



UNIVERSITÀ DEGLI STUDI DI MILANO
FACOLTÀ DI STUDI UMANISTICI



Ph.D. Thesis

LEARNING, FORECASTING AND CAUSATION AS
COMPRESSION.

Advisor:
Hykel Hosni

Author:
Francesco Facciuto

June 2023

Abstract

Two main guidelines can be adopted to investigate a system and forecast its future behavior. The first kind of strategy emphasizes the role of general principles, which guide us in building models that embody the background knowledge available. The second class of techniques refers to phenomena ruled by unknown laws and directly probed by data-driven protocols. While the *Scientific Method* encodes the first kind of procedure, *Data Science* embraces more inductive schemes. In the last twenty years, many scholars have developed growing expectations about the impact of the latter family of methods, and the role of *inductivism* seems to be re-evaluated. This enthusiasm is - to some extent - justified. Despite the numerous successes achieved by model-driven science, many systems seem to resist being understood through a modeling approach. Conversely, valuable advances in performing practical tasks have been obtained by adopting *Machine Learning* and *Pattern Recognition* techniques. While a detailed analysis of these protocols' performances is beyond the scope of this discussion, some methodological aspects can be considered on a conceptual level, keeping in the background the possibility of proceeding with formally rigorous arguments. The comparison between modeling and automatic methods will allow for a better framing of the role of *intelligence* in their respective fields of application.

Research Plan. This dissertation starts by examining the properties of a generic database based on simple results from Dynamical Systems. The Analogs Method is considered an archetypal case to illustrate the consequences of an inductivist approach from a broader viewpoint. In this regard, Statistical Learning theory will provide some essential ingredients, allowing us to reformulate the *Principle of Induction*. This perspective will be elaborated more by using informational language. The *compression-generalization* trade-off is adopted as a general paradigm, discussing the effectiveness of Deep Learning protocols and drawing parallelisms with coarse-graining procedures. The analysis proposed will naturally lead us to re-read causality as a tool to manage information when some distribution shift comes into play. Moreover, some formal ways of characterizing cause-effect relations will be critically examined, and the potential connections that alternative frameworks may have will be explored to establish a more unified viewpoint. As we critically discuss the challenges inherent in inductive protocols, we aim to shed light on the following general questions. *To what extent is mathematical modeling still a necessary pursuit? What role could be played by Artificial Intelligence in this regard?*

Sommario

Due linee guida principali possono essere adottate per investigare un sistema e prevederne il comportamento futuro. Il primo tipo di strategia enfatizza il ruolo dei principi generali, guidandoci nella costruzione di modelli che incorporano le conoscenze di base disponibili. La seconda classe di tecniche si riferisce a fenomeni regolati da leggi non conosciute ed esplorati mediante protocolli basati su dati. Mentre il *Metodo Scientifico* guida il primo tipo di procedura, la *Scienza dei Dati* abbraccia schemi più induttivi. Negli ultimi vent'anni, molti studiosi hanno coltivato aspettative crescenti sull'impatto di quest'ultima famiglia di metodi e il ruolo dell'*induttivismo* sembra essere stato rivalutato. Questo entusiasmo è in parte giustificato. Nonostante i numerosi successi ottenuti dalla scienza model-based, molti sistemi manifestano una particolare resistenza all'essere compresi mediante un approccio di modellizzazione. Al contrario, sono stati ottenuti progressi significativi nell'esecuzione di compiti pratici adottando tecniche di *Apprendimento Automatico* e *Pattern Recognition*. Mentre un'analisi dettagliata delle performance di questi protocolli si pone al di là della presente discussione, analizzeremo alcuni aspetti metodologici ad un livello concettuale, lasciando sullo sfondo la possibilità di procedere con argomentazioni formalmente rigorose. Il confronto tra i metodi di modellizzazione e i metodi automatici permetterà di inquadrare meglio il ruolo dell'*intelligenza* nei rispettivi campi di applicazione.

Piano di ricerca. Questa dissertazione inizia esaminando le proprietà di un database generico adoperando alcuni semplici risultati nell'ambito dei Sistemi Dinamici. Il Metodo degli Analoghi viene considerato un caso archetipico per illustrare le conseguenze di un approccio induttivo. A tal proposito, la teoria dello Statistical Learning fornirà alcuni ingredienti essenziali, consentendoci di discutere una particolare formulazione del *Principio di Induzione*. Questa prospettiva sarà elaborata ulteriormente utilizzando il linguaggio della Teoria dell'Informazione. Il compromesso tra *compressione* e *generalizzazione* viene adottato come paradigma generale, discutendo l'efficacia dei protocolli di Deep Learning e tracciando parallelismi con le procedure di coarse-graining. L'analisi proposta ci porterà naturalmente a reinterpretare la causalità come uno strumento per gestire l'informazione quando si verifica uno shift distribuzionale. Inoltre, analizzeremo alcune caratterizzazioni formali per le relazioni causa-effetto e ne esploreremo le potenziali connessioni. Attraverso una discussione critica delle sfide intrinseche ai protocolli induttivi, ci proponiamo di fare luce sulle seguenti domande generali. *In che misura la modellizzazione matematica è ancora necessaria? Quale ruolo potrebbe svolgere l'Intelligenza Artificiale a tal riguardo?*

Contents

I	The overall picture	5
1	Mathematical Models	6
2	Machine Learning Protocols	13
3	Causality	24
II	Forecasting and Learning	31
1	Predictability	32
2	The Analogs Method	37
3	Interlude. The Relevance Objection.	44
4	Representation and Learning	49
III	Causality	77
1	Mathematical Framework	78
2	Causal Pipeline	83
3	Cause-Effect Pairs	87
4	State of the Art	99
5	Causality with time	107

Part I

The overall picture

This introduction aims at providing a comprehensive epistemological framework to contextualize the analysis presented in the following pages. Some preliminary connections between *prediction*, *information*, and *causation* will be explored as we proceed. Remarkably, the concept of *compression* will play a central role. Starting from the discussion of how scientific practice can be conceived, the problems of *forecasting* and *learning* will be reconsidered in light of some elementary results. An overall - albeit preliminary - viewpoint on the notion of *intelligence* will follow with its consequent effects in terms of the modeling-inductivism pair.

1 Mathematical Models

From Galileo onwards, observations are employed to grasp general principles, select relevant variables, and create models via mathematical equations to exhibit testable forecasts. Modeling and its applications through the *Scientific Method* have enabled breakthroughs that are difficult to overestimate, providing accurate predictions from microphysics to the cosmological scale. Due to these achievements, the idea that Physics can serve as a reference for other scientific disciplines has become increasingly prevalent, with some consequences at least at two analysis levels.

- Pragmatically - in the hope of obtaining similar performances in terms of reliability and predictive power - many efforts have been made in establishing an *isomorphism principle* between physical and non-physical entities to import into other areas of study the techniques and results that have been developed in Physics [13]. This perspective does not necessarily require any ontological commitment and can be based on the correspondence between the formal structures used. Looking at the past century only, *Social Physics* provides an explicit example of this approach, with further impacts spanning across disciplines from Social Sciences to Linguistics [174][175][197][198].
- Theoretically, the *reductionist program* posits the possibility of explaining all phenomena in terms of fundamental physical laws, thereby offering a unifying perspective of great significance [153]. Paradigmatic examples in Physics include the relationship between Thermodynamics and Statistical Mechanics [109][164] or between Classical Mechanics and the theory of Relativity. More generally, reductionist paradigms can be found in numerous disciplines - even in approaching long-standing philosophical problems - from Economics [192][65][66] to Cognitive Sciences [39][38]. Despite the practical difficulties in implementing this viewpoint, its cultural influence remains robust in various fields of inquiry, sometimes assuming the status of dogma. As a corollary, the *unity of nature* should have its counterpart in the *unity of science*, also in terms of the methodology adopted. If there was no need to admit independent levels of reality, there should be no independent theories but only more or less detailed descriptions. Therefore, assuming that the more detailed level is well-founded, we should be able to justify the techniques and the results at the higher ones hierarchically.

Models as Intelligent Representations

Both of the attitudes mentioned above are put into practice by employing mathematical models as abstract and simplified *representations*, which provide the basis for new conjectures to be submitted to additional investigations. In the first instance, mathematical models can be seen as the result of *intelligent agent processing*, through which observations and experimental outcomes are considered in light of formal frameworks compatible with the model-maker *cognitive bounds* [84][85]. More precisely, the model-maker solves a trade-off problem by providing enough accurate

representations that are also sufficiently simple to be concretely useful with respect to the task at hand [84][85][191]. Consequently, quantitative models can be available only if phenomena are weakly dependent on many microscale or context-specific details. Moreover, the ability to distinguish relevant from irrelevant details in the absence of a predetermined recipe - i.e. trying to employ the limited cognitive resources to achieve goals optimally - comes into play at least in two directions.

- Horizontally, different and independent research areas can communicate through models, allowing for the *establishment of analogies* between seemingly disparate phenomena. For example, concepts like energy, relaxation, or phase transition can admit analogous in neurophysiological studies, suggesting using similar models to characterize magnets and brain activity [4][17], thus omitting the specific details that make the two scopes distinct. Consequently, progress in one field of research can benefit the other, implying advantages in the face of the progressive specialization in scientific research and establishing disciplinary connections that contribute to frame strengths or possible limitations.
- Vertically, modeling via *coarse-graining* procedures makes the hierarchy of description levels a well-defined line of research [153][30]. In a nutshell, too detailed models can be unproductive for forecasting or inefficient in representing a phenomenon at a given spatiotemporal scale. For example, the Navier-Stokes equations emerge from complex underlying microscopic interactions. If the macroscopic fluid motion strongly depended on the shape of the constituent molecules or their detailed behavior, compact continuum laws for Fluid Mechanics would not be achievable. Contrarily, we can often move to a *more compressed* level of description, considering a set of effective equations to employ regularities that may not be explicit in the *fine-grained* representation: details negligible to our purposes - as small fluctuations or fast-timescale changes - are removed via some average or aggregation operation.

In a nutshell, *relevance criteria* are significant in selecting adequate analogies and adopting coarse-graining procedures by eliminating details that are irrelevant to the selected goals. Remarkably, this scenario immediately leads to methodological difficulties, as becomes clear when we try to formalize the role played by *relevance* in scientific practice. To elaborate on this point further, we can briefly refer to two conceptual issues that we will reconsider in the following discussion.

Moving across resolution scales. The *Renormalization Group* approach - which underlies the work on critical phenomena and many other achievements in Physics - studies how a system's space is mapped onto itself through coarse-graining procedures [30]. In this regard, the technique proposed by L. Kadanoff for spin lattices offers a clear example [154]. In a nutshell, given a spin-lattice of dimension d , the system is described at a progressively coarser *resolution* by considering blocks of 2^d spins as basic units and computing the effective interactions between these spin-blocks. This coarse-graining is then combined with a rescaling step, and the procedure is repeated iteratively, providing a constructive way to integrate all the small-scale

features into their increasingly large-scale results. Remarkably, the renormalization operates at the level of models, expressing how a given model - i.e. a particular representation at a fixed scale - has to be modified when the viewpoint on the system changes. At any rate - although admitting several technical variants - the essence of the Renormalization Group recipe lies precisely in choosing some degrees of freedom among the others to *compress* the irrelevant ones up to an effective model. Nonetheless, these degrees of freedom need to be identified first, and this step can provide a substantial barrier from a technical and conceptual point of view.

Employing causality. *Causal models* refers to an area of modeling traditionally broader than Physics. They characterize the relevant relationships among the macrovariables of interest in the Social Sciences and in all those cases where evaluating some policy arises naturally [72][115][117], from Medicine [54] to Economics [66]. From an epistemological point of view, two preliminary questions arise.

- First, it is not clear the degree to which causality is a crucial factor in our comprehension of a particular research area and whether it can be reduced to more fundamental levels [190][48]. This point emerges immediately by making a rough distinction between *theoretical* and *phenomenological* sciences, as proposed by N. Cartwright [28][29]. In the former, explanations in light of a set of principles - within a functionalist scheme - should play a predominant role. In the latter, causality should tend to be a productive concept in framing how systems respond to external stimuli and *manipulation*.
- Second, The question arises as to under what conditions causation can be established. In this regard, the *Russo-Williamson Thesis* states that a causal claim can be established only if it can be established that there is a difference-making relationship between the cause and the effect, and that there is a mechanism linking the cause and the effect that is responsible for such a difference-making relationship [141][72]. The implications of this perspective are debated in relation to biomedical research [53][188] and Social Sciences [51]. Its *normative* dimension reduces the possibility of bias that might lead to inferential mistakes and helps researchers to establish more reliable causal claims.

Moreover - as we will discuss in the following pages - causal language presents concrete limitations challenging to overcome, with impacts on our ability to identify cause-effect relationships and also leading to oversimplification of complex phenomena.

Circumventing Models

In *The Ethics of Artificial Intelligence*, L. Floridi pointed out that Artificial Intelligence research aspires to develop two different scientific programs [47]. The *engineering* program aims to reproduce the *results* of intelligent biological behavior by nonbiological means; on the other hand, the *cognitive* program aims to produce the nonbiological equivalent of intelligence, that is, the *origin* of intelligent behavior.

Moreover, L. Floridi claims that the engineering successes of Artificial Intelligence follow from a clear separation with respect to the cognitive perspective:

Today, Artificial Intelligence splits effective problem-solving and proper task execution from intelligent behavior, and it is because of this split that it can relentlessly colonize the endless space of problems and tasks whenever these can be performed without understanding, awareness, acumen, sensitivity, concerns, sensations, semantics, experience, bio-incorporation, meaning, even wisdom, and any other ingredient of human intelligence. Briefly, it is precisely when we stop trying to produce human intelligence that we can successfully replace it in an increasing number of tasks [47].

An intriguing question arises as to whether this separation can also manifest itself in the realm of mathematical modeling. At this point, it may be beneficial to delve deeper into some drawbacks that frequently arise in modeling phenomena, thus motivating the introduction of inductive methods as a valuable alternative. Three general remarks can be considered as follows.

- First, many systems are difficult to model, and establishing well-defined laws can be problematic. Although this issue already arises in Physics, it becomes even more challenging in fields such as Biology, Economics, and Social Sciences. In these contexts, determining the system dimension, identifying first principles, selecting relevant variables, and controlling experimentation create difficult-to-overcome complications. Moreover, the possibility for Physics to adopt an analytical approach stems from the homogeneity, isotropy, and *nature of interactions*, which remain constant over time [177]. This condition typically allows us to write evolution equations and study their solutions analytically or via numerical techniques. However, things can be fundamentally different in other fields of inquiry. For example, the interactions between components may change over time due to dynamics, and the system's constituents may have an internal state that significantly influences the overall behavior. Additionally, isolating the system from the environment in which it is embedded may not be possible, and this last point poses a challenge in determining initial or boundary conditions [177].
- Second, the mechanisms by which expert researchers can develop effective models that incorporate the available *background knowledge* remain somewhat unclear. Consequently, significant obstacles exist when attempting to formalize the analysis and discovery processes, as *model construction is more akin to a subtle art* that demands substantial cognitive resources rather than a strictly rigorous and replicable recipe [68]. To put it differently, building models entails a creative ability that combines discernment, intuition, and the capacity to synthesize heterogeneous data into a coherent, meaningful, and straightforward structure, the efficacy of which is challenging to predict in advance. All these elements provide strong constraints in developing automated modeling methods to support researchers and decision-makers in managing and overcoming their cognitive limitations.

- Finally, how an intelligent agent performs an assigned task in a reasonable amount of time can be challenging to represent effectively starting from a fixed set of principles, and modeling limitations that emerge are more general than those recognizable at the analytical level. In this regard, a system that dynamically changes its internal structure can be considered a machine that updates its internal state as it operates. This machine can be described as an algorithm or a list of rules that prescribe how the system’s dynamics determine the internal state and interactions at the next step. However, these *look-up-tables procedures* must face prohibitive computational costs, requiring a characterization for a huge number of possible instances. In this regard, the *rule-based* and *expert-knowledge* framework led us to the Artificial Intelligence Winter in modeling systems such as vision, language, and computationally costly games. This last point suggests there may be contexts where modeling starting from some background knowledge is not necessarily the cheapest or best-performing solution, sometimes leading us down unproductive paths.

The Shortcut Perspective. The recognized obstacles may be arduous - or even impossible - to overcome: the limitations in our cognitive faculties or the inherent inefficacy of modeling may lead to a theoretical and practical *progress plateau*. Additionally, exercising informed judgments and managing risks reasonably are crucial tasks in contexts where timely decision-making is required despite the lack of a comprehensive theoretical framework and suitable models. In each case, inductive protocols may represent a possible shortcut to circumvent these difficulties and could supplement the Scientific Method. More specifically, the question at hand pertains to the efficacy of inductive protocols in overcoming the problem of working out an effective phenomena representation in the absence of a priori guidance by means of relevance criteria - e.g., without the need for integrations to facilitate coarse-graining procedures and enable the application of causal knowledge. In recent years, this perspective seems to find some confirmations. The increasing availability of large databases has been joined by investing considerable resources in developing robust and efficient inductive strategies to infer the input-output dependency from data. In other words, Machine Learning and Pattern Recognition have come into play. Many successful applications are developed using *black-box protocols*, from speech recognition [64] and natural-language processing to image classification [83][59] and playing Go [161]. This widespread trend has also involved sectors in which mathematical modeling has a well-established tradition [27][103], such as weather and climate forecasting where an interesting debate is in progress [33][90][150].

The Big-Data Scenario. One relevant point to frame the shortcut perspective concerns the role played by the *data deluge* and the specular *data-centric* approach [5][87]. Some preliminary considerations may be suggested focusing on Table 1, where a high-level viewpoint is summarized. With some degree of approximation, the amount of data available is related to the corresponding availability of a theoretical framework and the consequent ability to build suitable models. To clarify the overall picture, let us proceed by points.

Data Regime	Background	Scales	Models	Cognitive costs
Small-data	Few principles Lots of theory	Single-scale	✓	Often tractable
Intermediate	Some theory, Phenomenology	Multi-scale	✓	Sometimes tractable
Big-data	Few Knowledge	Scaling Cascade	?	Typically prohibitive

Table 1: Three possible *Data Scenarios* - in terms of *volume* and *variety* - are schematically illustrated. Machine Learning and Pattern Recognition techniques are typically employed for large databases in the absence of a satisfactory model.

- In a *small-data scenario*, we typically have access to a limited set of variables and principles, which allows for a manageable cognitive cost in representing phenomena and making inferences. Data are primarily used to provide boundary conditions or determine the coefficients of a well-known differential equation at the scale of interest. At this level, the *hypothetico-deductive* approach has often proved promising, and the greater availability of observations can require a progressive increase in model complexity.
- At the *intermediate scenario* - where multiscale phenomena can occur and some parameter values or entire terms can be missed in modeling - procedures such as coarse-graining may be necessary to bridge the gap between available data and the underlying fundamental theory. Moreover, phenomenological considerations can also be valuable in guiding the development of theories compatible with our practical needs, and *effective models* consequently come into play.
- Finally, in the *big-data scenario*, recognizing the underlying mechanisms may not be feasible due to the vast amount of available data. The cognitive cost of developing models to explain such large datasets could be prohibitively high. In this regime, Machine Learning techniques can identify patterns within the data across multiple scales and *climbing down the ladder of complexity* without necessarily requiring a comprehensive understanding of the underlying mechanisms.

In the next chapter, the discussion will be mainly concentrate on *Large-scale learning*, for which the big-data scenario is a natural prerequisite. Problems with a large-dimensional, high-entropy input will be considered. On the other hand - in the last part of this dissertation - causal structures will be re-read as tools for dealing with cases for which massive databases are typically unavailable or even useless. In this respect - on the other side - learning effective representations may be the central

goal, even given a small dataset. At this point, it might be helpful to make two general comments to delimit the scope of the subsequent analysis.

- The overall framework mentioned above may appear strongly dependent on the specific disciplinary context. In this regard, *Physics-Informed Machine Learning* is an exemplificative perspective: data-augmentation protocols adhering to conservation laws, architectures incorporating symmetries, and cost functions promoting convergence towards physically consistent solutions are employed [78][134]. This approach addresses a specific issue: namely, to ascertain the extent to which Machine Learning techniques - when appropriately endowed with internal constraints and principles of the particular discipline - may be applied within the regime demarcated by the question mark in Table 1. Interesting as it may be, we will not explore this point in what follows.
- On the other side, each research community develops strategies and guiding principles inspired by the experience gained in addressing concrete problems, so positioning at a higher analysis level may appear unproductive. Although this concern is justified, Machine Learning and Pattern Recognition demand a methodological reflection beyond the single discipline boundaries. As already highlighted, this need becomes more pressing when considering these techniques' potential pervasiveness - even in socially sensitive areas - where the availability of automated procedures can introduce unexpected criticalities in decision-making processes. Additionally, the belief that a paradigm shift in favor of inductive protocols is needed cannot be underestimated if scientific practice as a whole is considered. For these reasons, the following analysis will recall some disciplinary contexts for illustrative purposes only.

Let us conclude this paragraph by briefly reconsidering the inductivist approach from a general viewpoint, leaving the following section to discuss Machine Learning in some detail. Adapting the analysis proposed by W. Pietsch [130][131], a few methodological guidelines in contrast with the hypothetical-deductive perspective can characterize inductivism. First, scientific laws derive from observations and can be highly probable when describing a narrow range of phenomena. In other words, there is a natural aversion to general hypotheses, inherently preliminary and never entirely established. Second, background knowledge and modeling assumptions play a minor role, as empirical data frequently justify them. Conversely, laws are obtained by *varying circumstances*: knowledge improves by continually expanding the case history recorded. Lastly, inductivism establishes a hierarchy of laws, starting with simple observation statements combined to form low-level phenomenological laws and gradually building towards laws of greater abstractness. The scientific agenda is then conceived in terms of a bottom-up process.

In summary, inductivist protocols are *data-inference*, *localized*, *context-dependent*, and *bottom-up* procedures, which prioritize predictions rather than providing phenomena representations.

