acknowledgements

I would like to acknowledge the committee members for reading my manuscript and providing feedback.
Lo sviluppo di questa tesi non sarebbe stato possibile senza il supporto pragmatico ed emotivo di tutte le persone che mi sono state vicino.

Vorrei ringraziare di cuore il professor Diego Garlaschelli, che mi ha sempre guidato ed incoraggiato nel mio percorso di ricerca. Le stimolanti conversazioni, il tempo e le attenzioni dedicatomi sono state di incommensurabile valore per la mia crescita professionale e per lo sviluppo del mio lavoro. Lo ringrazio inoltre per essere stato un ammirevole esempio di professionalità, umanità ed empatia, qualità che ambisco a tenere con me nel futuro.
Sono grato anche al professor Paolo Ferragina, per le discussioni costruttive e la condivisione di idee ed esperienza che hanno portato allo sviluppo del paper su cui si basa il Capitolo 4.
Grazie a tutta l'Unità di Networks dell'IMT di Lucca per gli scambi di idee, il calore ed i momenti gioviali passati insieme.

Al di fuori dell'università, ci sono una serie di persone che hanno arricchito la mia vita, con cui ho condiviso riflessioni, gioie e difficoltà.
Grazie a tutti i Rubbera per il tempo speso insieme. È stato un faro in dei periodi bui. Grazie per la costante presenza, vicinanza, ascolto.
Paolo, senza di te berrei solo vino di scarsa qualità. Grazie per i momenti passati insieme a mangiare bene e bere bene.
Luca, mi ricordo ancora quando a Trieste ti insegnai a preparare la carbonara, ormai tuo cavallo di battaglia. Sei e sarai sempre il mio coinquilino preferito.
Martino, ne è passato di tempo da quando a Parigi mi hai iniziato al mondo del fitness. Grazie per per non stancarti mai di bacchettarmi quando me lo merito.

Giuseppe, compino, a te devo dire grazie su ogni piano. Grazie per avermi aiutato ad affrontare i miei momenti difficili, tramite il tuo ascolto e i tuoi consigli; mi hai aiutato a fare pace con me stesso, la tua vicinanza è stata fondamentale. Grazie per condividere con me la passione per la ricerca, per le nostre conversazioni e per il lavoro che abbiamo svolto insieme, che ha portato al paper su cui si basa il secondo Capitolo di questa tesi. Grazie per tutte le parentesi di svago che abbiamo vissuto, le chiacchere sugli anime, gli allenamenti, le partite a ping pong, l'invenzione di storie surreali.

Voglio inoltre ringraziare gli amici di sempre, capisaldi della mia vita.
Niccolò, sono più di dieci anni che per me ci sei sempre, la distanza non ha mai intaccato la nostra amicizia. Grazie per essere stato presente in ogni fase della mia vita, condividendo con me le gioie e le soddisfazioni. Il tuo ascolto attento è sempre stato prezioso, aiutandomi a riflettere su me stesso e a superarare a testa alta i momenti peggiori.
Alessandro, sei probabilmente la persona con cui abbia passato più tempo a coltivare i miei hobby. La tua empatia e delicatezza, che ti hanno sempre caratterizzato, sono qualità magnifiche da trovare in un amico. So che su di te posso sempre contare.
Carlotta, dopo ventitre anni di amicizia è un eufemismo dire che per me sei come un familiare. Nonostante la distanza che ci separa, ogni volta che ci rivediamo è come se fosse passato un attimo dalla precedente.

Per ultimo, grazie a tutta la mia famiglia, che non mi ha fatto mai scordare dove fosse *casa*.
Nonna Luciana, grazie per la leggerezza e la tranquillità che mi offri.
Nonno Luciano, grazie per la tua sensibilità, silenziosa empatia e rumorosa cura.
Zia Roberta, grazie per l'amore, la comprensione e la sensibilità.
Zio Angelo, grazie per l'affetto, l'interesse e ed supporto.
Michela, sorellina, grazie per la protezione.
Mamma e Papà, grazie per la libertà.