2 Machine Learning Protocols

The field of Machine Learning refers to the wide range of tools that can deal with data inference problems - including regression, classification, forecasting, and control - providing an algorithm designed to select a rule into a class of possible functions by adjusting a set of free parameters. Algorithms as *meta-rules* that solve learning tasks based on semantically annotated data are said to operate in a *supervised learning* mode [103][108]. Although the upcoming insights may apply to a broader range of frameworks, the domain of supervised protocols is primarily considered in favor of a more concise presentation. The starting point consists of an input-output database

$$\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\},$$

provided by an unknown - deterministic or stochastic - rule $y = f(x)$ that we want to reconstruct with some degree of approximation. Once learned, these rules are used to classify documents or images [108][7][59], predict the price of a stock [168], diagnose a disease based on a patient's medical record [160]. To achieve this goal, a predefined architecture

$$\mathcal{A}_\theta : x \mapsto y$$

is selected, with some free parameters $\theta \in \Theta$ to be determined. The general recipe in this respect consists of three steps. First of all, \mathcal{D} is randomly partitioned into two subsets - \mathcal{D}_{train} and \mathcal{D}_{test} - whose data must be distributed according to the same probability distribution. Secondly, the machine receives data from \mathcal{D}_{train} and the *training phase* consists of determining $\theta^* \in \Theta$ such that

$$\theta^* = \arg \min_{\theta \in \Theta} \mathcal{C}(\theta; \mathcal{D}_{train}), \quad (1)$$

where \mathcal{C} is a predefined *cost function*. In this regard - fixed the *learning rate* α and the initialization θ_0 - *Gradient Descent*

$$\theta_{t+1} \longleftarrow \theta_t + \alpha_t \nabla_{\theta} \mathcal{C}(\theta_t; \mathcal{D}_{train})$$

is a standard iterative algorithm to optimize \mathcal{C} . Finally, the ability of \mathcal{A}_{θ^*} in generalizing outside of \mathcal{D}_{train} is evaluated on the subset not employed in the training phase, \mathcal{D}_{test} . The main goal - both on the theoretical and application levels - is to ensure that the trained algorithm \mathcal{A}_{θ^*} with a small cost on \mathcal{D}_{train} also admits a small *test error*. In this regard, the *over-fitting* problem can come into play: if the trained algorithm is too specialized on \mathcal{D}_{train} , it may have learned the noise on the data, thus degrading the predictive performance on \mathcal{D}_{test} . To contain this general difficulty, additional steps can be taken between training and testing phases by proceeding with *fine-tuning* over some *hyper-parameters*, often making decisions based on intuitions, previous experiences, and by surrogating tests on the *cross-validation* set $\mathcal{D}_{val} \subset \mathcal{D}$, distributed in accordance with \mathcal{D} [110]. A standard technique to avoid θ too-specialized on \mathcal{D}_{train} consist of equipping the cost function with a *regularization term* [20][108], thus replacing (1) with

$$\theta^* = \arg \min_{\theta \in \Theta} \left\{ \mathcal{C}(\theta; \mathcal{D}_{train}) + \gamma \|\theta\| \right\}, \quad (2)$$

where $\gamma \geq 0$ and $\|\cdot\|$ denotes some norm. Choosing the hyperparameter γ is equivalent to selecting an upper bound for $\|\theta\|$, thus introducing a penalty that effectively constrains the magnitude of θ .

Postponing further details to the following pages, let us conclude this paragraph with two remarks. First, the question of *data preparation* arises. Inputs x are typically stored as real-value vectors. This representation is straightforward whenever the data consist of a set of homogeneous quantities, such as pixels in an image or word frequencies in a document, so the information the labels y carry hopefully admits a geometrical counterpart in the input space. In general, this representation may not be natural: some encoding procedures become needed and the corresponding input-space geometrical properties might not correlate well with labels. In what follows, we will omit further details on this point, and only the first case will be considered. Second, as briefly discussed above, the described Machine Learning procedure aligns with classical inductivism. In this regard, \mathcal{A} and \mathcal{C} are considered instrumental aids to effectively manage \mathcal{D} without any assumptions about the phenomenon under investigation; a progressive database expansion via varying circumstances should enhance generalization; the trained algorithm \mathcal{A} is typically applied within the context and for the purposes for which it was trained.

Deep Learning as a prominent example. Although there are some significant variants, only the feed-forward and multi-layered settings will be considered [7][59]. Briefly, multiple hidden layers h_1, h_2, \dots, h_L are defined so that each node of h_l has incoming edges only from nodes in the previous layer and outgoing edges only to nodes of the next one. Consequently, the model admits the functional form

$$y = h_L \circ h_{L-1} \circ \dots \circ h_1 \circ h_0(x),$$

$$h_0 := id, \quad h_l := \sigma(W_l h_{l-1} + b_l),$$

where the matrices W_l and the vectors b_l collect the weights, and σ is some non-linear activation function. As recalled above, training the network consists of learning W_l and b_l via *Gradient Descent* and *Back-propagation* in minimizing the cost \mathcal{C} . Moreover, some specialized regularization techniques for Deep Learning come into play. Instead of equipping the cost function with a regularization term as in (2), the most commonly employed optimizer is *Stochastic Gradient Descent*. The stochasticity introduced by the use of *mini-batches* helps in preventing overfitting and can also be supported by techniques such as *dropout*, *batch-normalization*, and *early stopping* [20][7][59]. At any rate, two comments are needed.

- Regarding the network's representation power, a network with a single hidden layer can theoretically compute every binary classifier and approximate any real-valued function under rather general assumptions [45][67][193]. However, the required size of the hidden layer grows exponentially in the input-space dimension, and things do not necessarily improve for a multi-layer structure [7].

- Conversely, it is still unclear how this type of protocol achieves so impressive performance, as much empirical evidence has confirmed. For example, if the classification dataset labels are randomly shuffled, the expressivity power can still empower them with a near-zero training error [193]. More in general, Deep Learning architectures are heavily over-parametrized, and the classical *bias-variance* trade-off predicts that a good generalization should not be possible [158][179][193]. This point will be discussed in some detail.

At this point, it is worth noting that Deep Learning protocols have generated high expectations and subsequent disappointments due to various theoretical analyses put forth over time. After the introduction of *Perceptron* [101] and its concrete implementation highlighting some relevant limitations [137], M. Minsky and S. Papert argued formally that a multilayer structure would not have worked in the face of the too-tricky cost-function [105], with no chances to generalize well. To delve into more detail, the learning algorithms adopted can get stuck in local minima, i.e. suboptimal solutions that are distant from the optimal one; the minima obtained for a particular training set may not generalize well to unseen data in the test set; identifying high-quality minima may necessitate extremely long convergence times, which can be impractical or inefficient. These kinds of arguments - although convincing - could not at that time be supported by empirical evidence. In recent years, this perspective has been completely reversed in practice: adequate computer performance, *Stochastic Gradient Descent*, and enough data are sufficient in many applicative contexts. On the other hand, a satisfactory theoretical understanding is not yet available [193][7].

Three general problems for ML

As part of a comparison with the modeling approach, it may be useful to recall some epistemic downsides arising when we employ a Machine Learning protocol, as much for fundamental research as for possible applications.

- The *applicability problem* concerns the difficulty of ensuring that Machine Learning algorithms are suitable for a specific task a priori. It is unclear how to predict whether a specific problem can be studied using Machine Learning protocols. The only truly effective way is through direct testing as well as a blind succession of attempts, and it can be challenging to understand how to intervene when these protocols do not show satisfactory effectiveness [194][143]. In this regard, it would be highly beneficial to have general criteria that can guide us a priori on whether a given protocol is reasonable to expect to work effectively for the problem at hand. Factors to consider in this decision should include the algorithm, the amount and type of training data used, and some parameters that characterized the phenomena observed.
- The *opacity problem* refers to the challenge of comprehending the inner workings of Machine Learning algorithms. This difficulty extends to the semantic level, as the algorithms' nature and the enormous quantities of data they handle can make it hard to pinpoint the exact factors that affect their decisions, leading

to difficulties in interpreting their outputs and verifying their precision. This problem is especially prominent in Deep Learning protocols, where we currently do not have a widely accepted explanation for how they operate [142].

- The *explainability problem* pertains to the difficulty of offering transparent justifications for why Machine Learning algorithms produce a specific output [91]. In certain contexts, such as medical diagnoses [3], legal proceedings, or public policies, this point may be particularly critical and strongly connected to the responsibility problem [112][42]. Intuitively, making progress in this direction seems to require integrating the adopted inductive protocol with some background knowledge, for example, by equipping the algorithm with a module that employs an appropriate notion of causality [148].

The scholarly discourse surrounding inductivism is extensive. While we will not delve into its various manifestations and admonitions, a reverse course will be considered by showing how specific challenges in formulating the Problem of Learning have epistemological consequences. This, in turn, necessitates a process of *conceptual design* oriented towards properly framing the potential of inductive protocols. The starting point consists of an observation recently advanced by H. Hosni and A. Vulpiani on the *problem of forecasting* in the Dynamical Systems context [68][69][70].

Dynamical Systems and Predictability

Given a deterministic system, the future state is entirely determined by the current one. However, additional conditions are necessary to ensure the practical possibility of accurate forecasting, and some obstacles make it necessary to reconsider Laplace's approach to predictability [22]. The first limitation occurs when the evolution law is perfectly known. Two key properties come into play: the stretching of infinitesimal displacements and the wide variety of alternative orbits. Although these two aspects may seem unrelated, they are tightly connected when formally characterized via *Lyapunov Exponent s* and *Sinai-Kolmogorov Entropy* [22][32][80][113]. Tiny inaccuracies about the initial state can result in significant future deviations within the time frame of interest. Furthermore, every model can be seen as a compression-procedure outcome, allowing for appropriate generalization. If the compression process provides the optimal model, the previously mentioned limitation can be re-read in terms of the initial condition incompressibility, in the sense of *Algorithmic Complexity* à la Kolmogorov [22][35][70]. Moreover, if a lossy compression takes place by removing relevant details, *Perturbation Theory* offers a dual viewpoint on the predictive capacity: slight model variations can result in significant phase-portrait changes, with severe consequences for forecasting [32][80]. These limitations might suggest looking at inductivism as an alternative.

A first limitation for inductivism. In the context of the ergodic systems, Kac Lemma estimates the expected recurrence time [32][76][80]. Traditionally, this result is invoked to provide a rigorous answer to the paradoxes that arise from a classic

theorem due to H. Poincaré [132] and contributes to the Statistical Mechanics foundation. On the other side - within the comparison between modeling and inductive approach - Kac Lemma limits the ability to make predictions inductively. To clarify this point, we consider the *Analogs Method* proposed by E.N. Lorenz [92][93][94][95]. Given the actual state z , we want to forecast its evolution after an amount of time t , say $\phi^t z$. Retrieved a past state a in our database \mathcal{D} similar to z , we approximate $\phi^t z$ with $\phi^t a$:

$$a \sim z \implies \phi^t a \sim \phi^t z,$$

$$a, \phi^t a \in \mathcal{D}.$$

Theoretically, the Analogs Method only requires a *regularity assumption*: if a system behaves in a certain way, it will do so again. It seems to be a natural claim at least for numerous practical tasks. Operationally, the algorithm that provides a forecast is straightforward and admits advantages that will be discussed later. Contrarily, Kac Lemma enables us to prove that the required time-series length L scales as

$$L \sim \frac{1}{\varepsilon^d},$$

where d is the system's dimension and ε is the *resolution* at which nearby points are identified [31][69]. Given the exponential law in d , the required \mathcal{D} size becomes prohibitive for moderately high dimensions. This limitation could temper the prospects of a purely inductive approach to forecasting. While for modeling the most significant constraints pertain to the chaotic nature of the system and the instability of the phase portrait under small perturbations, the primary obstacle for the Analogs Method lies in the system's dimension. Remarkably - as we will discuss later - the dimension involved can be also the *effective* one. However, it remains to be established to what extent this limitation is relevant for Machine Learning. In this respect, two general observations are significant:

- Although the Kac Lemma argument provides a dynamic justification for the *curse of dimensionality*, this issue is practically overcome whenever an inductive protocol achieves performances compatible with our desiderata. Consequently, a genuine understanding of the limit mentioned above should include a justification for which cases where the curse appears to dissolve.
- The Analogs Method is a simple *memorization and search procedure*, while the *generalization level is assigned a priori* via the resolution ε . This observation casts doubt on the relevance of what is derived from Kac Lemma for Machine Learning, where very different techniques come into play to deal with the generalization problem.

We will examine these points in detail, also considering some objections based on the distinction between *enumerative* and *variational* inductivism [130][131]. Let us take a first step by reconsidering the problem of inductivism within the general framework of Statistical Learning.

Statistical Learning and Information

The classical Statistical Learning theory [184][183][181] allows us to formulate the generalization problem in a mathematically precise way by returning a trade-off between the approximation accuracy provided by the algorithm \mathcal{A} from a dataset \mathcal{D} and the complexity of the approximating function f , chosen from a functional space \mathcal{F} [106]. Intuitively, we expect that as the size of the dataset increases, the approximation selected by \mathcal{A} will become more accurate. This perspective is supported by the results of *universal consistency*. However, while universal consistency tells us that everything can be learned within the limit of infinite data, it does not imply that every problem is learnable from a finite dataset with a reasonable amount of data. In practice, a consistent algorithm may not be preferred over a non-consistent one. More in particular, the phenomenon of *slow rates* may come into play, and a *No Free Lunch Theorem* states that for any learning algorithm, there are problems for which the learning rates are arbitrarily slow. The problem of supervised learning allows us to clarify this idea [183][108].

Let us consider a random generator of occurrences X distributed according to a probability distribution P_X ; a supervisor that returns an output Y according to a distribution $P_{Y|X}$; an algorithm that returns a function $f \in \mathcal{F}$, selected to approximate the supervisor's response with the highest possible accuracy. Let us assume that the distribution $P_{X,Y}$ is unknown. The algorithm takes as input a training set $\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$ consisting of n independently and identically distributed instances sampled from $P_{X,Y}$. The quality of the selected approximation f must be evaluated on instances distributed as $P_{X,Y}$, but not necessarily included in the training set. In other words, we want to ensure that f is not specialized to the available dataset \mathcal{D} , but captures the underlying regularity of the problem instead. For simplicity, we will consider the case of *binary classification*: given an occurrence for X , it is to be determined a rule that associates the correspondent label in $\mathcal{Y} = \{0, 1\}$. In other words, we seek to learn $f : \mathcal{X} \rightarrow \mathcal{Y}$ based on observations available. Given a rule f and an occurrence (x, y) , the corresponding error is defined via some metric fixed a priori, say $err[f(x), y]$. In principle, we assume that the optimal rule should minimize the *expected error*

$$R[f] := \int err[f(x), y] dP(x, y)$$

over the class of functions \mathcal{F} . Since the distribution $P_{X,Y}$ is unknown, we can not directly minimize the functional $R[f]$.

Induction Principle. To avoid the difficulty mentioned above, we consider instead of $P_{X,Y}$ the empirical counterpart \hat{P} on \mathcal{D} , which is the distribution concentrated on the available data. Through this particular choice, the functional $R[f]$ is therefore replaced by the *empirical risk*

$$\hat{R}_n[f] := \frac{1}{n} \sum_{i=1}^n err[f(x_i), y_i],$$

so that the algorithm \mathcal{A} will proceed by minimizing $\hat{R}[f]$. However, this procedure does not necessarily provide the necessary guarantees. To see this, it is sufficient to consider the following problem: if we consider as \mathcal{F} the entire space of binary functions defined on the finite range as \mathcal{X} of X , our algorithm could in principle select one of the $2^{|\mathcal{X}|-n}$ possible rules for which $\hat{R}_n[f] = 0$, without being able to establish a priori which of these best generalizes on the new instances. In other words, by minimizing the empirical risk, we could select a function in the *overfitting regime*. The classic strategy to overcome this difficulty consists of restricting the space of functions \mathcal{F} , controlling its element's complexity.

Capacity measures. By introducing a measure of complexity for the functional space adopted, it is possible to obtain a uniform control of the difference between empirical risk and expected error, establishing a convergence rate that depends on the size of the dataset. The main idea is to exploit the concentration in probability for the empirical risk around the expected error due to the *Chernoff-bound* effect. If \mathcal{F} has finite cardinality, it is possible to demonstrate that the following inequality holds for any $f \in \mathcal{F}$, with probability $1 - \delta$:

$$|R[f] - \hat{R}_n[f]| \leq \sqrt{\frac{\log(2|\mathcal{F}|/\delta)}{2n}}. \quad (3)$$

This inequality establishes a relation between *generalization-gap* $|R[f] - \hat{R}_n[f]|$ and the function complexity measured by the capacity $|\mathcal{F}|$, controlled by the dataset size n . Suppose we find a sufficiently small set of functions \mathcal{F} to achieve a small empirical risk. In that case, we can guarantee - with high probability - that those functions will also have small errors in future data from the same distribution. A non-trivial trade-off ensues. On the one hand, we would like to work with a large class of functions to allow for a small empirical risk. On the other hand, the functional class should be small to control the generalization gap. Optimizing this trade-off is a well-defined problem also for other capacity measures, e.g. when \mathcal{F} is an infinite set. At any rate, three remarks are essential in our views.

- First, the function space \mathcal{F} has a central role. This set represents the *background knowledge* a priori available and corresponds to some hypothesis of regularity about the data. The capacity measure quantifies this aspect. However, it is not necessarily the case that our knowledge of the problem is sufficient to select \mathcal{F} such that it is both relevant - i.e. containing the true function of the model - and has a limited capacity. Methodologically, this object constitutes a *spurious* element in a properly inductive scenario.
- Second - continuing along the line of the previous point - the inequality reported above is independent of the probability distribution P . The particular problem at hand has no role, giving us a *worst-case* result that does not exploit possible distribution regularities. This perspective does not seem satisfactory: we would like a bound that depends on the complexity of P , establishing some classification that distinguishes tractable distributions from intractable ones.

- Finally, on a theoretical level, the derived bound does not seem to be able to explain the results obtained from recent developments in Machine learning. As we have already said, Deep Learning models - despite the enormous number of parameters and the consequent huge \mathcal{F} -capacity - show that they can learn complex problems with much less data than the bound would predict. In many cases, the number of parameters exceeds the size of the dataset by orders of magnitude, rendering the classical control over the generalization gap meaningless.

Given the aforementioned critical issues, *Information Theory* provides significant conceptual tools for reformulating the accuracy-complexity trade-off. Traditionally, this research area is associated with the problems of coding and communication through a noisy channel [157][98]. Let us briefly focus on how information-theoretic tools are valuable in the context of learning and the following implications, leaving the task of presenting a comprehensive discussion to Part II of this dissertation.

Information bounds. Let us consider the binary classification problem for the space $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \dots \times \mathcal{X}_m$ of binary strings of length n . The set of the possible classification rules $f : \mathcal{X} \rightarrow \mathcal{Y}$ has cardinality $2^{|\mathcal{X}|} = 2^{2^m}$. Identifying the correct rule - without further assumptions - is clearly an intractable problem for large m . On the other hand, the generalization gap inequality (3) discussed above is meaningless since the right-hand term is of order $O(1)$. Instead of considering a proper subset of functions $\mathcal{F} \subset \mathcal{Y}^{\mathcal{X}}$ a priori as in the classical approach, replacing \mathcal{X} with an effective input space \mathcal{X}_{eff} can provide a considerable simplification. Let us explore this perspective by assuming P_X factorized,

$$P_X = \prod_{i=1}^m P_{X_i},$$

where P_{X_i} refers to a single bit of X and $P_{X_j} = P_{X_k}$ for all j and k . The normalized sum of independent terms concentrates - for large n - around the mean:

$$-\frac{1}{m} \log p(x_1, x_2, \dots, x_n) = -\frac{1}{m} \sum_{i=1}^m \log p(x_i) \approx H[X_k],$$

$$H[X_k] := - \sum_{x_k \in \mathcal{X}_k} p(x_k) \log p(x_k).$$

In other words, a *typicality argument* comes into play. When the input variable X is a huge vector of independent and identically distributed components the *Shannon Entropy* $H[X]$ dominates the effective size of \mathcal{X} , i.e.

$$|\mathcal{X}_{eff}| \approx 2^{mH[X_k]} = 2^{H[X]}$$

Moreover, the Asymptotic Equipartition Principle in the spirit of Statistical Physics holds when only typical patterns are considered, each with probability $2^{-H[X]}$. Given

these simple observations, the *compression-generalization* trade-off will be adopted as a general paradigm in place of the capacity measures adopted in the classical Statistical Learning approach. From a general viewpoint, more compressed representations will allow us to control the generalization gap more robustly, as will be discussed in the following three steps.

- First, we shall derive and examine a rigorous generalization-gap bound independent of the function class \mathcal{F} , but instead rely on the complexity of the input variable X , as quantified by $H[X]$. This preliminary result will be based on the characterization of the set of *non-negligible* patterns, providing a first insight into to what extent a more compressed input-space representation X_ϵ can enhance the efficacy of the inductive protocol adopted.
- In the *Large Scale Learning* regime - where the number of training examples and the protocol complexity are both huge - we will discuss a *typical-case bound* obtained by restricting the probability distributions to a suitable class. This result suggests a criterion for determining which problems can be effectively treated using an inductive protocol.
- As highlighted by N. Tishby [179][158], this approach provides some insights into explaining the functioning of Deep Learning protocols as procedures that provide *compression-in-representation*, shedding light on the overfitting problem and the benefits of adopting a layered structure [159]. Moreover, some points of contact with modeling are considered, suggesting an analogy with coarse-graining techniques in developing an effective model.

The role of compression. The analysis proposed will suggest the following general picture. On the one hand - by adopting the *Empirical risk minimization* principle - Machine Learning techniques are understood in terms of *curve-fitting* procedures. In this respect, some quantitative guarantees about the control of the generalization gap are provided as long as the hypothesis space is selected a priori. On the other hand, Deep Learning protocols seem to respond to a partially different logic by automatically implementing some coarse-graining procedure. More precisely, information irrelevant to the assigned task is progressively compressed - layer by layer - to provide coarser and coarser representations. However, the *compression-in-representation* implemented by Deep Learning protocols involves a procedure of *compression-in-resolution* through which points in suitable ϵ -spheres are progressively identified. Schematically, the learning apparatus can be represented as a *Markov chain*

$$Y \longleftrightarrow X \longrightarrow X_\epsilon \longrightarrow \hat{Y}$$

where the compressed representation X_ϵ - selected via the training phase to obtain \hat{Y} as close as possible to the correct label Y - consists of a *clusterized version* of X . More precisely, the *Bottleneck Principle* characterizes the optimal compression via the variational problem

$$\min_{P_{X_\epsilon|X}} I(X : X_\epsilon) - \beta I(X_\epsilon : Y),$$

where *mutual information* $I(X : X_\varepsilon)$ and $I(X_\varepsilon : Y)$ corresponds to compression and accuracy respectively [178]. On the other side, whenever compression-in-resolution has a central role, the ε -partitions involved in solving the variational problem must satisfy certain general conditions to work correctly and generalize well. More in particular:

- First, the ε -spheres that partition the input space \mathcal{X} must be coarse enough to admit a typicality argument and contain sufficient training instances. Conversely, suppose ε is too small. In that case, typicality does not work, and no data are contained in too many partition elements. Consequently, the label assigned to each empty ε -sphere is random, affecting the accuracy.
- At the same time, the partitions adopted must be fine enough to admit labels that are as homogeneous as possible in Y for the elements of the training set. This point requires an exponential increase in the number of the ε -spheres needed, providing us with a limitation in the spirit of that suggested by Kac Lemma. In this respect - as we will discuss later - the incompressibility of the initial data will thus remain a robust constraint in the forecasting context.

While these observations pertain to whether guarantees can be offered regarding the ability to generalize, the question arises as to within what terms the compression-in-resolution is actually activated. This point - rather subtle - can be discussed at a preliminary level by considering the role of regularization. In a nutshell, more robust regularization is equivalent to lower resolution in determining the decision bound for a binary classification problem. Classically, regularization of the cost function as in (2) is achieved by manually fixing the corresponding hyperparameter γ and thus inserting a spurious element into the inductive learning process [108]. As will be discussed in Part II, *Stochastic Gradient Descent* could make this process component implicit and automatic [16][167].

Compressing to learn and forecast. On the general level, the discussion proposed is oriented to justify the correspondences represented in Figure 1, thus suggesting a parallel between learning and forecasting, for which some kind of coarse-graining procedure - as in modeling - is adopted. Forecasting and learning occur within a processing infrastructure - the brain or the machine, via a suitable representation - that is incapable of capturing all potentially relevant information. The external world can be too complex, and information bottlenecks prevent the transmission of every detail. In other words, although learning and forecasting could be implemented by finding ways to encode the system observed efficiently, coarse-graining procedures come into play and respond to the system's complexity by discarding information via a *lossy-compression mechanism*. This perspective could support the expectation that it is in principle possible to reach through Machine Learning performance at least comparable with that achievable via modeling. Moreover - considering the three levels schematically illustrated in Table 1 - the presence of a scales cascade could be handled more effectively by a procedure capable of automatically compressing irrelevant details without the aid of relevance criteria difficult to implement in modeling.

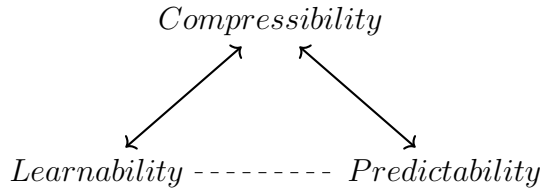


Figure 1: Compression as a common paradigm for Learning and Forecasting, thus establishing a parallel with modeling via coarse-graining.

Some additional comments can provide a clearer understanding of Figure 1. First, the relationship between compression and predictability can be interpreted as follows. Compressing corresponds to selecting a specific system’s scale to determine a level of resolution beyond which further refinement becomes irrelevant to the problem at hand. Assuming we are dealing with a chaotic system - where even a slight error in the initial condition leads to a rapid degradation of predictive capacity - it is possible that this chaotic behavior only pertains to certain scales, ideally those at higher resolutions. In such cases, effective compression can provide a representation free from chaos, wherein the predictability horizon assumes an entirely different regime. To make this type of procedure feasible, there are at least two essential conditions: the compression process must be capable of identifying the system’s scales and their corresponding variables; it is crucial that the chaotic behavior of the system is indeed confined to scales considered irrelevant for modeling the system behavior.

Second, compression also plays a central role with respect to learnability. Learning a problem is a task that differs from memorization to the extent that its achievement can rely on the protocol’s ability to forget. This point, although somewhat vague for now, will assume a precise meaning in Part II, when the compression-generalization trade-off is also discussed in terms of the Bottleneck Principle. By reformulating the bound on the generalization gap presented in (3) in terms of information, the mutual information $I(X : \hat{X})$ between the initial input X and its compressed representation \hat{X} will play a role similar to the cardinality of the hypothesis space. In fact, it will determine the level of complexity of the problem and its degree of learnability.

Finally, the compression mechanism activated should be taken into consideration in more detail. On the one hand, compression-in-resolution involves some limitations that are difficult to circumvent, recognizable as much for the forecasting problem as for the learning problem. On the other hand - from our point of view - the question arises as to whether it is possible to equip Machine Learning protocols with *compression-in-representation procedures that do not pass through compression-in-resolution*. This solution might allow us to circumvent the limitations imposed by the scarcity of data, providing a possible improvement in terms of applicability, opacity, and explainability. Causality could be a useful tool in this direction.

3 Causality

In the preceding section, we explored the task of extracting information from a dataset, which can be regarded as a collection of observations that are both independent and identically distributed. However, observations are not always enough: our epistemological expectations can be more challenging, especially in social sciences, for policy evaluation and engineering issues [72][54][6][140], where understanding and managing concrete problems may depend on identifying *causal-effect relations* [115][117]. Specifically, we are interested in determining how the system’s behavior changes in the face of *external interventions* or in establishing to what extent a *counterfactual statement* is true. This degree of knowledge shelters us from the deluge of *spurious correlations* [26] and controversial conclusions such as the Simpson Paradox [162] - which appear when only observational data are considered - can be correctly interpreted [116].

The Machine Learning community shows a growing interest in causality, with efforts to integrate causal structures into inductive protocols [151][146][152][129]. In this regard, the most established Machine Learning techniques are primarily associative rather than causal. From a pragmatic standpoint, causal information is often challenging to obtain, and the available datasets are typically observational. Theoretically, while the *i.i.d. assumption* is considered the classical gold standard in learning, estimating the system’s response - then moving beyond the observational level - takes into consideration some *distribution shift* [115][129]. It is essential to realize that the richer epistemic picture just sketched entails additional and not trivial to overcome difficulties. Adopting a metaphor proposed by J. Pearl [117][10], more profound questions force to climb the *Ladder of Causality* as represented in Table 2.

Level	Typical Question	Model	Data	ML Methods
3. Counterfactual	What would Y have been if I had acted on X?	$\hat{u} := y_j - f(x_j)$ $x_j := \hat{x}$ $y_j^{\hat{x}} := f(\hat{x}) + \hat{u}$	No Data	No ML methods
2. Interventional	How would acting on X change Y?	$x_j := \hat{x}$ $y_j := f(\hat{x}) + u_j$	$(\hat{x}, y_i) \sim P_{\hat{x}}$ <i>i.i.d.</i>	?
1. Observational	How would seeing X change my belief on Y?	$y_j = f(x_j) + u_j$	$(x_i, y_i) \sim P$ <i>i.i.d.</i>	Supervised/ Unsupervised Learning

Table 2: *The Ladder of Causality for the static-bivariate case.* For simplicity, the system is assumed to be governed by an additive-noise model $Y := f(X) + U$. The symbol $:=$ should be read as an assignment, not a symmetric equation.

Some preliminary remarks can be provided to enhance clarity in anticipation of a more comprehensive examination in the subsequent pages. Let us proceed by points, following the ladder from the step upwards for the bivariate case.

- *How would seeing X change my belief in Y ?* On the first step, observational data (x_i, y_i) are identically and independently distributed according to the a priori probability P , and questions at this level admit answers as conditional probabilities, say $P(y|x)$. We assume that a functional mechanism f governs the relationship between X and Y , excluding measurement errors, unmodelled external factors, and additional stochasticity - all captured by the additive noise U -. For simplicity, we can think of the Machine Learning protocol as aiming to recognize f , for example, by adopting a regression procedure.
- *How would acting on X change Y ?* Moving up to the next step, if we intervene by manually fixing the value of X , data are generally no longer distributed as P . The functional mechanism can be modified, and for each possible external intervention, say $X := \hat{x}$, data are distributed as the corresponding new distribution $P_{\hat{x}}$, to be determined. Generally, observational data are insufficient to extrapolate the system's cause-effect relations, and only a partial causal structure can be uniquely determined. No matter how elaborate Machine Learning protocols may be, the observational distribution and the set of independence relations inferred from data are compatible with a non-trivial class of alternative causal structures [115]. On the other side, causal inferences require causal assumptions: some conditions under which it is possible to answer interventional questions by examining observational data are available, but assuming a causal graph can correctly capture the cause-effect relationships between variables involved.
- *How would Y have been if I had acted on X ?* Finally - climbing the last step - data are unavailable by definition at the counterfactual level, and no experiment in the world will be sufficient to answer counterfactual questions. More precisely, counterfactual inferences require assumptions about the mechanism underlying the observed phenomenon: the model $Y := f(X) + U$ is necessary to make counterfactual conclusions. In a nutshell, the noise actual value \hat{u} is determined via the observational couple (x_j, y_j) , providing $\hat{u} := y_j - f(x_j)$. Consequently, the counterfactual quantity $y_j^{\hat{x}}$ is obtained by intervening with $X := \hat{x}$ on the modified equation $y = f(x) + \hat{u}$ [115][129].

The points above show how causality is essential for accessing information not necessarily obtainable on the observational level, regardless of the size of the system under consideration. Accordingly, the integration of causal structures within inductive protocols bears the potential to enhance their *expressive capacity*. In this regard, the cause-effect problem in the *bivariate* context may be considered both the most fundamental from a conceptual point of view and the least realistic in modeling concrete situations. Given only two observable and *correlated* variables - X and Y - it is intended to establish in which of the two possible alternatives the eventual causal link is directed [73][129]. Although we can not expect to describe a realistic

system as a collection of bivariate relations, the scientific literature contains many discussions about variables in pairs, particularly in social sciences: taxes and government spending, education and income, armaments and aggression, inflation rate and unemployment, social status and political preferences. Moreover - and beyond its practical applications - the discussion of this elementary topic clarifies some general features of causality, its role in making practical inferences, and its possible relations with Machine Learning methods. For these reasons, a formal analysis of the *static model* will be considered in some detail by expanding further the discussion briefly summarized in Table 2. At this point - considering only the interventional level of the ladder - Table 3 collects the two causal alternatives available when the *confounder* - i.e. a third not-observed variable which causes both X and Y providing their non-zero correlation - is excluded a priori [135].


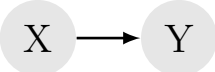

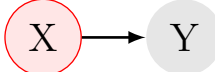

Observational Level	Causal Structure	Post-Interventional Level	Interventional Distribution
	 	 	$p_{\hat{x}}(y) = p(y \hat{x})$ $p_{\hat{x}}(y) = p(y)$

Table 3: *Cause-Effect Pairs*. Whenever the presence of a confounder is ruled out, the intervention $X := \hat{x}$ makes it possible to distinguish which case the system belongs to by comparing the interventional distribution $P_{\hat{x}}$ with the observational one.

Nonetheless - before this investigation - we wish to address a distinct yet informative perspective that concerns once again the comparison between modeling and inductivism. Specifically, we aim to examine how the notion of causality could aid in mitigating the limitations inherent in inductive protocols regarding applicability, opacity, and explainability, as discussed in the previous section. Moreover, causality could explain why Machine Learning methods are rather limited at some crucial feats where *natural intelligence* excels. For example, the learning processes characteristic of biological intelligence do not necessarily require exposure to huge amounts of data, so the elimination of irrelevant details may not involve only a compression mechanism in resolution. In this regard, let us consider some conceptual steps necessary for properly framing causal language as an alternative to making compression. In what follows, the bivariate case is abandoned in favor of the more general case, where m variables and their dependence relations are considered.

Canonical Factorization. Transparency and explainability are difficult to define formally, although they are closely related to how a human being understands the world. Constructing a representation compatible with the agent’s cognitive capacity in terms of memory and computational resources calls into question the analyst’s

ability to determine whether the model assumptions are sound or need additional prescriptions, whether they are aligned with the available data, and to identify which assumptions require improvement. In this regard, *Graphical Models* provide a compact representation that is qualitatively compatible with how humans think about relationships between variables, providing a valuable model for cognition [119]. In a nutshell, the variables X_1, X_2, \dots, X_m are ordered according to a *directed acyclic graph* (DAG) \mathcal{G} , and the probability distribution admits the correspondent *canonical factorization*

$$P_X = \prod_{i=1}^m P_{X_i|Pa_i}, \quad (4)$$

where Pa_i denotes the set of parents of X_i in \mathcal{G} , and can be interpreted as the set of *direct causes* for X_i [82][115]. Originally, this formal apparatus was designed to effectively represent the variables involved through the use of independence relations, achieving benefits on the cognitive, computational, and statistical learning fronts [82][118]. Remarkably, a given set of independence relations can be compatible with alternative DAG structures. The *causal graph* \mathcal{G} - in accordance with Occam's razor-type principle - could then be the one that represents in the most compressed way the information available given a set of independencies. This insight will require the use of a nonstatistical concept of information, such as *Kolmogorov Complexity* [74][129].

Modularity to generalize. When environmental conditions change, a machine trained in one environment cannot be expected to perform well unless localized changes are identified. For example, in computer vision changes in the distribution may come from aberrations, geometrical transformations and quality compression. Problems of this nature can be addressed by employing data augmentation, pre-training, and other techniques that employ some data bias. However, these fixes may not be sufficient, and some structural knowledge about the system could be needed. Artificial Intelligence researchers have recognized this problem and identified subtasks such as *domain adaptation*, *transfer learning*, and *life-long learning* to alleviate the general problem of robustness. In this regard, *modularity* plays a central role and makes it possible to some levels of generalization [115][146]. More precisely, it is assumed that starting from the canonical factorization (4) with respect to the graph \mathcal{G} , a single factor $P_{X_i|Pa_i}$ can be modified by providing a new factor $\hat{P}_{X_j|Pa_j}$ without affecting the others, namely obtaining

$$P_X := \prod_{i=1}^m P_{X_i|Pa_i} \longrightarrow \hat{P}_X := \hat{P}_{X_j|Pa_j} \cdot \prod_{i \neq j}^m P_{X_i|Pa_i}. \quad (5)$$

More in particular, the *distribution shift* (5) formalizes the functional and semantic independence of the mechanisms that characterize the system, then allowing for the identification of the specific mechanism that is being modified, either due to changes in external conditions or the application of a model to a similar but not precisely equivalent problem. Consequently, the generalization achieved is not from one dataset to another - sampled from the same distribution - but from one problem to

another. This *out-of-distribution* generalization is a central ability for *natural intelligence*, which heavily employs interventions, domain shift, and temporal structure to build a modular representation. In this regard, causal models can be a relevant step toward a more transparent and robust Artificial Intelligence [121][117].

Explicit mechanisms. So far the following perspective has been adopted: a causal model consists of a graph structure \mathcal{G} that defines which variables Pa_i directly determine the values of the variable X_i , thus establishing a partial order on X_1, X_2, \dots, X_m . Moreover, modularity allows us to compute *interventional* distribution as in (5). For example, when a variable X_j is intervened upon by fixing its value to \hat{x} , we disconnect it from its parents Pa_j and fix

$$\hat{P}_{X_j|Pa_j} := \delta_{\hat{x}},$$

that is, the Dirac distribution concentrated in \hat{x} . However, the underlying mechanisms that characterize the system at hand have remained implicit and are represented by the probability distribution $P_{X_i|Pa_i}$. This approach is enough when we are interested in answering a causal query in terms of a probabilistic one, placing us at the second level of Table 2. In this regard, a causal graph can be considered an intermediate description between mechanistic models as differential equations in Physics and statistical ones for which the i.i.d. hypothesis holds. A further step is to make the nature of the underlying mechanisms more explicit by considering the language of the *Structural Equation Modeling* (SEM) [115][151][146][152]. In place of the factor $P_{X_i|Pa_i}$, the following rule is considered

$$X_i := f_i(Pa_i, N_i),$$

to be read as an assignment from right to left - and not as a symmetric equation - and for which N_i collects all the exogenous factors, i.e. a random noise on X_i . At this level, the modularity assumption allows us to change f_i - e.g. by fixing X_i to a constant value - without changing f_j . Thanks to this finer-grained representation, we may be able to discuss intervention queries not representable in probabilistic terms [61] and answer additional queries such as the counterfactual ones [115][61].

In our intentions, this further conceptual step makes possible a comparison with other causal indices - including *Granger Causality* [60][66][62], *Transfer Entropy* [149][25], and *Linear Response* [8][48] - contributing to developing a more unified perspective. These points will be discussed in the last part of this dissertation by re-considering the language of dynamical systems. Remarkably, we have so far adopted the language of cause-and-effect relationships without stating that the cause must be temporally antecedent to the effect. In the language of causal graphs, the presence of time constitutes an element of simplification if it is assumed that cause-effect relationships must necessarily satisfy the arrow of time. In the following, we will not dwell on this aspect. Instead, we will present a more detailed analysis of the concept of *information flow*. Although informational quantities are usually understood at the observational level, we will show how causal semantics can also be employed in the

case of Transfer Entropy. It will be done by considering the simplest possible case - the linear case - establishing a comparison with the other indices already recalled.

Causal Pipeline. Having discussed some conceptual elements of causality, it remains to consider its effectiveness with respect to Machine Learning methods. Although causality brings the hope of significant improvements, new critical issues come into play when trying to integrate causality into inductive protocols. On the general level, understanding a system via causal language must necessarily pass through progressive procedural bottlenecks, each with theoretical and practical consequences. The more one intends to climb the ladder of causality - firstly passing through the interventional level and then accessing the counterfactual one - the more critical the barriers become. Each jump involves new assumptions - some of which are unverifiable - showing a significant trade-off between the power of the causal tool and its range of applicability. This point will be discussed in the *causal-pipeline* section. At this level, it may be useful to recall only two general points:

- *Causal Discovery* methods provide inductive procedures to recognize the causal graph \mathcal{G} , and can be analyzed via statistical learning theory [55][115]. Inferring the system's causal structure can be a particularly difficult task, both in terms of computational resources and the amount of data required. More precisely, conditional independence tests - which are central for many causal discovery algorithms - require an exponential amount of data [195]. Computationally, identifying the optimal causal structure is an NP-hard problem [82].
- Even assuming an infinite amount of data is available, multiple different causal graphs may exist associated with the underlying distribution P . Consequently, given a distribution P , there is a degree of uncertainty that limits our ability to infer the correct system's causal structure. The *non-identifiability* of a unique causal structure provides a significant bottleneck on the ability to draw *causal inferences* effectively [121][129].

The two points above will be reconsidered by presenting a section on *state of the art* regarding available algorithms. At any rate - once a causal model is fixed - *causal inference* allows us to draw conclusions on the effect of interventions, counterfactuals, and potential outcomes. In this regard, J. Pearl has proposed *Do-calculus* - i.e. a *complete* set of rules - that allows us to identify the interventional distribution starting from \mathcal{G} and P , also when \mathcal{G} admits unobserved common causes [115][120][121][71].

The inferential conclusions can be checked a posteriori, given the data available and the causal structure can be consequently updated. However, standard causal model assumptions are rarely satisfied in real systems: the acyclicity condition implies a strict hierarchy among the variables, and feedback mechanisms between them are not permitted; in addition, the modularity condition requires that the mechanisms can not communicate with each other and draw on completely disentangled processes. More in general, the problem arises of determining under which conditions an intervention on the system is - in principle - conceivable and what limitations

arise in intervening. All these remarks can significantly limit this framework’s validity range. At any rate, the *causalization* of a system - describing it as admitting a causal structure that satisfies the correspondent causal assumptions - can provide a coarse-grained representation that maintains a good level of fidelity without encoding all the system’s details. Even assuming the *eliminativist* perspective illustrated by B. Russell [139][48], causality provides a compression-in-representation compatible with a notion of intelligence *embedded* in the environment and that takes into account the agent’s ability to intervene and generalize in the face of the distribution shift. From a general viewpoint, Figure 1 can be updated by equipping it as in Figure 2, where red links require to be explained. To argue for this perspective, we will discuss a preliminary issue: the causal graph can be thought of as the most compressed of those that are compatible with the probability distribution [129].

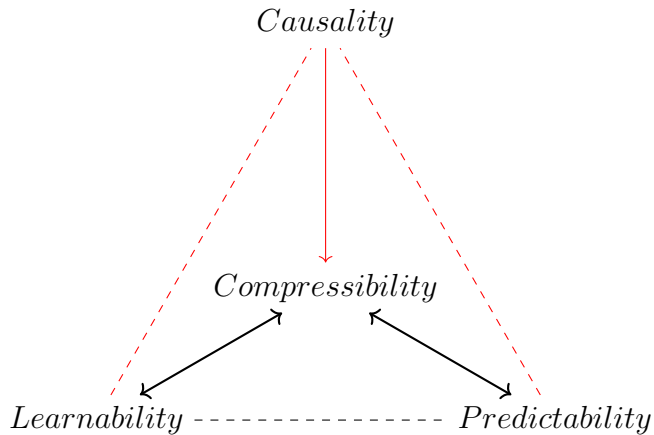


Figure 2: Causation as compression-in-representation with benefits for predicting and learning, also when the generalization problem involves a distribution shift.

To complete the picture offered so far, it would be necessary to identify a general principle that - as for the bottleneck framework in the observational case - characterizes the most appropriate causal representation. A proposal in this regard could take the following form:

$$\min_{P_{\hat{X}|X}, \mathcal{G}_{\hat{X},Y}} I(X : \hat{X}) - \beta I(\hat{X} : Y) + \gamma K[\hat{X}, \mathcal{G}_{\hat{X},Y}],$$

where $K[\hat{X}, \mathcal{G}_{\hat{X},Y}]$ is some information/complexity measure - e.g. Bayesian Information Criterion as in [37] and Algorithmic Mutual Information [74] - and γ corresponds to a complexity/accuracy trade-off parameter for the causal representation adopted. By minimizing K , the optimal canonical representation should be causal, assuming that the causal representation corresponds to the most economical one among those compatible with the observations. This possibility will be explored - at a preliminary and conceptual level - in Part III.

Part II

Forecasting and Learning

This second part aims to discuss the problems of forecasting and learning, which can be reread in terms of a compression procedure. The following discussion examines the Method of Analogues as exemplifying procedures based on *compression-in-resolution*. Beginning with some simple results from Statistical Learning theory, the trade-off between compression and generalization is discussed for Machine Learning techniques.

1 Predictability

Reliability and accuracy in making forecasts reflect the soundness of the scientific knowledge acquired. Conversely, predictions incompatible with experimental evidence challenge the theoretical framework adopted, leading to conceptual revisions and - sometimes - paradigm shifts. Moreover, predicting the future starting from the available information on the past informs our decisions in public and private spheres, often resulting in significant benefits for applications. Thus it is essential to consider - both from a theoretical and practical viewpoint - whether there are general conditions that determine the legitimacy of our predictions, also focusing on their inherent limitations. In what follows, this point will be briefly examined by adopting the language of Dynamical Systems. In this regard, let us start with considering two commonly used indicators to measure the system predictability in relation to the long-term evolution of a small uncertainty: the *Lyapunov Exponent* and the *Kolmogorov-Sinai Entropy*. In what follows, these objects are recalled by adopting [22][32][30][113] as the main references. The discussion carried out will allow us to consider a perspective that calls into question the level of detail in the description of phenomena. In this regard, the presence of different scales - as mentioned in Table 1 - will assume a positive role in terms of predictability.

Preliminary concepts. A dynamical system can be defined as a pair (M, ϕ^t) where the set M collects all the possible system's states, and the function $\phi^t : M \rightarrow M$ works as a transition rule from each state $x \in M$ to its temporal evolute after a time t . Remarkably, (M, ϕ^t) is not a simple input-output machine: the input is given only as an initial condition so that the system changes according to the evolution law ϕ^t . The mathematical properties of these objects depend on the requirements that arise in modeling. For example, given a continuous-time and deterministic process with m degrees of freedom, M is naturally equipped with a smooth m -manifold structure, requiring that $\{\phi^t\}_{t \in \mathbb{R}}$ is a one-parameter group of diffeomorphisms in t , so the group properties

$$\phi^0 = id_M, \quad \phi^{t+s} = \phi^t \circ \phi^s$$

hold. In this case, ϕ^t is associated to a differential equation $\dot{x} = v(x)$, where the bijective correspondence between ϕ^t and v is locally determined via the relation

$$\frac{d}{dt} \phi^t x|_{t=0} = v(x).$$

In what follows, ϕ^t will also denote discrete-time maps, for which $x_t = \phi^1 x_{t-1} = \dots = \phi^t x_0$ and $t \in \mathbb{Z}$. Moreover, other mathematical structures - in addition to differential and topological ones - may come into play: a probability measure μ on M can be introduced to represent the state uncertainty, or a metric structure compares two states by measuring their distance. At any rate, one of the main remarks in this context can be phrased as follows: while the dynamic rule ϕ^t may be relatively simple, its iterated application can lead to behaviors prohibitive to forecast based solely on the rule itself, as the following predictability indices quantified.

The Lyapunov Exponent. Let us consider the separation between two trajectories - say $x(t)$ and $x'(t)$ - which start from two close initial conditions, $x(0)$ and $x'(0) = x(0) + \delta x(0)$ respectively. As long as $\delta x(t) = x'(t) - x(t)$ remains sufficiently small, it can be regarded as a tangent vector of M . Consequently - under rather general hypothesis [22] - the tangent dynamics can be decomposed as

$$\delta x(t) = \sum_{i=1}^m c_i v_i e^{\lambda_i t}, \quad (6)$$

where $\{v_i\}_{i=1}^m$ is an orthonormal basis for the tangent space, the coefficients $\{c_i\}_{i=1}^m$ can depend on the initial condition, and the exponents $\{\lambda_i\}_{i=1}^m$ are supposed to be ordinated as

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m.$$

To visualize the role of the exponents λ_i , we can imagine centering a small sphere in the initial state to consider its evolution. After a time t , this sphere is mapped to an infinitesimal ellipsoid, and the ratio of its radius i to the initial radius will be of the order $e^{\lambda_i t}$. In this regard, Osledec's *Multiplicative ergodic theorem* ensures that this picture does not depend on the initial state [113][32][80]. More in particular, if the dynamical system has an ergodic invariant measure - i.e. an ergodic natural measure on the system's attractor considered - the spectrum $\{\lambda_i\}_{i=1}^m$ does not depend on the initial condition almost surely, and the distance between the two trajectories considered evolves - given the tangent approximation in (6) - as

$$\|\delta x(t)\| \approx \|\delta x(0)\| e^{\lambda_1 t} \{1 + O(e^{-(\lambda_1 - \lambda_2)t})\}. \quad (7)$$

From the perspective of the system predictability, the exponential amplification of the initial error $\delta x(0)$ comes into play whenever the largest Lyapunov Exponent $\lambda := \lambda_1$ - called the system's Lyapunov Exponent - is positive. In this case - given an initial uncertainty $\delta_0 := \delta x(0)$ - the system behavior is predictable within a tolerance Δ only up to a *predictability-horizon*

$$T_p \sim \frac{1}{\lambda} \ln \left(\frac{\Delta}{\delta_0} \right). \quad (8)$$

This relation shows that T_p is basically determined by the Lyapunov Exponent λ , while its dependence on δ_0 and Δ only logarithmically. In other words, λ gives quantitative information on how rapidly the ability to predict the evolution of a system is lost. After a time enough larger than $1/\lambda$, the system state may be found almost everywhere. Remarkably, the Lyapunov Exponent is a global quantity that characterizes the system's fine-grained properties, namely about the generic trajectory for a long time and infinitesimally small perturbation [22][30].

The Kolmogorov-Sinai Entropy. The stretching of infinitesimal displacements - as quantified by the Lyapunov Exponent - is a well-established characterization for the chaotic dynamics. Additionally, the unpredictable nature of chaotic systems can be re-read in terms of the wide variety of alternative orbits, as pointed out by

Kolmogorov in applying Shannon's information characterization. In a nutshell, given a finite partition $\mathcal{U} = \{U_i\}_{i=1}^u$ of M , the entropy function

$$H(\mathcal{U}) = - \sum_{i=1}^u \mu(U_i) \log \mu(U_i) \quad (9)$$

quantified the average information gains in knowing that the orbit belongs to one of the elements of \mathcal{U} . The general idea is simple: given the initial condition x with some inaccuracy, we wish to measure how much additional information ϕ generates in reducing the uncertainty on x . In other words, the information (9) can be refined by taking into account the state evolution via iterated applications of ϕ , namely by adopting the finer partition

$$\mathcal{U}_\phi^k := \{U_i^k := U_{i_0} \cap \phi^{-1}U_{i_1} \cap \dots \cap \phi^{-k+1}(U_{i_{k-1}}) : U_{i_s} \in \mathcal{U}\}$$

and considering $H(\mathcal{U}_\phi^k)$ in place of (9). At this point the quantity

$$h(\phi, \mathcal{U}) := \lim_{k \rightarrow \infty} \frac{1}{k} H(\mathcal{U}_\phi^k) \quad (10)$$

can be regarded as the average information rate achieved - asymptotically - by considering the finer partition at stage $k + 1$ in place of the coarser one at stage k . Remarkably, definition (10) depends on \mathcal{U} and the Kolmogorov-Sinai Entropy of ϕ is defined removing this dependence via the supremum operation

$$h(\phi) := \sup_{\mathcal{U}} h(\phi, \mathcal{U}). \quad (11)$$

In view of the discussion that follows, we adopt a more constructive definition than (11) by considering a compact phase space M equipped with a metric d . First, fixed a resolution ε , two points of M distant less than ε are considered indistinguishable. However - after a time k - their iterations could lie in two not-intersected ε -spheres and then turn out to be distinguishable. In other words, orbits indistinguishable up to an instant k may begin to be distinguishable, thus allowing us to access additional information about the initial condition. To formalize this point, let us introduce a new metric d_k and its correspondent ε -spheres

$$d_k(x, y) := \sup_{0 \leq r \leq k-1} d(\phi^r x, \phi^r y),$$

$$B_\varepsilon^k(x) := \{y : d_k(x, y) < \varepsilon\}.$$

Remarkably, two distinct points in $B_\varepsilon^k(x)$ result - by definition - indistinguishable at resolution ε at least until $k - 1$ iterations, so that the chain of inclusions

$$B_\varepsilon^1(x) \supseteq B_\varepsilon^2(x) \supseteq \dots \supseteq B_\varepsilon^k(x) \supseteq B_\varepsilon^{k+1}(x)$$

holds. This observation allows us to define a family of coverings in ε -spheres that provides a *filtration* for M . First of all - for each k - we define the M -covering

$$\mathcal{U}_\varepsilon^k := \{B_\varepsilon^k(x) : x \in M\},$$

so that $\mathcal{U}_\varepsilon^{k+1}$ is finer than $\mathcal{U}_\varepsilon^k$: for every $B_\varepsilon^k \in \mathcal{U}_\varepsilon^k$, there exists an ε -sphere $B_\varepsilon^{k+1} \in \mathcal{U}_\varepsilon^{k+1}$ such that $B_\varepsilon^{k+1} \subseteq B_\varepsilon^k$. Moreover, given any other covering \mathcal{U} , there exists a k for which $\mathcal{U}_\varepsilon^k$ is finer than \mathcal{U} . Considering the smallest finite sub-covering of $\mathcal{U}_\varepsilon^k$ and denoting by $\mathcal{N}_k(\varepsilon)$ its cardinality, the quantity

$$h_\varepsilon(\phi) := \lim_{k \rightarrow \infty} \frac{1}{k} \log \mathcal{N}_k(\varepsilon) \quad (12)$$

measures the growth rate of ε -spheres needed to cover M , at a given resolution ε . Remarkably, $\mathcal{N}_k(\varepsilon)$ can be interpreted as the minimal number of distinguishable k -orbits $\{\phi^t x\}_{t=0}^{k-1}$ needed to represent all the possible k -orbits at the fixed resolution ε , so that (12) is an index for the correspondent grow rate. The *Topological Entropy* is then defined via a limit operation on the resolution:

$$h_{top}(\phi) := \lim_{\varepsilon \rightarrow 0} \lim_{k \rightarrow \infty} \frac{1}{k} \log \mathcal{N}_k(\varepsilon) \quad (13)$$

However, $\mathcal{N}_k(\varepsilon)$ in (13) also takes into account orbits that are less and less probable as k increases. By considering again the role played by the measure μ , our analysis should be restricted to the *typical orbits*. At this point, it is sufficient to consider the following observation: if μ is the natural measure on a chaotic attractor, for a sufficiently large k the only observable orbits are those that live on the attractor. The substitution

$$\mathcal{N}_k(\varepsilon) \longmapsto \mathcal{N}_k^{eff}(\varepsilon)$$

in (13) - considering only the effective number $\mathcal{N}_k^{eff}(\varepsilon)$ for sufficiently large k - allow us to re-obtain the Kolmogorov-Sinai Entropy [22] as in (11) by removing the dependence on resolution ε through the limit operation

$$h(\phi) = \lim_{\varepsilon \rightarrow 0} \lim_{k \rightarrow \infty} \frac{1}{k} \log \mathcal{N}_k^{eff}(\varepsilon). \quad (14)$$

It can be shown that $h(\phi) \leq h_{top}(\phi)$. On the one hand, if $h_{top}(\phi) > 0$, the system may have non-chaotic attractors and a non-attracting chaotic invariant set. On the other hand, the condition $h(\phi) > 0$ tells us that at each sufficiently small ε scale - also under the assumption of considering only typical orbits for sufficiently long times - the number of ε -spheres needed to cover M grows exponentially with k . For systems of this type, the elimination of details always produces a limit on the predictability horizon, whatever resolution is chosen. This fact can also be interpreted as sensitive dependence on the initial condition in the sense of chaotic dynamics - i.e. when the Lyapunov Exponent is positive - although the divergence mechanism is combined to a converging one due to the compactness of M .

Large-Scale Predictability. While the Lyapunov Exponent quantifies the system's unpredictability regarding the stretching of infinitesimal displacements, on the other hand, Kolmogorov-Sinai Entropy measures the growth rate of the typical alternative orbits. Remarkably, the *Pesin-Ruelle formula* establishes a link between these

two viewpoints [22][32]. In the simplest case,

$$h(\phi) \leq \sum_{\lambda_i > 0} \lambda_i$$

so that for $h(\phi) > 0$ the estimate (8) for the predictability-horizon come back into play. However, Lyapunov Exponent and Kolmogorov-Sinai Entropy - as remarked above - are fine-grained indices and quantify a global characterization of the dynamical system at hand, considering all the system's degrees of freedom. In this sense, they provide a *worst-case* estimate about the system predictability, even under the assumption of considering only typical orbits. The fine-grained representation does not take into consideration the possible existence of *different characteristic scales*: no advantages of restricting to a particular description level are employed by removing details irrelevant to the predictions of our interest. From a general viewpoint, a dynamical system can exhibit both chaotic behavior and simplifying tendencies, where certain degrees of freedom dominate and constrain others. In other words - although forecasting can be challenging when considering the most detailed description level - some fluctuations can average out when we change the resolution scale, revealing a more regular behavior for less detailed description levels so that the estimate (8) for the predictability-horizon can result too pessimistic for many concrete problems. To clarify this point, let us briefly consider an example proposed in [22][23][32], for which the following three-dimensional coupled map is defined:

$$\begin{cases} x(t+1) = R_\theta x(t) + \varepsilon f(y(t)) \\ y(t+1) = 4y(t) \cdot [1 - y(t)] \end{cases}$$

where $x \in \mathbb{R}^2$, $y \in \mathbb{R}$, R_θ is a rotation of angle θ , $f(y) := (y, y)$ and $\varepsilon \geq 0$. For $\varepsilon = 0$ the system provides two decoupled maps: the rotational in x , for which the Lyapunov Exponent is $\lambda_x = 0$; the chaotic map for y , that admits $\lambda_y := \ln 2$. On the other side, given a small coupling $\varepsilon > 0$, the correspondent three-dimensional map is chaotic, and its largest Lyapunov Exponent is

$$\lambda = \lambda_y + O(\varepsilon).$$

Consequently, the predictability horizon for x would scale as $1/\lambda_y$ if the estimate (8) were employed naively. This conclusion would conflict with the basic understanding of the system. In this regard, we expect predictability for x to be unaffected by the behavior of y , which is negligible for small ε . This incongruity derives from the mistaken application of (8), as explained in [32][23]. While (8) is valid only for both δ_0 and Δ infinitesimal - as soon as the error becomes large - the full nonlinear error evolution has to be taken into account. More in particular, at the beginning both $|\delta x(t)|$ and $|\delta y(t)|$ grow exponentially according to the tangent-space dynamics in (7). After a time $O(1/\lambda_y)$, the evolution of $|\delta y(t)|$ leads the saturation due to the boundedness of the available phase space for y , i.e. $|\delta y(t)| \sim O(1)$. Consequently, two y -subsystem realizations are completely uncorrelated and their difference δy acts as noise on the equation for $\delta x(t)$. As a consequence, the growth of $|\delta x(t)|$ becomes

diffusive: $||\delta x(t)|| \sim \varepsilon t^{1/2}$ and the predictability-horizon for x scales as

$$T_p^x \sim \frac{\Delta^2}{\varepsilon^2}. \quad (15)$$

What has just been discussed exemplifies a general mechanism in accounting for how unpredictability and predictability can coexist when considering different scales. In these situations, considering a too-accurate observation scale can be both unnecessary and misleading. Conversely, large-scale predictability is achievable by moving across resolution scales and removing small-scale details. In the example discussed, the model delivers a representation that allows us to naturally distinguish which details are irrelevant for forecasting. On the other side, recognizing the underlying mechanisms may not be feasible, as illustrated in Table 1. In the realm of big data, Machine Learning techniques may have the capability to uncover patterns across various scales within the data, traversing the hierarchy of complexity without necessitating a comprehensive understanding to determine the relevant scales for the given task at hand. In what follows, the shortcut via inductive procedure will be discussed in more detail.

2 The Analogs Method

Red sky at night, sailor's delight. Red sky in the morning, sailor's warning. Every culture abounds with propositions of this nature, and grandmothers who have long lived in the same place typically declare that they can predict the weather based on what they have already observed. Nevertheless, no one should adopt the above statement as having scientific validity. Although some epistemological considerations typically come into play in this regard, some serious limitations emerge when considering a rigorous formulation of this inductive procedure. Let us start by formalizing the grandmother's approach via the Analogs Method, which has been used in meteorology since Lorenz's work [92][93][94][95]. Given the actual state z of which we want to forecast the evolution after an amount of time t , say $\phi^t z$, we consider in our database S a past state a similar to z , possibly the best one, and approximate $\phi^t z$ with $\phi^t a$:

$$\begin{aligned} a \sim z &\implies \phi^t a \sim \phi^t z, \\ a, \phi^t a &\in \mathcal{D}. \end{aligned}$$

The Analogs Method does not require the functional form of ϕ^t . Operationally, the concrete algorithm that provides a forecast is straightforward and reduced to a search and sorting procedure in the database \mathcal{D} . Theoretically, the Analogs Method requires an underlying *regularity assumption*. After all, it seems to be a natural claim that if a system behaves in a certain way, it will do so again, at least for numerous practical tasks. Moreover, in the conceptual scheme offered by the Ladder of Causality, it is clear that Analogs Method is placed at the first level: interventions and counterfactuals have no role, while the data available are observational records of the underlying process. On the other side, we have not yet speculated on the mechanism that generated our database \mathcal{D} . In this regard, two major viewpoints of forecasting based

on historical datasets are available. The first assumes that an observed time series is a specific realization of a random process, where the randomness can be typically amplified by the interaction between many degrees of freedom. The second states that random behavior may be generated by deterministic systems also with a few degrees of freedom but interacting non-linearly, emerging from a not infinitely precise knowledge of the current state [24]. In our notation, ϕ^t stays for the *deterministic evolution law* specified by the underlying system’s equations in the manner of the Dynamical Systems theory.

Considering the statistical literature for classification tasks, Analogs Method shares many characteristics with KNN. This protocol looks at the K points in the training set that are *nearest* to the input x , counts how many members belong to each class, and returns an estimation of the probability $p(y|x)$ or an output y through a majority vote [103][108]. In what follows, we will emphasize the role of some dynamical ingredients, such as *recurrence* and *ergodicity*, leaving the parallelism above in the background. Nevertheless, while KNN is considered one of the most simple Machine Learning algorithms, Analogs Method represents an archetypal example among inductive protocols for forecasting. Consequently, it deserves a more detailed analysis to discuss its assumptions, strengths, and limitations.

Main Assumptions and Advantages

In the standard Analogs Method procedure only one single analog is used, although slightly different variants employ the K best analogs, for example performing weighted averages or random selections. However, all alternative protocols are based on at least three shared assumptions.

A.1. Adequate feature space and measure of similarity δ are available.

This first assumption concerns our background knowledge. Generally, every system’s state can be characterized through a D -dimensional vector, where each element corresponds to one observable. Determining the correct number D and identifying the correspondent observables are not obvious tasks. Sometimes, practical considerations can lead us to map the D -dimensional *phase space* into a lower d -dimensional *feature space*. In this regard, a relevance criterion is adopted, and dimensional reduction procedures such as Principal Components Analysis are commonly used in data-oriented applications [103][108]. On the other hand, if only a few variables are observed, time-embeddings techniques are supported by Takens’ results: the *attractor* can be identified through the *delay reconstruction map*, at least theoretically [176]. In short, given a time series for a one observable f , say $\{f_t\}_{t=1}^T$, there is a finite integer d such that the delay coordinate d -vector at time τ

$$x_\tau^d = (f_\tau, f_{\tau+1}, \dots, f_{\tau-d+1})$$

can faithfully reconstruct the properties of the underlying dynamics. At any rate, the adequate feature space is not enough. We will also assume that a *measure of similarity* between states is introduced. The Analogs Method employs a notion of closeness,

for which the Euclidean distance is a standard option, although other possibilities are available in principle. For example, we can select a metric that gives more weight to some features than others; again, the analogs similarity can be quantified via statistical quantities. This choice is typically motivated by considerations that are relevant to the specific problem to be addressed. In what follows - for our conceptual analysis - it will suffice to assume that the appropriate space M and metric δ are known.

A.2. Close analogs can produce good approximations. This second assumption is about the system’s behavior. Given a generic dynamical system, we can provide an a priori estimation of its orbits divergence. If δ_0 is the distance between two states at $t = 0$, as defined in A.1, the *a priori* estimate of its evolution after an amount of time T is

$$\delta_T \approx \delta_0 e^{\lambda T}, \quad (16)$$

where λ is the system’s Lyapunov Exponent [32][80]. As discussed above, if $\lambda > 0$, the error will grow exponentially in time. Fixing the error tolerance δ^* and the predictability time T of our interest, we consequently select an accuracy ε for which we should take $\delta_0 < \varepsilon$ to have sufficiently accurate previsions. Remarkably, the evaluation of the error growth (16) provides - at least in principle - a way to empirically determine λ starting from a long time series. More recently, also ML approaches have been proposed for obtaining Lyapunov Exponents from data, as in [114]. We remark that the estimate (16) can be too pessimistic, and deterministic systems without chaotic behavior admit the error growing polynomially in time. It is well known how chaos constitutes a significant limitation to our ability to forecast, even for low-dimensional systems. What we are interested in highlighting is how the pair (T, δ^*) returns an estimate for ε , given the system properties quantified by λ .

A.3. Close analogs can be found examining the known history. This third assumption is about data being concretely available. Every database is to some extent representative of the system under study. More precisely, given the metric δ provided by A.1, we can cover M with spheres of radius ε as defined in A.2. We will say that the database explores all the space M at a *resolution* ε if it contains at least one point for each sphere. At this stage, a counterintuitive property of high-dimensional spaces comes into play. Suppose d increases, the relative measure of the ε -spheres definitively decreases, and more data will be required to have M entirely explored by the database in question. Quantitatively, the number $\mathcal{N}(\varepsilon)$ of ε -spheres needed to cover M scales as $1/\varepsilon^d$. The Analogs Method is based on the assumption for which ε -similar points will produce comparable evolutions. However, \mathcal{D} may not be large enough to contain all the analogs needed, at least if d is sufficiently large.

Despite its simplicity, the Analogs Method has significant advantages both from the practical and theoretical sides.

S.1. It does not require an underlying model. The Analogs Method has a *non-parametric* nature and can reproduce nonlinear relationships requiring only a

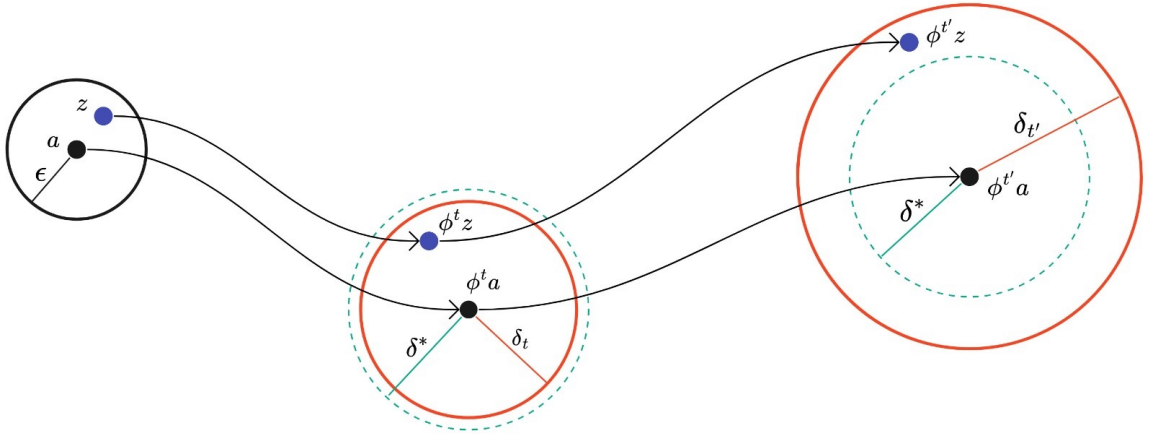


Figure 3: The Analogs Method illustrated. After the predictability time T the error δ_t overcomes the tolerance δ^* . The a-priori trade-off between δ , T and ϵ is governed by (16).

few *meta-parameters* to work, such as the choice of distance or the number of analogs used at each iteration. In other terms, the *background knowledge* needed to implement this procedure is more contained. Moreover, it is in principle capable of providing realistic simulations without any simplifications about the underlying system.

S.2. It is computationally affordable and easily interpretable. Even when a model is available, it is not necessarily tractable from a computational point of view, given our finite resources and fixed constraints. Conversely, the Analogs Method is supported by fast algorithms in researching suitable analogs in large databases, assessing uncertainties by performing ensemble forecasts [88]. Moreover, the art of building good models is very far from being clearly understood, and model-driven approaches require many expertises and a high cognitive cost. On the other side, the simple principles behind AM make it easily interpretable, both from an operational and theoretical point of view.

Although the Analogs Method was initially proposed for weather predictions, forecasting is primarily based on physical models combined with data assimilation techniques. Nevertheless, *the prominence of the modeling approach over the grandmother's philosophy is not due only to epistemological considerations*. On the contrary, the main obstacle for the latter concerns the properties of the available databases. More precisely, assumption A.3 seems falsified in many factual contexts. As Lorenz himself realized [93][94], the main issue is to find suitable analogs, and this drawback arises from the need for unmanageable massive data [182], regardless of the system's chaotic behaviour[31].

Limitations

One of the main ingredients supporting AM is the pervasive role of recurrence in deterministic systems [32][80]. In this regard, Poincaré proved the fundamental recurrence theorem [132]: *for every subset A of non-zero measure and every probability-preserving transformation ϕ , almost every point of A will return to A , after a sufficiently long but finite number of iterations of ϕ .* In other terms, every orbit will come arbitrarily close to a previous state, even for chaotic behaviors. At the same time, although this existence theorem is good news, some limitations arise unavoidably.

L.1. A suitable analog can require a too-large database in many concrete situations. As stated above, Poincaré’s result is an existence theorem and does not provide a quantitative estimation of the recurrence time. A classical result proved by Kac fills this gap, stating that the average recurrence time to a subset A is the inverse of the probability of A [76]. This elementary fact produces a strong limitation: the expected recurrence time increases exponentially with the system dimension d [31][69]. More precisely, given the resolution ε in searching analogs in the past, the database size L scale as

$$L \sim \frac{1}{\varepsilon^d}. \quad (17)$$

In other terms reducing our resolution to a factor α leads to a factor α^d for the database size. When d is moderately big, the needed L can become unmanageably large. The estimate (17) justifies in a dynamic context the a priori estimate for the number n_ε of ε -spheres necessary to cover M , providing an argument for the *curse of dimensionality* in the forecasting framework. It is important to emphasize that the estimation above is an average result, while the waiting time can depend on the system’s local properties [111]. Consequently, a high variability around the mean value could justify why AM works in some regimes and does not in others.

L.2. A given database can not typically explore the feature space with a satisfactory resolution. The problem of how long we must wait when searching for adequate analogs can be replaced by a more applications-oriented one. Fixing the dataset length L , we are interested in identifying *the best analogs possible* to evaluate AM performance. Denoted as $\{a_i\}_1^n$ the n best analogs of the actual state z , we define the distance $r_k := \delta(z, a_k)$ assuming that $r_i < r_j$ for $i < j$. The random variable R_k associated to r_k admits the following estimation for its expected value and variance:

$$\langle R_k \rangle \approx \left(\frac{k}{L} \right)^{1/d}, \quad (18)$$

$$\sigma_{R_k}^2 \approx \frac{1}{d^2 L^{2/d}} k^{\frac{2-d}{d}}. \quad (19)$$

The relations above state that for very low dimension, say $d < 2$ the first distance R_1 has a lower variability than R_k , while the variance is constant when $d = 2$. On the other hand, the inverse phenomenon happens for higher dimensional systems. The

successive analogs in the available catalog are progressively pushed far from the actual state z with decreasing uncertainty. This *distances concentration* would justify why a lower number of analogs is used in performing good forecasts in low-dimensional space, while we expect that high values of k would not greatly impact the performance in higher dimensions.

To obtain (18) and (19), let us evaluate how the best possible analogs are distributed, considering a dynamical system (M, ϕ^t, μ) , where μ is a conserved probability measure. We consider a database with L data points, corresponding to L random variables X_1, X_2, \dots, X_L . Under the ergodic assumption and if we assume that enough time passes between one measurement and another, it is natural to consider these random variables as independent and identically distributed. Remarkably, the i.i.d. hypothesis is shared with the standard ML framework. Let $z \in M$ be the actual system state, with $\delta_i := \delta(z, X_i)$ for a given metric δ on M . Let $R_1 := \min_i \{\delta_i\}$ be the random variable that corresponds to the distance between the actual state z and the best available analog. By definition, we have

$$\begin{aligned} \text{Prob}(R_1 > r) &= \text{Prob}(\delta_1, \delta_2, \dots, \delta_L > r) \\ &= \prod_{i=1}^L \text{Prob}(\delta_i > r) \\ &= [1 - \mu(B_{z,r})]^L, \end{aligned}$$

where $B_{z,r}$ is the z -centered ball with radius r . It is well-known that if μ is ergodic on a given attractor and the limit for $r \rightarrow 0$ exists, then the effective dimension is well-defined:

$$d_{z,r} := \frac{\ln(\mu(B_{z,r}))}{\ln r} \xrightarrow{r \rightarrow 0} d,$$

where d is independent from z and $\mu(B_{z,r}) = r^d$. Putting it all together, for $L \gg 1$, we have

$$\text{Prob}(R_1 > r) = [1 - r^d]^L \approx e^{-Lr^d}.$$

Given the cumulative function $F_{R_1}(r) = 1 - \text{Prob}(R_1 > r)$, the probability distribution calculation is straightforward:

$$p_{R_1}(r) = dLr^{d-1}e^{-Lr^d}. \quad (20)$$

This calculation is generalizable by considering the n best analogs a_1, a_2, \dots, a_n , where $\delta_1 < \delta_2 < \dots < \delta_n$ and $n < L$. Let R_k be the random variable associated to the distance $\delta_k := \delta(z, a_k)$, where a_k is the k -th nearest analog. What follows aims at determining the density distribution $p_{R_k}(r)$, with $k \leq n$. For simplicity, we start with $k = 2$, observing that

$$\begin{aligned} \text{Prob}(R_2 > r) &= \text{Prob}(a_1 \in B_{z,r}; a_2 \dots a_L \in B_{z,r}^c) + \text{Prob}(R_1 > r) \\ &= \binom{L}{1} (1 - r^d)^{L-1} r^d + (1 - r^d)^L \\ &\approx (1 + Lr^d)e^{-Lr^d}, \end{aligned}$$

where the last approximation holds for $L \gg 1$. Given the equations as follows

$$\begin{aligned} \text{Prob}(R_k > r) &= \text{Prob}(a_1 \dots a_{k-1} \in B_{z,r}; a_k \dots a_L \in B_{z,r}^c) + \text{Prob}(R_{k-1} > r) \\ &= \binom{L}{k-1} (1-r^d)^{L-k+1} (r^d)^{k-1} + \text{Prob}(R_{k-1} > r), \end{aligned}$$

and proceeding inductively, we obtain

$$\text{Prob}(R_k > r) = \sum_{i=0}^{k-1} \frac{(Lr^d)^i}{i!} e^{-Lr^d}.$$

Consequently, (20) is generalized as

$$p_{R_k}(r) = \frac{dL^k r^{dk-1}}{(k-1)!} e^{-Lr^d}, \quad (21)$$

with expected value and variance

$$\begin{aligned} \langle R_k \rangle &= \frac{1}{L^{1/d}} \frac{\Gamma(k + \frac{1}{d})}{\Gamma(k)}, \\ \sigma_{R_k}^2 &= \frac{1}{L^{2/d}} \left[\frac{\Gamma(k + \frac{2}{d})}{\Gamma(k)} - \frac{\Gamma(k + \frac{1}{d})^2}{\Gamma(k)^2} \right], \end{aligned}$$

where $\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt$ is the standard Gamma function. By adopting standard approximation techniques, we obtain (18) and (19). The average value of R_k scales as $k^{1/d}$ and the approximation will be increasingly valid as k grows when $d > 2$, with R_k progressively concentrated around its maximum

$$R_k^* = \left(\frac{k - \frac{1}{d}}{L} \right)^{1/d}.$$

To summarize, it is well known that chaos is the main limiting factor to predictability in deterministic systems when the evolution laws are known. This limitation can be formalized by using Lyapunov exponents or through Kolmogorov-Sinai Entropy, for which the notion of resolution plays an explicit role in the definition. On the other side, the first bottleneck for a purely inductive strategy lies in Poincaré recurrences, so good analogs are not available if the system's *effective dimension* is moderately large. The effective dimension can be understood in terms of the attractor's dimension or as the number of degrees of freedom that persist if we take advantage of the possible multiscale structure. In the latter case, the relationship between T and ε can be polynomial as in (15). However, for moderately large effective d , the size of the dataset remains prohibitive in many applicative contexts. A similar limitation emerges in considering the performance of a given dataset, and the probability distributions for the best-analog distances are derived analytically as in (21). Moreover, other limitations can come into play. For example, the method's performance is compromised when *extreme events* - typically rare - are considered: it is easy to imagine situations in which Analogs Method is suboptimal. All these limitations naturally emerge in a data-driven context where neither an explicit model nor other information on the system's causal structure is available.

3 Interlude. The Relevance Objection.

It is a popular claim that Machine Learning methods are essentially inductivist, although many details of such inductivism remain in the dark. Consequently - given the algorithmic differences between the Analogs Method and Machine Learning protocols - it could be argued that the analysis proposed above is irrelevant in discussing other inductive procedures and their possible limitations. A position of this type has recently been supported by the epistemologist W. Pietsch:

A central problem with the argument based on Poincaré's recurrence theorem and Kac's lemma is that it presupposes a mistaken picture of how Machine Learning algorithms reason inductively from data ([130], p. 58).

To properly frame how Machine Learning protocols work - as opposed to the Method of Analogs - Pietsch distinguishes at least two categories of induction. On the one hand, *enumerative* protocols such as Analogs Method would be based on the regularity of recorded events; on the other, the *variational* approach should have its foundation in changing circumstances. In the former case, pattern recognition would become more robust as the number of observed co-occurrences and similarities increases. In the latter, the ability to predict would be based on the numerous variations available and thus on the protocol's capacity to recognize causal links. To make this last point clear, let us go into a bit more detail by stating *Variational Induction* as proposed in [130][131], where the *methods of difference and agreement* have a central role.

Methods of difference and agreement. First of all, a general premise is necessary. All the system variables are assumed to be known, and three objects are considered: the *observables* X and Y , for which the causal relationship is investigated; the set of *background variables* B_1, B_2, \dots, B_n , that is to say, all other variables in the system. Additionally, it is essential to highlight that *the system's causal structure is at the beginning unknown*. At this point, we can proceed by introducing the methods mentioned above.

Method of difference. Given the same background B , two instances are observed. In the first, circumstance X is present and phenomenon Y is present; in the second, circumstance X is absent and phenomenon Y is absent. Then X is causally relevant to Y with respect to background B if and only if B guarantees homogeneity.

Method of agreement. Given the same background B , two instances are observed. In the first, circumstance X is present and phenomenon Y is present; in the second, circumstance X is absent and phenomenon Y is still present. Then X is causally irrelevant to Y with respect to background B if and only if B guarantees homogeneity.

Both methods invoke the *Homogeneity Condition*, which captures the intuition that all circumstances causally relevant to the phenomenon must be fixed, except those

explicitly under examination. In the author’s intention, this condition allows one to compare different instances and identify causal relationships between observables.

Homogeneity Condition. Context B guarantees homogeneity with respect to the relationship between X and Y if and only if context B is thus defined that only circumstances that are causally irrelevant to Y can change, except for X and for circumstances that are causally relevant to Y in virtue of X being causally relevant to Y.

The second exception allows changes for circumstances that lie on a causal path from X to Y or effects of circumstances that lie on it. This remark will be discussed in the last section of this paper. At any rate, the author’s main thesis states that the most successful types of Machine Learning algorithms all implement Variational Induction precisely because of its causal nature. In other words, while not giving us an explicit model, some Machine Learning protocols would be able to capture causal relationships between system variables, thus circumventing the limitations discussed in the previous section.

Relevance Objection. By virtue of the elaboration schematically proposed in the previous paragraph, W. Pietsch goes into the merits of the argument based on Kac’s lemma.

Any estimate of the amount of data required for predictions based on Kac’s lemma is misguided since this lemma takes into account the phase space in its full dimensionality, mistakenly considering all circumstances to be equally relevant ([130], p. 59).

This remark on the theoretical level is accompanied by a more applications-oriented claim about the role of the system’s dimensionality:

Big data approaches have been successful not only for low-dimensional problems but also for extremely high-dimensional problems, as in the case of the prediction of skin cancer from images. In this example, a large number of pixels of these images were taken into account to predict, wherein the number of pixels determines the order of magnitude of the dimensionality of the problem ([130], p. 58).

In what follows, we will argue in favor of a broader interpretation for the already discussed limitations by adopting the *informational language* as a unifying framework, also considering a comparison with Deep Learning protocols via a high-level analysis. However, before proceeding further, it may be helpful to focus on some preliminary misconceptions contained in the recalled quotations.

- First, there is a substantial difference between the number of pixels and the number of features needed to classify a set of images correctly. The latter is closer to dimensionality in the Dynamical Systems framework, where dimension corresponds to the number of relevant variables or effective degrees of freedom.

- Second, Kac’s lemma neither considers the full dimensionality of the phase space nor assigns the same relevance to all possible circumstances. On the contrary, the analysis proposed above is applicable when the system has some attractor where the measure is concentrated or when the multiscale structure is exploited.

Hopefully, the previous sections should have clarified all these preliminary points. At a general level, the argument based on Kac’s lemma concerns the number of data required to explore a phase space of dimension d at a resolution ε . In this sense, it corresponds to a rigorous formalization of the *curse of dimensionality* in the context of forecasting. While Machine Learning protocols employ procedures that are undoubtedly different from Analog Method, the question arises as to what terms ε -resolution plays a role in the context of learning. This point will be discussed in the following pages. First, however, let us consider the constructive part of the argument proposed by W. Pietsch.

Variational Induction

In this section, we turn to Pietsch’s theoretical proposal, for which the Variational Induction framework justifies Machine Learning algorithms’ performances precisely because of its causal nature. If this were the case, the epistemological picture offered by the Ladder of Causality would have to be revised. We could say that an inductive protocol that operates on observational data without making causal assumptions works well precisely because it can recognize cause-effect relationships, a prerogative of the ladder’s second rung. On the other side, the weakness of Pietsch’s proposal lies in the role played by homogeneity. If Pietsch was right, then *the Homogeneity Condition should be satisfiable when considering a protocol without the system’s causal structure being known*. On the contrary, homogeneity is satisfiable only if we can a priori distinguish between variables placed along a chain from X to Y and variables that are not. In other words, Variational Induction could say something about causality only if the causal graph was a priori known - at least partially - against the argument that causality can instead be captured by Variational Induction. If the system’s causal structure is unknown and we consider a purely observational level, we can admit at most a *weak homogeneity condition*, thus modifying the methods proposed.

Weak Homogeneity Condition. Context B guarantees homogeneity with respect to the relationship between X and Y , iff context B is thus defined that only circumstances that are causally irrelevant to Y can change, except for X .

However, the weak homogeneity condition does not allow us to extract causal information from observational data. Clearly, the whole discussion can be done in terms of a causal graph, so we will say that X is causally relevant to Y if the underlying causal graph admits a chain from X to Y . Adopting this perspective, it becomes almost automatic to show the point via extremely simple examples.

Method of difference. Let us consider a causal graph consisting of the nodes $X, Y, B = \{B_1, \dots, B_n\}$, with unknown skeleton and arrows' directions. Suppose we fix the values of each individual variable in B , $B_1 = b_1, \dots, B_n = b_n$. We say that X is causally relevant for Y if

$$\Delta X \neq 0 \Rightarrow \Delta Y \neq 0. \quad (22)$$

If we consider the purely observational level, the validity of the relation (22) can only be investigated by studying the dependence between variables X and Y under the assumption of conditioning on B . In other words, we must verify that the conditional dependency condition is satisfied:

$$P(X, Y|B = b) \neq P(X|B = b)P(Y|B = b), \quad (23)$$

where $B = b$ means $B_1 = b_1, \dots, B_n = b_n$. However, this type of condition is insufficient to identify a causal relationship, as is easy to verify in low-dimensionality cases. Two examples should be enough. The first one is about the *causal direction*. If we consider the simplest case, assuming $B = \emptyset$, the Method of difference is in no way effective in determining in which of the two directions the causal link is oriented. The second one is known as *selection bias*. Suppose for simplicity that X and Y are connected exclusively by a path which passes through a node $B_i \in B$. Let also B_i be fixed to a specific value b_i . It is possible to show that in the case where one has a collider, the (23) may turn out to be true regardless of whether X actually plays a causal role in determining the value of Y . Not only such a dependence turn out to be undesirable in the framework described by Pietsch, but it emerges *because* one has fixed the value of B_i .

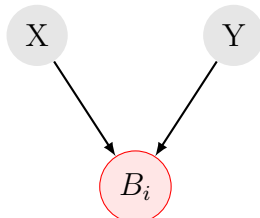


Figure 4: Collider. Fixing B_i introduces a dependency between X and Y .

Method of agreement. Mirroring what was done above, we fix the values of each individual variable in B , $B_1 = b_1, \dots, B_n = b_n$, saying that X is causally irrelevant for Y if

$$\Delta X \neq 0 \Rightarrow \Delta Y = 0. \quad (24)$$

As in the previous section, if we consider the purely observational level, the validity of the relation (24) can only be investigated by studying the dependence between variables X and Y under the assumption of conditioning on B . In other words, we must verify that the conditional independence condition is satisfied:

$$P(X, Y|B = b) = P(X|B = b)P(Y|B = b). \quad (25)$$

where $B = b$ means $B_1 = b_1, \dots, B_n = b_n$. However, this type of condition is insufficient to identify an existing causal relationship. All chains from X to Y containing nodes B are blocked by conditioning on B . Let us consider the simplest case, where a single node B_i blocks the chain from X to Y . Conditioning on B_i allows us to conclude that X is causally irrelevant to Y , despite the existence of a causal chain.

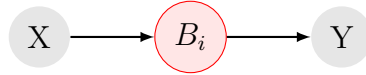


Figure 5: Mediator. Conditioning on B_i blocks the causal chain.

Locality. The topics presented so far involve elementary graphs. Someone might argue that these low-dimensional examples are irrelevant: Machine Learning algorithms typically deal with complex problems with many relevant variables. However, this objection is incorrect. A simple observation can clarify this point. Given variables X and Y , all paths connecting them via at least two other nodes are blocked once B is fixed. The only way to open a path with all nodes fixed is to have a collider between X and Y . Once the collider is placed, any choice of remaining arrows blocks the path.



Figure 6: Conditioning on B_1, B_2 blocks the path, whatever is the direction of the edge between B_2 and Y .

To summarize the discussion articulated so far, Machine Learning protocols represent an undeniable opportunity for scientific and technological progress, with significant consequences for many aspects of our lives, for better or worse. In our view, the conceptual framework provided by the modeling approach can help us correctly frame these opportunities, also highlighting some of their potential limitations. The theory of Dynamical Systems offers some indications, starting by analyzing a remarkably naive inductive protocol, the Analogs Method. A significant bottleneck - under rather general assumptions - lies in the unavailability of a sufficiently large database. The limit recognized at this first level can be relevant in a broader sense, and this point will be discussed in the following section. At any rate, taking Pietsch's objections and theoretical proposal into account, we have shown how a naive interpretation of the causality role is insufficient to circumvent the difficulties highlighted, nor can it offer a solid framework to justify on a methodological level the extraordinary functioning of specific inductive procedures. Although the integration of causal relations - as will be discussed in the last part of this dissertation - may help in improving Machine Learning protocols, Variational Induction does not seem to be the right path in this direction.

4 Representation and Learning

At a high level of analysis, two key issues can be considered in the context of Machine Learning methods. First, the *representation problem* involves determining whether there exists a set of free parameters - within a given architecture - that allows us to accomplish a specific task successfully. For example, given a complex input X we ask whether it is possible to provide a simpler representation $T = T(X)$ that allows us to recognize the correct corresponding output Y . Second, the *learning problem* consists of developing an automated procedure that can identify the representation mentioned above from a finite dataset, assuming its existence. Reconsidering the example above, we ask how to employ a given set of instances $\{x_1, x_2, \dots, x_n\}$ to identify the representation T with an appropriate level of accuracy.

Representation. In the classical theory of Statistical Learning, the role of representation is played by a space of functions \mathcal{F} defined a priori. Consequently, the problem arises of establishing which among these functions in \mathcal{F} best represents the regularity of the assigned dataset. This approach presents some criticalities - both conceptual and practical - as discussed in the Introduction of this dissertation. The following is aimed at showing how it is possible to obtain a generalization-gap bound that does not depend on the class of functions \mathcal{F} selected a priori. This approach involves a significant shift in perspective, arguing in favor of an analogy between learning and coarse-graining rather than the idea for which learning is a fitting procedure. In this regard, informational quantities will play a central role.

For simplicity, only the binary classification problem will be considered: given the i.i.d dataset $\mathcal{D} := \{(x_i, y_i)\}_{i=1}^n$ and $err[f(x), y] := \mathbb{1}_{\{f(x) \neq y\}}$, the expected error and empirical risk are defined as

$$R[f] := \mathbb{E}_{p(x,y)}[\mathbb{1}_{\{f(x) \neq y\}}],$$
$$\hat{R}_n[f] := \frac{1}{n} \sum_{\mathcal{D}} \mathbb{1}_{\{f(x_i) \neq y_i\}},$$

where $\mathbb{1}_A$ is the characteristic function on set A . In a nutshell, the following paragraphs are organized as follows. Starting from the classical bound for the $|\mathcal{F}|$ -finite case, the covering argument for the $|\mathcal{F}|$ -infinite generalization suggests introducing a covering on \mathcal{F} by identifying functions that agree on an appropriate subset of \mathcal{X} . The result obtained is a first step toward the informational interpretation by adopting the compression-generalization trade-off as a general paradigm. To elaborate on this point further, the notion of representation is explored, and the optimal case is formally characterized. Finally, the typicality argument is adopted to discuss the Large-scale Learning regime. This step will suggest drawing parallels with the predictability problem, for which an arbitrarily high-resolution description - at a too-fine observation scale - may be not only unnecessary but also misleading [32].

The classical bound. For the sake of clarity in view of the following steps, let us briefly review the classic *Probably Approximately Correct* (PAC) learning framework [106][155][181]. Fixed $f \in 2^{\mathcal{X}}$, the expected error is a deterministic quantity. On the contrary, the empirical risk is random and depends on the sample realization \mathcal{D} . Thanks to the i.i.d assumption the identity

$$\mathbb{E}_{\mathcal{D}}[\hat{R}_n[F]] = R[f]$$

holds and the empirical risk converges to a Gaussian variable centered in $R[f]$ with a variance of order $O(1/m)$, as the Central Limit theorem prescribes. Moreover, the *Chernoff-bound* allows us to control the probability of the Gaussian tails - i.e. the quantity to be minimized - via the inequality

$$\text{prob}_{\mathcal{D}}(|R[f] - \hat{R}_n[f]| > \varepsilon) \leq 2e^{-2n\varepsilon^2}. \quad (26)$$

At this point, relation (26) can not directly provide a generalization gap bound. Indeed, given the learning algorithm

$$\mathcal{A} : \mathcal{D} \mapsto f_{\mathcal{D}} \in \mathcal{F},$$

the function $f_{\mathcal{D}}$ is a random variable dependent on the dataset \mathcal{D} . Therefore, a uniform control in $f \in \mathcal{F}$ is needed. In this regard, the *union-bound* trick is a standard tool in providing the following relations [106]:

$$\begin{aligned} \text{prob}_{\mathcal{D}}(\exists f^* \in \mathcal{F} : |R[f^*] - \hat{R}_n[f^*]| > \varepsilon) &\leq \text{prob}_{\mathcal{D}}\left(\bigcup_{f \in \mathcal{F}} \{f : |R[f] - \hat{R}_n[f]| > \varepsilon\}\right) \\ &\leq \sum_{f \in \mathcal{F}} \text{prob}_{\mathcal{D}}(|R[f] - \hat{R}_n[f]| > \varepsilon) \\ &\leq 2|\mathcal{F}|e^{-2n\varepsilon^2}. \end{aligned}$$

Consequently - given the confidence threshold $\delta := 2|\mathcal{F}|e^{-2n\varepsilon^2}$ with $\delta \in (0, 1)$ - if the sample cardinality satisfies the inequality

$$n > \frac{\log |\mathcal{F}| + \log \frac{2}{\delta}}{2\varepsilon^2}, \quad (27)$$

the generalization gap is controlled by ε with probability $1 - \delta$ and uniformly in $f \in \mathcal{F}$. Remarkably - this time as f varies in \mathcal{F} - the empirical risk is expected to be a more irregular function than the expected error, the former being dependent on the specific dataset realization. In this regard, the uniform bound in f offers assurance about controlling the overfitting problem in minimizing the empirical risk. More precisely, when the cardinality $|\mathcal{D}|$ is sufficiently large with respect to $\log |\mathcal{F}|$, the empirical risk becomes a good estimator for the expected error simultaneously for all $f \in \mathcal{F}$, as in (3). In other terms,

$$R[f] \leq \hat{R}_n[f] + O\left(\sqrt{\frac{\log |\mathcal{F}|}{n}}\right) \quad (28)$$

and minimizing $\hat{R}_n[f]$ consequently becomes a good strategy to minimize $R[f]$.

The covering argument for $|\mathcal{F}| = \infty$. The union-bound trick provides a meaningful result only if $|\mathcal{F}| < \infty$ and the previous sample lower-bound is uninformative when dealing with an infinite function class. Generalizing (27) to $|\mathcal{F}| = \infty$ is possible by reducing the infinite case to the analysis of the finite one, adopting concepts as the *Rademacher complexity*, *grow function* or *VC-dimension* [106][155]. In what follows, we will discuss only the *covering argument* [155], although all these alternative approaches can be related to each other. As in the next paragraph, this perspective will also help determine a \mathcal{F} -independent bound. Let \mathcal{F} be a compact space with respect to the topology induced by the metric

$$\Delta(f_1, f_2) := \text{prob}_p(f_1(x) \neq f_2(x)).$$

Therefore, the number of ε -spheres needed to cover \mathcal{F} - denoted by $N(\varepsilon)$ - is finite and scales as $1/\varepsilon^d$, where d is the dimension of \mathcal{F} . Remarkably, the dimension d can be related to the concepts of VC-dimension, Hausdorff dimension, or topological dimension. In what follows, it will be thought of as related to the number of parameters needed to characterize an element f in \mathcal{F} . Moreover, for the binary classification problem, we have

$$R[f] = \text{prob}_p(f(x) \neq y)$$

and the following relation holds

$$|R[f_1] - R[f_2]| \leq \text{prob}_p(f_1(x) \neq f_2(x)) = \Delta(f_1, f_2). \quad (29)$$

In other words, if two functions belong to the same ε -sphere then the expected error of the former is approximated by the latter with an error at most ε . The same relationship is true also for the empirical risk by considering the empirical distribution \hat{p} instead of p . At this point - given the center of each ε -sphere - the learning algorithm $\mathcal{A} : D \mapsto f_D$ can be replaced by an approximate version

$$\mathcal{A}_\varepsilon : D \mapsto f_\varepsilon \in \mathcal{F}_\varepsilon,$$

where f_ε is the center of the ε -sphere in which f_D lives and $|\mathcal{F}_\varepsilon| < \infty$. Consequently, the substitution

$$|\mathcal{F}| \longrightarrow |\mathcal{F}_\varepsilon| := N(\varepsilon/2)$$

can be adopted in (27), obtaining a new lower-bound for n which depends linearly on d . Again, the main idea is the empirical risk concentration around the expected error via the Chernoff-bound. In this respect, the probability p comes into play only via the metric Δ . On the other side - once Δ is defined - the only object with a role is \mathcal{F} via its dimensionality. Moreover, for a given n the generalization gap is uniformly controlled with probability $1 - \delta$ as follows

$$|R[f] - \hat{R}_n[f]| \leq \sqrt{\frac{d \log(2/\varepsilon) + \log \frac{2}{\delta}}{2n}}, \quad (30)$$

which is meaningless in the Deep Learning context, for which $d \gg n$. This latter observation provides a relevant problem for Statistical Learning theory. Deep Learning models work with much fewer data than the classical bound would predict, also when the number of parameters exceeds the dataset size by orders of magnitude.

\mathcal{F} -independent bound via covering. The classical bounds (3), (27), (28) and (30) as discussed above require a class of function \mathcal{F} - or \mathcal{F}_ε - narrower than $2^{\mathcal{X}}$. On the contrary, if $\mathcal{F} = 2^{\mathcal{X}}$, the inequality (27) would require a number of data on the order of all possible realizations of X . In that case, the generalization-gap control is meaningless. More in general, the classical PAC approach provides bounds that do not depend on the specific learning procedure adopted or sample distribution, thus characterized by the worst-case behavior. Remarkably, the previous paragraph offers a useful tool to rediscuss this point, with the difference that we assume $|\mathcal{X}| < \infty$ to obtain a bound independent of \mathcal{F} , and taking into account the properties of the probability distribution p . More in detail, a covering argument will be adopted by constructing an *effective partition* on $2^{\mathcal{X}}$ by identifying functions that agree on appropriate sets. Let us proceed step by step. Firstly, the following objects are defined as represented in Figure 7:

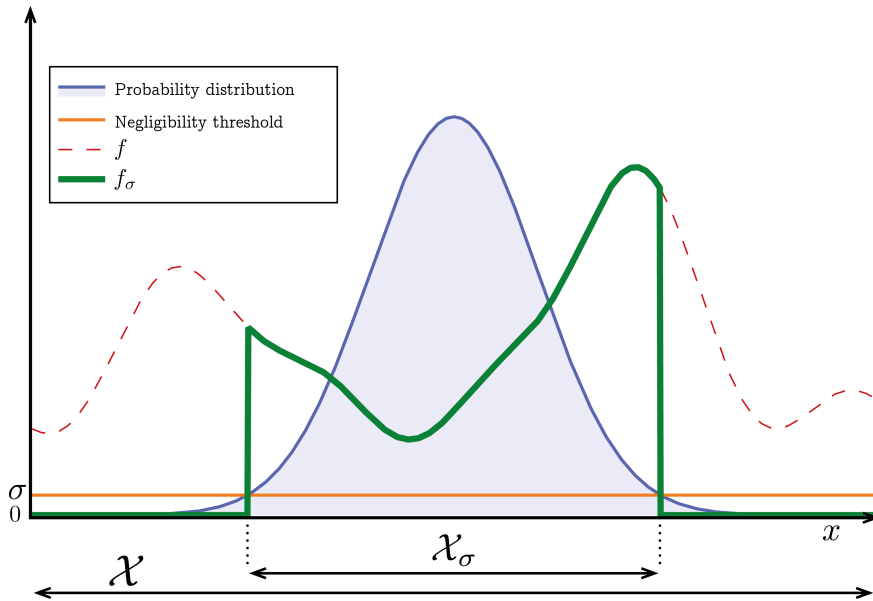


Figure 7: *Non-negligible patterns \mathcal{X}_σ and quotient on \mathcal{F} .* The procedure is schematically illustrated. For clarity, the above functions are continuous with \mathbb{R} -values.

$$\mathcal{X}_\sigma := \{x \in \mathcal{X} : p(x) > \sigma\} \quad \text{with } \sigma \in (0, 1),$$

$$\mathcal{F}_\sigma := \{f_\sigma : f \in 2^{\mathcal{X}}\},$$

$$f_\sigma(x) := \begin{cases} f(x) & \text{if } x \in \mathcal{X}_\sigma \\ 0 & \text{o.w.} \end{cases}.$$

In other words, the set of *non-negligible patterns* \mathcal{X}_σ is introduced by setting a threshold σ on the probability of each pattern. Moreover, the functions that agree on \mathcal{X}_σ are identified by renouncing in characterizing them for negligible patterns: hopefully, this choice enables the adjustment of worst-case behavior, characteristic of the PAC approach. Secondly, the generalization gap is decomposed into three pieces as an immediate consequence of the triangular inequality and the union-bound trick:

$$\begin{aligned} \text{prob}_{\mathcal{D}}(|R[f] - \hat{R}_n[f]| > \varepsilon) &\leq \text{prob}_{\mathcal{D}}\left(|R[f] - R[f_\sigma]| > \frac{\varepsilon}{3}\right) \\ &\quad + \text{prob}_{\mathcal{D}}\left(|R[f_\sigma] - \hat{R}_n[f_\sigma]| > \frac{\varepsilon}{3}\right) \\ &\quad + \text{prob}_{\mathcal{D}}\left(|\hat{R}_n[f_\sigma] - \hat{R}_n[f]| > \frac{\varepsilon}{3}\right) \end{aligned}$$

At this point, the three terms reported above can be separately considered. The general idea is to determine the threshold σ to set the first term to zero and simultaneously control the other two. Let us consider the first term by employing (29) and the *Markov inequality* as follows:

$$\begin{aligned} |R[f] - R[f_\sigma]| &\leq \text{prob}_p(f(x) \neq f_\sigma(x)) = \text{prob}_p(x \notin \mathcal{X}_\sigma) \\ &= \text{prob}_p\left(-\log p(x) > \log \frac{1}{\sigma}\right) \\ &\leq \frac{\mathbb{E}_p[-\log p(x)]}{\log \frac{1}{\sigma}} = \frac{H[X]}{\log \frac{1}{\sigma}}, \end{aligned}$$

where the Shannon Entropy $H[X]$ is assumed to be finite. Consequently, by defining

$$\frac{H[X]}{\log \frac{1}{\sigma}} := \varepsilon'$$

and choosing σ such that $\frac{\varepsilon}{3} \geq \varepsilon'$,

$$\text{prob}_{\mathcal{D}}\left(|R[f] - R[f_\sigma]| > \frac{\varepsilon}{3}\right) = 0.$$

The second term includes only $f_\sigma \in \mathcal{F}_\sigma$. Consequently, the classical bound holds. More precisely - observing that $|\mathcal{X}_\sigma| \leq 1/\sigma$ and adopting σ as above - the cardinality $|\mathcal{F}_\sigma|$ is controlled as follows

$$|\mathcal{F}_\sigma| = 2^{|\mathcal{X}_\sigma|} \leq 2^{\frac{1}{\sigma}} \leq 2^{2^{\frac{H[X]}{\varepsilon'}}$$

and then

$$\begin{aligned} \text{prob}_{\mathcal{D}}\left(|R[f_\sigma] - \hat{R}_n[f_\sigma]| > \frac{\varepsilon}{3}\right) &\leq 2|\mathcal{F}_\sigma|e^{-2n\varepsilon'^2} \\ &\leq 2^{2^{\frac{H[X]}{\varepsilon'}+1}}e^{-2n\varepsilon'^2}. \end{aligned}$$

For the third term, the inequality (29) with the empirical distribution \hat{p} allows us to use the standard Chernoff-bound again. More precisely - as already done for p - we have

$$|\hat{R}_n[f] - \hat{R}_n[f_\sigma]| \leq \text{prob}_{\hat{p}}(f(x) \neq f_\sigma(x)) = \text{prob}_{\hat{p}}(\hat{p}(x) < \sigma).$$

Consequently,

$$\begin{aligned} \text{prob}_{\mathcal{D}} \left(|\hat{R}_n[f_\sigma] - \hat{R}_n[f]| > \frac{\varepsilon}{3} \right) &\leq \text{prob}_{\mathcal{D}} \left(\text{prob}_{\hat{p}}(\hat{p}(x) < \sigma) > \frac{\varepsilon}{3} \right) \\ &\leq \text{prob}_{\mathcal{D}} \left(\text{prob}_{\hat{p}}(\hat{p}(x) < \sigma) - \mu > \frac{\varepsilon}{3} - \mu \right) \\ &\leq e^{-2n(\frac{\varepsilon}{3} - \mu)^2}, \end{aligned}$$

where μ is defined as

$$\mu := \mathbb{E}_{\mathcal{D}}[\text{prob}_{\hat{p}}(\hat{p}(x) < \sigma)] = \text{prob}_p(x \notin \mathcal{X}_\sigma).$$

Remarkably, $\mu \leq \varepsilon' \leq \frac{\varepsilon}{3}$. Therefore - choosing σ such that $\varepsilon' = \varepsilon/6$ and putting it all together - the generalization gap is controlled by ε with probability $1 - \delta$ and uniformly in $f \in 2^{\mathcal{X}}$ if n scales - unless a multiplicative constant - as

$$n > \frac{2^{\frac{6H[X]}{\varepsilon}} + \log \frac{2}{\delta}}{\varepsilon^2} \quad (31)$$

The relation (31) is independent of the function class \mathcal{F} , i.e. a spurious element in a properly inductive scenario. Remarkably, $\log |\mathcal{F}|$ can be interpreted as the number of bits needed to represent \mathcal{F} so that bounds (3), (27) and (30) can consequently be interpreted in terms of the *Occam's Razor* principle. On the other side, (31) requires two general remarks.

- First, this outcome shall provide a first insight into how a *more compressed* representation \mathcal{X}_{eff} of \mathcal{X} can enhance the efficacy of the inductive protocol adopted: reducing $H[X]$ exponentially decreases the number of data required to achieve the desired control on the generalization gap. In what follows, this point will be clarified further.
- On the other hand, the exponential dependence on $1/\varepsilon$ in (31) introduces a significant difficulty: for small ε the dataset size n required for learning becomes unreasonably large, regardless of how small the entropy $H[X]$ is. Remarkably, factor $1/\varepsilon$ appears by construction from the definition of the threshold σ , used to define the set \mathcal{X}_σ of *non-negligible* patterns.

To elaborate on this point further, it will be necessary to consider Information Theory as a valuable tool. In this regard, the following paragraph collects some standard definitions and elementary relationships, and the main references can be found in [157][98][43]. After concluding the following parenthesis, we will reconsider the role of representation and rediscuss the results presented above.

Elements of Information Theory. Let X be a random variable with discrete range \mathcal{X} and probability distribution $p(x)$. Except for a multiplicative constant, $h(x) := -\log p(x)$ can be interpreted as a measure of the information content associated with the occurrence $x \in \mathcal{X}$. Averaging over \mathcal{X} , the *Shannon entropy* $H[X]$ is then defined as

$$H[X] := - \sum_{x \in \mathcal{X}} p(x) \log p(x). \quad (32)$$

Remarkably, $H[X]$ can be axiomatically derived by defining three reasonable properties of uncertainty measures. More precisely, $H[X]$ is the only function that satisfies the *continuity*, *additivity*, and *monotonicity* requirements [157][43]. Entropy can also be read as the minimum description length of X : it corresponds to the minimal number of bits or binary questions needed - on average - to determine the value of X . In this respect, $H[X]$ is a concave function in p that reaches its maximum $\log |\mathcal{X}|$ when p is uniform on \mathcal{X} , and the outcome of a random experiment is guaranteed to be the most informative as possible. Conversely, $H[X] = 0$ when X is deterministic. Moreover, (32) can be extended to the *joint* and *conditional* cases via

$$H[X, Y] := - \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y) \log p(x, y),$$

$$H[X|Y] := \sum_{y \in \mathcal{Y}} p(y) H[X|Y = y] := - \sum_{y \in \mathcal{Y}} p(y) \sum_{x \in \mathcal{X}} p(x|y) \log p(x|y),$$

which admit analogous interpretations. By adopting the definitions mentioned above, the next relationships follow immediately and lead back to the information additivity when X and Y are independent:

$$H[X, Y] = H[X] + H[Y|X] = H[Y] + H[X|Y].$$

Clearly, conditioning allows us to isolate the information shared by X and Y . In operational terms, given a particular input-output couple - i.e. X and Y - a measure of how much information the output Y conveys about the input X is desired. In this respect, *mutual information* $I(X : Y)$ is defined as

$$I(X : Y) := H[X] - H[X|Y] = H[Y] - H[Y|X],$$

and can be interpreted as the mean information attributed to a realization of X given complete knowledge of Y , or vice versa. Remarkably, mutual information is a useful index for measuring statistical dependence between variables and can be derived axiomatically [43]. Let us dwell on this point, also giving some definitions useful for what follows. First of all, the equality $I(X : Y) = H[X] + H[Y] - H[X, Y]$ provides us with the more explicit formula that follows

$$I(X : Y) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}, \quad (33)$$

to be read as a special case of the *Kullback-Leibler divergence*. More precisely, suppose we have a priori a distribution p for X and consider the distribution q in place of p ,

given a certain set of experimental measures. Under the assumption that $q(x) = 0$ implies $p(x) = 0$, we can define the *KL divergence*

$$D[p||q] := \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)},$$

to be interpreted as the information gap determined - on average - by using q instead of p . In the case in which $p(x) = q(x)$ for some x , then it will not contribute to $D[p||q]$. By employing *Jensen's inequality* for convex functions, *Gibbs' inequality* immediately follows [98]:

$$D[p||q] \geq 0,$$

where the equality holds if and only if $p \equiv q$. As a corollary - as mentioned just above - $I(X : Y) = 0$ if and only if X and Y are independent. Moreover, p and q are similar when they have a low KL divergence, while a high KL divergence indicates dissimilarity. In this regard, *Pinsker's inequality* is a standard tool and implies that the KL divergence upper bounds the L^1 -distance between distributions [43]:

$$\|p - q\|_1^2 \leq \frac{1}{2 \ln 2} D[p||q]. \quad (34)$$

Consequently - although KL divergence is not a metric because it is not symmetric and the triangle inequality does not hold - it is still useful to think of $D[p||q]$ as a natural distance between p and q . We conclude these recalls by mentioning the *data processing inequality* for mutual information [43], which implies that the information about X contained in Y cannot be increased by processing Y . Formally, X , Y , and Z form a Markov chain

$$X \longrightarrow Y \longrightarrow Z$$

if their joint distribution can be decomposed as $p(x, y, z) = p(z|y)p(y, x)$, so that Z is a function of Y , and independent of X given Y . Then, the data processing inequality states that

$$I(X : Z) \leq I(X : Y). \quad (35)$$

Remarkably, mutual information is a fundamental quantity within communication theory, allowing us to define the concept of *channel capacity*, characterizing the expected maximal number of bits that can be reliably sent in a discrete memoryless channel with a probability transition $p(y|x)$ [43][98]. On the other side, it remains invariant under bijective transformations [43]: given two bijections ϕ_1 and ϕ_2 ,

$$I(\phi_1(X), \phi_2(Y)) = I(X : Y).$$

Consequently, mutual information might seem an unsuitable indicator for a given representation in the context of learning. Instead - in what follows - mutual information will enable us to give a rigorous index for the *compression* and *accuracy* levels - respectively $I(X : T)$ and $I(T : Y)$ - of a given representation T .

Compression-in-representation. Machine learning protocols rely on internal input-based representations to enhance their performance [18][59][7]. While determining an appropriate representation can be essential for numerous practical domains, the precise criteria for establishing a satisfactory representation - as well as the relationship between representation and the learning process - continues to pose significant challenges [146]. From a general viewpoint, *representation learning* refers to recognizing data structures by transforming the high-dimensional input space \mathcal{X} into a lower-dimensional one \mathcal{T} , more effectively achieving the desired accuracy for the task at hand. In this regard, most of the input's entropy $H[X]$ turns out to be irrelevant, and the main challenge consists of extracting the relevant features to construct T . Schematically, the learning apparatus can be seen as a Markov chain

$$Y \longleftrightarrow X \longrightarrow T \longrightarrow \hat{Y}, \quad (36)$$

where the representation T is selected via the training phase to obtain the assigned label \hat{Y} as close as possible to the correct label Y . Remarkably, two different and separate components can be considered in this regard, although these may occur interdependently during the training phase:

- First, the *feature extraction* component corresponds to the encoding procedure represented by the chain arrow $X \longrightarrow T$, for which the *encoder* $p(t|x)$ must be determined. In this regard, the mutual information $I(X : T)$ quantifies the level of information shared between representations X and T , reducing to $H[T]$ when T is characterized deterministically by X .
- Second, the *label estimation* component corresponds to the decoding procedure represented by the chain arrow $T \longrightarrow \hat{Y}$, for which the *decoder* $p(y|t)$ should be as close as possible to the *optimal decoder*

$$p_{opt}(y|t) := \sum_{x \in \mathcal{X}} p(y|x)p(x|t). \quad (37)$$

The discussion so far has been concerned only with this second stage, without considering an effective representation and assuming the available one \mathcal{X} to be independent of the particular training set employed.

Let us start by assuming that the representation T is fixed, i.e. the encoder $p(t|x)$ is assigned. The mutual information $I(Y : T)$ can be interpreted as an accuracy index. To make this point clear, it is sufficient to consider the following relations, which are obtained by using Pinsker's inequality (34) and considering the optimality condition (37):

$$\begin{aligned} \mathbb{E}_{p(x,t)} [||p(y|x) - p(y|t)||_1^2] &\leq \frac{1}{2 \ln 2} \mathbf{E}_{p(x,t)} [D[p(y|x)||p(y|t)]] \\ &= \frac{1}{2 \ln 2} \{I(X : Y) - I(Y : T)\}, \end{aligned}$$

where mutual information for the problem distribution $p(x, y)$ - i.e $I(X : Y)$ - does not depend on the particular learning procedure. Consequently, to obtain an optimal

decoder $p(y|t)$ nearer to the original distribution $p(y|x)$, $I(Y : T)$ should increase regardless of the learning protocol adopted. On the other hand, to clarify how different representations can intervene in improving (31), let us consider \mathcal{T} as a partition of \mathcal{X} for which a *homogeneity condition* hold: all instances $x \in \mathcal{X}$ contained in the same $t \in \mathcal{T}$ share the same label y . Assuming the optimality condition (37), the expected error and empirical risk can be re-written as

$$\begin{aligned} R[f] &= \int err[f(x), y] p(x, y) dx dy = \int err[f(x), y] \int p(x, y, t) dx dy dt \\ &= \int err[f(t), y] \left\{ \int p(y|x) p(x|t) dx \right\} p(t) dy dt \\ &= \int err[f(t), y] p(t, y) dt dy, \end{aligned}$$

$$\hat{R}_n[f] = \frac{1}{n} \sum_{\mathcal{D}} err[f(x_i), y_i] = \frac{1}{n} \sum_{\mathcal{D}} err[f(t_i), y_i].$$

At this point, the argument presented above to obtain the \mathcal{F} -independent bound can be directly applied by replacing $H[X] \mapsto H[T]$ in (31). However, observing that Y is independent of T given X and weakening the homogeneity condition by introducing a soft partition via a non-deterministic $p(t|x)$ - we expect that bound (31) can be re-written as

$$n \sim \frac{2^{\frac{6I(X:T)}{\varepsilon}} + \log \frac{2}{\delta}}{\varepsilon^2}, \quad (38)$$

so concluding that this sample cardinality should be sufficient to control the generalization gap via ε . The estimate (38) suggest that - fixed a desired level of accuracy $I(Y : T)$ - we expect that a more compressed representation \mathcal{T} - as quantified by the mutual information $I(X : T)$ - improves the learning performance. At this point, two comments are needed.

- Although estimate (38) has not been rigorously proved, we will find a similar result when considering the typicality argument. Again, we will assume the T representation fixed and neglect *non-typical* patterns instead of negligible ones.
- The informational quantities considered so far are computable only if the distribution $p(x, y, t)$ is available. On the other hand, assuming only the empirical distribution $\hat{p}(x, y)$ and the encoder $p(t|x)$ are available - similarly to what was done for the theoretical distribution - it is possible to show that

$$\begin{aligned} \mathbb{E}_{\hat{p}(x,t)} [\|\hat{p}(y|x) - p(y|t)\|_1^2] &\leq \frac{1}{2 \ln 2} \mathbf{E}_{\hat{p}(x,t)} [D[\hat{p}(y|x) \| p(y|t)]] \\ &= \frac{1}{2 \ln 2} \{ I(X : Y) - \hat{I}(Y : T) \}, \end{aligned}$$

where the empirical quantity $\hat{I}(Y : T)$ plays the role of the empirical risk in the

classical PAC bounds setting via [156][179]

$$I(Y : T) \leq \hat{I}(Y : T) + O\left(\sqrt{\frac{2^{I(X:T)}}{n}}\right). \quad (39)$$

If (28) is considered, the analogy with the relation above is clear, and selecting the decoder $p(y|t)$ to minimize the empirical quantity $\hat{I}(Y : T)$ consequently becomes a good strategy to minimize $I(Y : T)$. Remarkably, whereas control in (28) improves only logarithmically in $|\mathcal{F}|$, a reduction in $I(X : T)$ provides an exponential improvement in (39).

Remarkably, the compression index $I(X : T)$ can always be reduced by ignoring further details in \mathcal{X} . An essential point in this regard is precisely to establish a link between compression and effects in terms of accuracy. Therefore, if the encoder $p(t|x)$ is not provided a priori but also the representation learning question is introduced, additional constraints on T are needed. The following paragraph will discuss this point by adopting the *Information Bottleneck Principle* perspective [178].

Information Bottleneck Principle. Let T be a random variable to be considered as in the learning apparatus (36). The encoder $p(t|x)$ induces a soft partition of \mathcal{X} , for which each value $x \in \mathcal{X}$ is associated with all $t \in \mathcal{T}$ via $p(t|x)$. In our intentions, T should be a compressed representation of X so that $|\mathcal{T}| \leq |\mathcal{X}|$ and $p(t|x)$ results appropriately concentrated. In this regard, $I(T : X)$ provides a compactness index for the representation T , although additional prescriptions are needed to evaluate the representation adequacy. In this regard, the *rate-distortion theory* offers a first type of *relevance criteria* by providing a metric

$$\Delta : \mathcal{X} \times \mathcal{T} \longrightarrow \mathbb{R}^+,$$

assuming that smaller values of $\Delta(x, t)$ imply a better representation \mathcal{T} [43][32]. Let us briefly consider this approach, preparatory to the following discussion. In a nutshell, the partition of \mathcal{X} induced by the encoder $p(t|x)$ corresponds to an expected distortion $\mathbb{E}_{p(x,t)}[\Delta(x, y)]$ and a trade-off with the representation compactness is naturally established via the constrained optimization of $I(X : T)$, so that

$$R(D) := \min_{\{p(t|x) : \mathbb{E}_{p(x,t)}[\Delta(x,y)] \leq D\}} I(X : T).$$

This problem can be re-formulated by introducing the Lagrange multiplier β and minimizing the Lagrangian functional

$$\mathcal{L}[p(t|x)] := I(X : T) + \beta \mathbb{E}_{p(x,t)}[\Delta(x, y)] \quad (40)$$

under the additional constraint $\sum_t p(t|x) = 1$ for all $x \in \mathcal{X}$. The variational problem $\frac{\delta \mathcal{L}}{\delta p(t|x)} = 0$ admits the implicit solution

$$\begin{cases} p_\beta(t|x) = \frac{p_\beta(t)}{Z(x, \beta)} e^{-\beta \Delta(x,t)} \\ p_\beta(t) = \sum_{x \in \mathcal{X}} p_\beta(t|x) p(x) \end{cases}$$

where $Z(x, \beta)$ is a normalization function and $\beta \geq 0$ satisfies

$$\beta = -\frac{\delta R}{\delta D}.$$

Clearly, $R(D)$ is defined with respect to a fixed set of representatives \mathcal{T} : given different values of $|\mathcal{T}|$, different distortion matrices $[\Delta(x, t)]_{x,t}$ are consequently defined, changing $R(D)$. Remarkably, the rate-distortion $R(D)$ provides a monotonic convex curve with slope $-\beta$ in the (D, I) -plane [43], to be determined numerically by choosing different values of β and applying the iterative *Blahut-Arimoto* procedure [165]. Theoretically, when β increases, minimizing the expected distortion becomes more and more relevant in (40), up to $\mathbb{E}_{p(x,t)}[\Delta(x, y)] = 0$ when $I(T : X) = H[X]$. On the other hand, the compression level matters more and more when β decreases, up to $I(T : X) = 0$ for a certain value of the distortion onward. At this level, two general comments are relevant to our discussion.

- First of all, the curve obtained provides two regions in the (D, I) -plane, establishing a limitation result. The region above the curve corresponds to all the achievable distortion-compression pairs, while the region below is in principle not achievable. In other words, given a distortion-compression pair (D^*, I^*) , there exists an encoder $p(t|x)$ for which $I(X : T) = I^*$ and $\mathbb{E}_{p(x,t)}[\Delta(x, y)] = D^*$ if and only if (D^*, I^*) is above the curve, on which the optimal encoder is placed.
- Secondly, the characterization of $R(D)$ relies on the spurious element Δ , which introduces an arbitrary and difficult-to-justify choice. In this regard - for a given $p(x)$ - different choices of Δ will yield different results and alternative rate-distortion curves so that selecting the appropriate one is far from trivial in many practical applications. In other words, the main drawback of the rate-distortion approach consists of considering Δ as a part of the problem.

The Information Bottleneck Principle [178][166][52] provides an alternative approach to overcome the abovementioned difficulties. Instead of considering only $p(x)$, the joint distribution $p(x, y)$ comes into play. Consequently, the following strategy can be introduced: determining a compressed representation of X that preserves the information on Y as high as possible. Formally, the optimal encoder is characterized by the problem

$$\min_{p(t|x)} I(X : T) - \beta I(T : Y), \quad (41)$$

where β corresponds to a resolution parameter that controls the compression level provided by T : small β implies more compression at the expense of informativeness, while bigger β corresponds to finer granularity in representation. The variational problem (41) admits the implicit solution

$$\begin{cases} p_\beta(t|x) = \frac{p(t)}{Z(x, \beta)} e^{-\beta D[p(y|x)||p_\beta(y|t)]} \\ p_\beta(t) = \sum_{x \in \mathcal{X}} p_\beta(t|x)p(x) \\ p_\beta(y|t) = \sum_{x \in \mathcal{X}} p(y|x)p_\beta(x|t) \end{cases}$$

where the KL-divergence $D[p(y|x)||p(y|t)]$ naturally assumes the role of the distortion measure [63] and

$$\beta = \frac{\delta I(X : T)}{\delta I(T : Y)}.$$

Similarly to the case of distortion - assuming that $p(x, y)$ is known - the three relations above determine self-consistently the optimal solution, and the alternating iterations of the implicit formula above converge to a local solution [165]. In contrast to rate distortion theory - where the representative selection is a separate problem - the optimization is also over the cluster representatives $p(y|t)$. If the decoder $p_\beta(y|t)$ is sufficiently near to $p(y|x)$, the encoder $p_\beta(t|x)$ is concentrated around the good t . In this regard, we have a complete correspondence with a *clustering problem* [108][165][166].

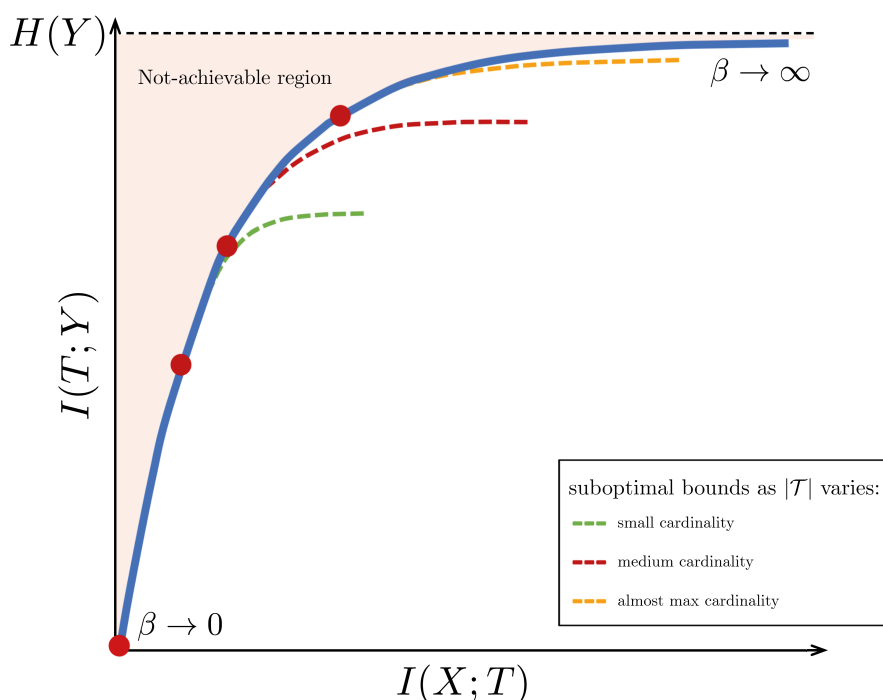


Figure 8: *Information Bottleneck curve in the Information Plane*. As β varies, a monotonic concave curve is provided. Only the region below the curve is achievable. Fixing the representation cardinality $|\mathcal{T}|$ provides sub-optimal curves. The optimal representation T for $p(x, y)$ is characterized by the compression-accuracy couple, regardless of the learning procedure adopted.

Recalling and mimicking what has already been discussed for $R(D)$, the *Information Bottleneck curve* is obtained by solving (41) as $\beta \geq 0$ varies and plotting the mutual information pair $(I_\beta(X : T), I_\beta(Y : T))$ given the optimal encoder $p_\beta(t|x)$. This curve in the *Information Plane* - i.e. the plane with $I(X : T)$ and $I(Y : T)$ as the x-axis and y-axis respectively [178][179][158] - is concave and monotonically

increasing, as shown in Figure 8. On the theoretical level - in contrast with the rate-distortion approach - the Information Bottleneck method is based purely on $p(x, y)$. Once the distribution $p(x, y)$ is available, the theoretical framework is completed. The optimal representation T is completely determined via a general principle without introducing additional spurious elements such as the distortion measure. In other words, the relevance criterion needed to encode X - i.e. to extract the relevant features - is assigned implicitly by the problem formulation through the second variable Y . Remarkably, the explicit calculation of the solution is generally prohibitive. In this regard, the Gaussian case is a significant exception, and its solution is strongly related to the *Correlated Component Analysis* [34][56].

Expanding on the observation above, the minimization as in (41) does not pertain to a particular learning procedure but refers to the possibility of an appropriate representation in general, in style inaugurated with Shannon's work with respect to the coding-decoding process for the transmission in a noisy channel. Moreover, this high-level principle can be central in characterizing intelligence in a broader sense [191], assuming the availability of an effective representation for carrying out a pre-determined task as an essential point for intelligent behavior. Similarly, it is also adopted as a general tool to characterize the forecasting problem, compressing away information about the past not useful for predicting the future [44].

The discussion considered so far requires that the distribution $p(x, y)$ is available. Conversely, whenever only a finite sample \mathcal{D} is given, the information curve previously introduced exhibits a different behavior, as shown in Figure 9. More precisely:

- If $I(X : T)$ is too small, *over-compression* comes into play, and a too much coarse representation compromises the accuracy $I(T : Y)$. In classical Statistical Learning terms, this situation would correspond to the underfitting case: a too-high compression level for \mathcal{T} produces the same effect as selecting a too-limited class of function \mathcal{F} to fit \mathcal{D} adequately.
- Conversely, if $I(X : T)$ is too near to $H[X]$, *under-compression* takes place, for which a too-detailed representation coexists with the sparsity of data, again compromising the accuracy level. In classical Statistical Learning terms - as when a too-large class \mathcal{F} is selected - the overfitting problem occurs.

Between the two cases mentioned above, there exists an optimal level of compression for which the highest possible level of accuracy is achievable in the finite-data regime. In this regard, the estimate (39) provides an indication of how both dataset size m and compression level can intervene in controlling the generalization gap. Having clarified in what terms it is possible to characterize the concept of effective representation by adopting the compression-generalization trade-off as a general paradigm, let us now return to the learning problem by taking a Large-scale perspective, where the number of training examples and the input dimension are both huge. As highlighted by N. Tishby [179][158][159], this approach provides some insights into explaining the

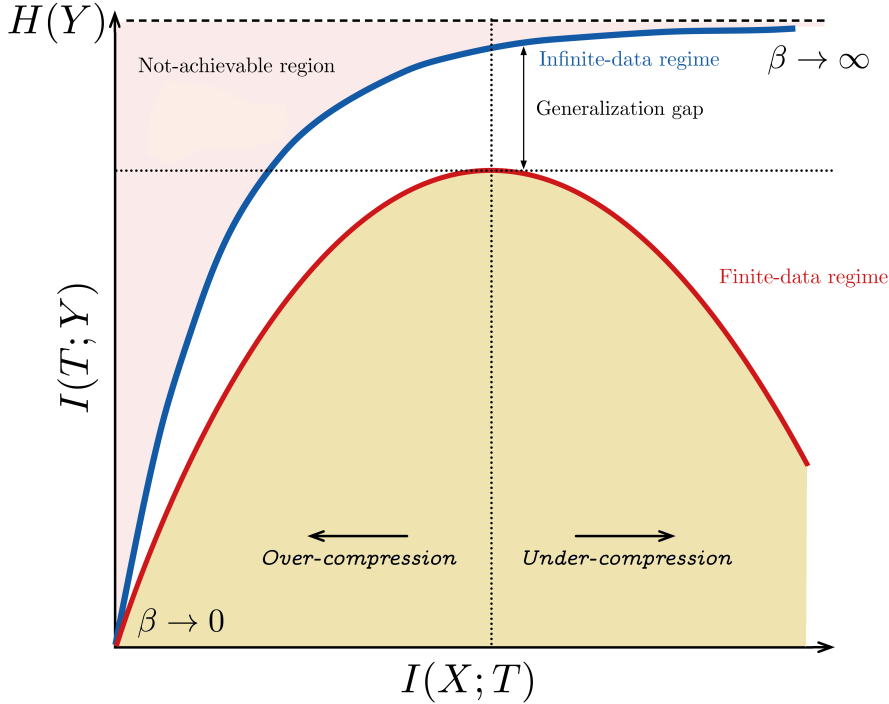


Figure 9: *Information Bottleneck curve in the finite-data regime*, adapted from [179]. Only encoders living in the yellow area can be learned from the finite dataset \mathcal{D} . Remarkably, the over-compression and under-compression regimes correspond - in the classical Statistical Learning language - to underfitting and overfitting conditions respectively.

functioning of Deep Learning protocols as procedures that provide compression-in-representation, shedding light on the overfitting problem and the benefits of adopting a layered structure. Moreover, some points of contact with modeling are considered, suggesting an analogy with coarse-graining techniques in developing an effective model.

Bound via typicality for Large-Scale Learning. The need to introduce the class of functions \mathcal{F} is so far circumvented by introducing an effective partition \mathcal{F}_σ constructed by identifying functions that agree on the non-negligible patterns \mathcal{X}_σ . In this regard, σ is the negligibility threshold related to the control parameter ε , and by adjusting the σ parameter appropriately an enough accurate partition is provided to have a generalization gap controlled by ε . This procedure - for which the assigned encoder $p(t|x)$ is a priori given - makes a factor $1/\varepsilon$ appear at the exponent by construction. In this paragraph, a different perspective is considered, for which the notion of information plays a more essential role. N. Tishby has proposed the following argument to justify Deep Learning's ability to generalize [179][158][159]. If the Large Scale Learning regime - where the number of training examples and the input dimension are both huge - is considered it is possible to introduce a typical-case

bound obtained by restricting the probability distributions to a suitable class and abandoning the worst-case perspective. Let us start by considering the simplest case, for which the probability distribution P_X is factorized as

$$P_X = \prod_{i=1}^m P_{X_i},$$

where P_{X_i} refers to a single bit of X and $P_{X_j} = P_{X_k}$ for all j and k . Defining the set of ε -typical input of size m , i.e.

$$\Lambda_\varepsilon^m[\mathcal{X}] := \left\{ x \in \mathcal{X} : \left| \frac{1}{m} \log \frac{1}{p(x)} - H[X] \right| < \varepsilon \right\}$$

the Chernoff-type inequality

$$\text{prob} \left(\left| \frac{1}{m} \log \frac{1}{p(x)} - H[X] \right| \geq \varepsilon \right) < e^{-2m\varepsilon^2}$$

ensures that nontypical patterns are negligible in probability for $m \rightarrow \infty$. Consequently, considering $\Lambda_\varepsilon^m[\mathcal{X}]$ instead of \mathcal{X} is justified. In other words, the Asymptotic Equipartition Principle holds. In this regard, some comments are needed.

- First of all, the equipartition holds only in a weak sense. Typical patterns are similar in probability only because their values of $-\log p(x)$ are within $2m\varepsilon$ of each other. As ε is decreased, m must grow as $1/\varepsilon^2$. If we write

$$\varepsilon \sim \frac{1}{\sqrt{m}},$$

then the most probable string in the typical set will be of order $2^{C\sqrt{m}}$ times greater than the least probable one, for some fixed C . On the other side, the typical set introduces a considerable simplification. Its elements have almost identical probability $2^{mH[X_k]}$, and the whole set has a probability of almost 1. Consequently, we have roughly $2^{mH[X_k]}$ elements, and the other ones have no relevant role in probability.

- Secondly, the i.i.d assumption on $X = (X_1, X_2, \dots, X_m)$ can be relaxed by requiring that a *Markov field* structure is satisfied. The general idea is based on the assumption that if the distribution P_X is factorized into many components, the Asymptotic Equipartition Principle holds. In this regard, let us assume that the probability measure P_X is factorized as

$$P_X = \prod_{i=1}^m P_{X_i|Pa_i},$$

where Pa_i denotes the X -components adjacent to X_i , for which the following independence condition holds

$$X_i \perp \neg Pa_i | Pa_i.$$

If this Markov random field is ergodic, the Shannon-McMillan-Breiman theorem comes into play, and typical patterns approximate the entire patterns space \mathcal{X} with high probability [43][81]. Remarkably, this *factorization hypothesis* seems reasonable in many application contexts - e.g. in image processing and speech recognition - where patterns comprise many local and weakly-dependent patches.

As previously discussed, the main idea in the classical PAC framework is to exploit the concentration in probability for the empirical risk around the expected error due to the Chernoff-bound effect. Instead - at this point - the typicality condition provides a new concentration in probability naturally applicable to representation. Let T be a mapping of X , as in the learning apparatus (36). Ignoring the not-typical patterns, only $2^{H[X]}$ realizations of X can be considered. Moreover - on average - $2^{H[X|T]}$ typical realizations of X are mapped to the same value of T . Mimicking the classic argument to justify the noisy channel coding theorem [98], the cardinality of the typical realizations of T can be estimated as

$$\frac{2^{H[X]}}{2^{H[X|T]}} = 2^{I(X:T)}.$$

Consequently - with some approximation - it is sufficient to consider the classification rules f defined for the representation of typical patterns, i.e. restricting ourselves to a proper subset $\mathcal{F}_T \subset 2^{\mathcal{X}}$. In other words, the substitution

$$|\mathcal{F}| \mapsto |\mathcal{F}_T| = 2^{2^{I(X:T)}}$$

can be adopted in (27) obtaining a new lower-bound for the dataset cardinality to have the desired generalization control ε with probability $1 - \delta$:

$$n > \frac{2^{I(X:T)} + \log \frac{2}{\delta}}{2\varepsilon^2}. \quad (42)$$

At this level, two comments are needed.

- The bound (42) indicates that a more compressed representation exponentially reduces the amount of data needed to control the generalization gap via ε . However, it is necessary to keep in mind that the Information Bottleneck curve controls this scenario. If $I(Y : T)$ is small, many possible representations with the same $I(X : T)$ are in principle available: the main point is to have an informative representation T on Y , and the estimate (42) is consequently significant when $I(Y : T)$ is sufficiently large.
- Reconsidering the expression (28) and introducing the set of effective functions $\mathcal{F}_T \subset 2^{\mathcal{X}}$ obtained via the representation T , the following relation follows

$$R[f] \leq \hat{R}_n[f] + O\left(\sqrt{\frac{2^{I(X:T)}}{n}}\right),$$

becoming significant when $I(X : T)$ is smaller than $O(\log n)$, and in this case k bits of compression are equivalent to an exponential factor of 2^k training examples.

At any rate, it remains to be clarified how a learning protocol can provide an appropriately compressed representation T starting with a finite dataset. In the following paragraph, this point is briefly discussed by considering Deep Learning protocols, for which the *stochastic gradient descent* plays a central role.

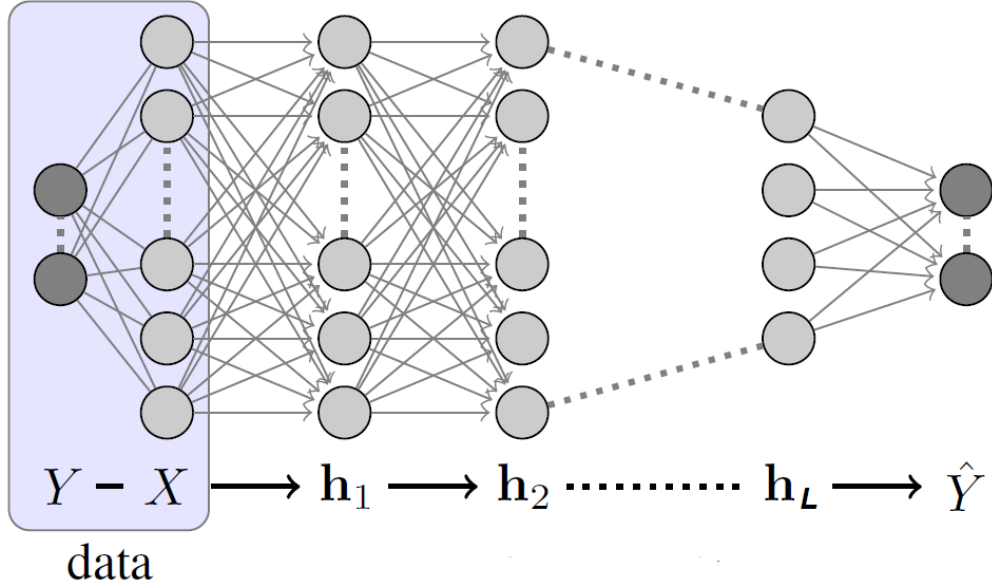


Figure 10: An example of feedforward Deep Neural Network with L hidden layers as represented in [179]. Each layer h_k corresponds to a compressed version of the input X and is theoretically determined via the minimization problem (41).

Deep Learning as compression-in-representation. With previous paragraphs, the role of a compressed representation has been discussed considering the abstract learning apparatus (36). A concrete implementation of this general view is offered by Deep Learning protocols - as highlighted by N. Tishby [158][179][159] - for which the dynamic in the informational plane associated to the training phase is examined. In a nutshell, the layered structure can be represented via a Markov chain - as in Figure 10 - where each layer h_k corresponds to an intermediate representation of X . Layer by layer, the encoder $p(h_k|x)$ and decoder $p(y|h_k)$ are considered, while the correspondent mutual information $I(X : h_k)$ and $I(h_k : Y)$ characterize the learning problem completely. Through training, each layer h_k is determined via weights adjustment trying to maximize $I(Y : h_k)$ and minimize $I(X : h_k)$, as stipulated by (41). Remarkably, by applying inequality (35) to the Markovian structure $X \rightarrow h_k \rightarrow h_l$, $I(X : h_k) > I(X : h_l)$ holds: neural networks provide a cascade of representations progressively more compressed. As represented by Figure 11, two phases can be empirically observed during learning via stochastic gradient descent [158][159][56][49][2], given an appropriate procedure in estimating mutual information [57]:

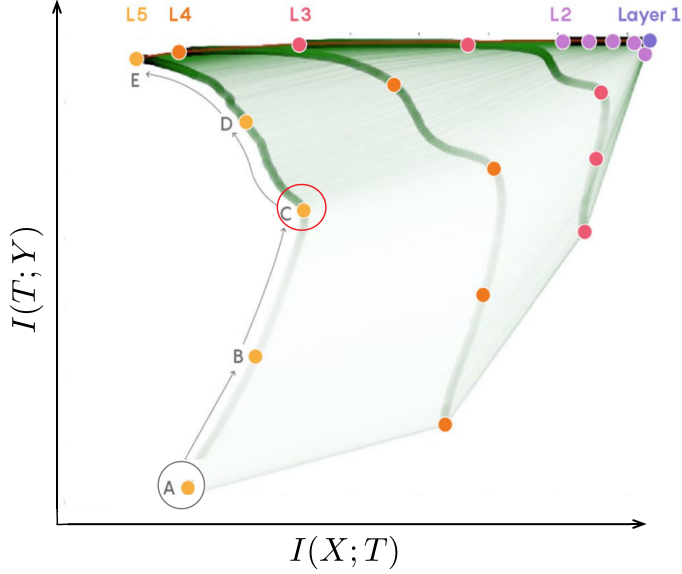


Figure 11: *Fitting and compression phases in DL*, as discussed in [179][158]. The fitting phase starts in A so that $I(X : T)$ and $I(T : Y)$ progressively increase. The compression phase begins in C, when $I(X : T)$ starts to decrease.

- First, the *fitting phase* enables the network to specialize increasingly in carrying out the task at hand, given a gradually larger and larger subset of the training set. During this first quick phase, $I(Y : h_k)$ grows rapidly. Moreover, a high signal-to-noise ratio for the stochastic gradient is found, i.e. the gradient means are much larger than the standard deviations. In other terms, the gradient drops into a flat minimum via a drift process.
- Second, the *compression phase* comes into play: the generalization component becomes dominant, and the network learns to ignore the irrelevant information while staying informative about Y . During this second phase, $I(X : h_k)$ decreases slowly, and a small signal-to-noise ratio for the stochastic gradient is found. In this case, the gradient seems to explore the reached flat minimum to irrelevant directions - i.e. without appreciably changing in cost function - via a diffusive process.

In other terms, the network adjusts as if trying to incorporate all information from the input required for well estimating the target but no more. The learning procedure that comes into play is different and richer than the classical curve-fitting perspective [117][21]. The compression phase has a central role: the representation compression is strongly connected to the generalization performance as prescribed in (38). The Empirical risk minimization is not enough, and interpolation would be insufficient to explain why the state-of-the-art algorithms exhibit so impressive performances in generalizing [9]. On the other side, the trained hidden layers tend to live near the Information Bottleneck theoretical bound in the information plane, so the corresponding encoder-decoder couple should satisfy the self-consistent equations

for (41). Moreover, the layered structure provides a computational benefit [159][2]. If generalization is obtained via compression, the Markovian structure allows each layer to compress the information already compressed in the previous one: all details compressed away in a lower layer are lost to the higher one so that each layer compresses independently of the others. This type of mechanism seems particularly suitable for applications such as image classification, where the relevant information seems highly distributed and available at different levels of abstraction. Consequently, it is reasonable that many redundancies for the specific classification task are typical so that Deep Learning can provide us with a representation with a smaller effective dimension.

Compression-in-resolution as a general constraint. The analysis proposed so far suggests the following general picture. On the one hand - by adopting the Empirical risk minimization principle - Machine Learning techniques are understood in terms of *curve-fitting* procedures. In this respect, some quantitative guarantees about the generalization gap control are provided by selecting a priori the hypothesis class \mathcal{F} . On the other hand, Deep Learning protocols seem to obey a partially different logic by automatically implementing a *coarse-graining* procedure: information irrelevant to the assigned task is progressively compressed - layer by layer - to provide coarser and coarser representations. As discussed above, this perspective works in the Large-scale Learning regime, for which the typicality argument comes into play whenever the factorization hypothesis on the probability distribution p holds. In that case, the generalization gap control can be explainable in terms of the informational quantities examined without introducing the class \mathcal{F} . This perspective suggests a parallel with the coarse-graining approach adopted in modeling, thus supporting the expectation that it is in principle possible to achieve performance at least comparable with that achievable through mathematical models. However, it is necessary to discuss in what terms coarse-graining is achievable from a dataset and what some inherent difficulties may be.

To better frame this point, let us briefly reconsider the Analogs Method, keeping in mind the problem of weather forecasting as an example. Given today's weather conditions, we would like to determine whether it will rain on a certain day thereafter, once again requiring an output in $\mathcal{Y} = \{0, 1\}$. The procedure is simple. Given the atmospheric configuration on a target day - say x^* - similar states x_i are identified in \mathcal{D} together with their corresponding predictand observation y_i . Then, the estimation of the predictand on the target day - say y^* - corresponds to the local observations that occurred on the closest analog atmospheric configuration. In formal terms:

$$y^* = y_k,$$

$$k = \arg \min_i \left\{ \delta(x_i, x^*) : (x_i, y_i) \in \mathcal{D} \right\},$$

where we assume that $\delta < \varepsilon$. Clearly, the Analogs Method is a simple memorization and search procedure: the generalization level is assigned a priori via the resolution ε , and it does not generalize outside the regime of the existing historical records. At this level, two observations are needed.

- The Analogs Method prescribes the theoretical compression threshold through the parameter ε , identifying all the points in every single ε -sphere. A higher ε value corresponds to a lower resolution and a more significant compression. Our expectations in performing adequate forecasting require a fixed predictability time t and the error tolerance δ^* , which can be interpreted as a measure of the generalization capacity. Fixed t and δ^* , the system responds by providing the *maximal admissible compression level* ε , given the predictability properties quantified by the maximal Lyapunov Exponent λ as in (16), or employing the multiscale structure as in (15).
- If the available database does not explore the feature space with resolution ε , it will be inadequate to capture the system behavior and provide accurate forecasts. In this regard, compression-in-representation has no role: raw data are used directly, and the information concretely used has a *local* nature. At the same time, no global patterns are taken into consideration.

These few remarks are enough to show how the Deep Learning compression mechanism works differently compared to the Analogs Method procedure. Nevertheless, there are good reasons to say that the limitations of the latter are also relevant in a broader sense than also providing a more precise answer to the relevance objection made by W. Pietsch. Let us proceed step by step, reconsidering the points discussed so far and introducing some additional elements. First, it is possible to propose a general argument by considering the compression procedure adopted. The *compression-in-representation* implemented by Deep Learning protocols involves a procedure of *compression-in-resolution* through which data points in suitable ε -spheres are progressively identified. Schematically, given the learning apparatus

$$Y \longleftrightarrow X \longrightarrow X_\varepsilon \longrightarrow \hat{Y},$$

the compressed representation X_ε - selected via the training phase to obtain \hat{Y} as close as possible to the correct label Y - consists of a *clusterized version* of X . From a general viewpoint, whenever compression-in-resolution has a central role, the ε -partitions injected by each layer must satisfy certain general conditions to work correctly and generalize well. More in particular:

- The ε -spheres that partition the input space \mathcal{X} must be coarse enough to admit a typicality argument and contain sufficient training instances. Conversely, suppose ε is too small. In that case, typicality does not work, and no data are contained in too many partition elements. Consequently, the label assigned to each empty ε -sphere is random, affecting the accuracy.
- At the same time, the partitions adopted must be fine enough to admit labels that are as homogeneous as possible in Y for the elements of the training set. This point requires an exponential increase in the number of the ε -spheres needed, providing us with a limitation in the spirit of that suggested by Kac Lemma. In this respect, the incompressibility of the initial data will thus remain a robust constraint in the forecasting context.

These remarks fit within the framework set by the argument of the typicality of input in the large-scale context. If typicality did not apply, we would not have the guarantees on the ability to generalize discussed above. However, the role of compression-in-resolution can also be considered from a somewhat different point of view, putting aside for a moment the problem of the guarantees offered by controlling the generalization gap. From a general perspective, the problem arises of justifying the fact that Deep Learning protocols necessarily involve some form of compression-in-resolution, thereby encountering the limitations already discussed for the Analogs Method. Let us briefly discuss how regularization plays a role in this regard.

Compression-in-resolution and regularization. To understand how compression in resolution works, the phenomenon of regularization is considered. As we have already mentioned, compression-in-resolution in a data-driven context corresponds to a partitioning process where all points belonging to a specific class are represented by a single realization, i.e. the centroid of ε -sphere. In this regard, two general options are available. On the one hand, Deep Learning protocols can undergo a hard partition - similar to the one described for the Analogs Method - at any point during the training process. However, this would not seem to be the case. The output typically exhibits a continuous and smooth function of the input - e.g. sigmoid of a sum of sigmoid - while a hard partition would involve discrete jumps when crossing the boundaries between partitioned classes. Alternatively, Deep Learning protocols could induce a process of compression-in-resolution through a soft partition. In this case, such a partition should be characterized by a characteristic resolution induced by regularization terms. In this respect, even when they are not explicitly included in the cost function as in (2), Stochastic Gradient Descent could introduce implicit regularization with similar consequences [16][167].

First of all, regularization and protocol's characteristic resolution are closely related. Both these concepts intervene in defining the effective hypothesis space, thus making the learning problem well-posed. Let us start by considering how resolution comes into play in terms of the decision boundary for a binary classification problem. Intuitively, one must look at the curvature of the boundary. Assuming to cover the space \mathcal{X} via ε -spheres, there is a maximum radius ε that allows the ε -spheres to delimit the decision boundary effectively. Smaller radii would require a larger number of spheres to cover \mathcal{X} effectively, demanding more data to learn the problem. On the other hand, larger radii would result in a loss of precision due to excessive compression. This trade-off is illustrated in Figure 12. At this level, a general problem arises naturally: how is a suitable resolution determined to learn a specific problem in the Deep Learning context effectively? Is it possible for the model to automatically adapt the resolution, allowing the hypothesis space to adjust during training? Preliminarily, we observe that the regularization term rewards small values for free parameters in the cost function. Consequently, regularizing will imply a decision boundary progressively closer to the linear behavior, with a corresponding decrease in the required resolution. In other terms, stronger regularization leads to a lower resolution in learning.

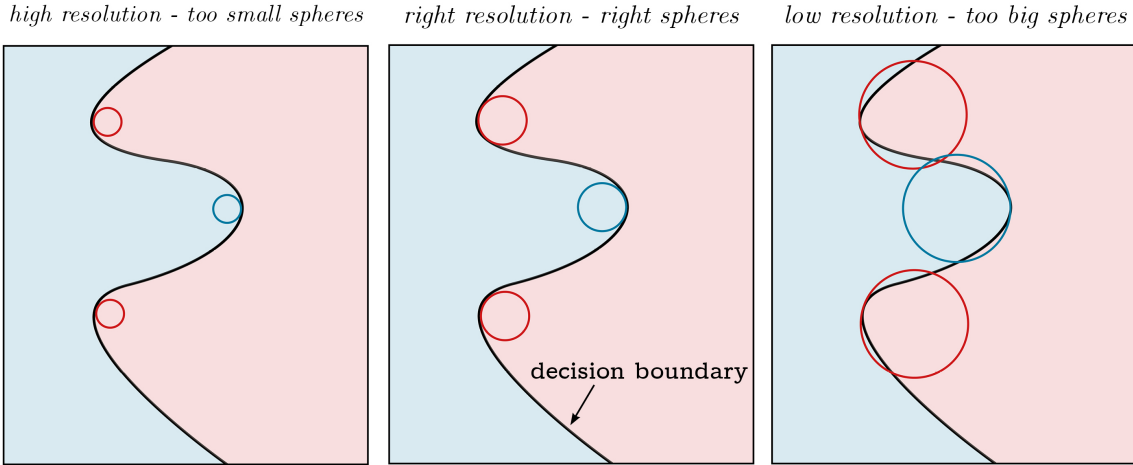


Figure 12: Visual depiction of the relationship between the curvature of the decision boundary and the radius of the covering spheres, directly related to the degree of compression-in-resolution applied to the prediction model. spheres that are too small can in principle maintain high accuracy but require a larger amount of data than necessary, otherwise inducing overfitting phenomena. spheres that are too large suffer from the opposite problem, not allowing sufficient expressive capacity of the model and causing underfitting.

On the one hand, *explicit regularization* as in (2) allows us to solve the problem empirically: different values for the hyperparameter γ are tried to choose the one that a posteriori has a better performance. On the other hand, *implicit regularization* could manage this process automatically. In a nutshell, the employing of Stochastic Gradient Descent in minimizing the cost function \mathcal{C} corresponds to equipping \mathcal{C} with the regularization terms as in the formula

$$\mathbb{E}(\mathcal{C}_{SGD}(\theta; \mathcal{D})) = \mathcal{C}(\theta; \mathcal{D}) + \frac{\alpha}{4} \|\nabla_{\theta} \mathcal{C}(\theta; \mathcal{D})\|^2 + \frac{(n-b)\alpha}{(n-1)4b} \Gamma(\theta; \mathcal{D}), \quad (43)$$

$$\Gamma(\theta; \mathcal{D}) := \frac{1}{k} \sum_{i=1}^k \|\nabla_{\theta} \mathcal{C}(\theta; \mathcal{D}_i) - \nabla_{\theta} \mathcal{C}(\theta; \mathcal{D})\|^2,$$

where α is the learning rate, n is the training set cardinality, b is the size chosen for mini-batches $\mathcal{D}_i \subset \mathcal{D}$, and the expected value in (43) corresponds to an average operation on all the possible choices for the mini-batches of size b .

The first term in (43) is due to the fact that the Gradient Descent update rule

$$\theta_{t+1} = \theta_t - \alpha \nabla_{\theta} \mathcal{C}(\theta; \mathcal{D}) \quad (44)$$

follows a trajectory of discrete steps that only approximately aligns with the ideal *Gradient Flow* curve defined via the differential equation

$$\dot{\theta} = -\nabla_{\theta}\mathcal{C}(\theta; \mathcal{D}).$$

Due to discretization error, for finite stepsize α , the discrete path may not lie exactly where the continuous one lies. Errors accumulate over time, making the paths more and more different. *Backward error analysis* comes into play in this regard: it aims at finding a differential equation $\dot{\theta} = -\nabla_{\theta}\tilde{\mathcal{C}}(\theta; \mathcal{D})$ such that its solution follows the approximate discrete path obtained from Euler’s method via 44. In other words, the goal is to reverse engineer the problem in order to find a modified \mathcal{C} such that the discrete iteration can be well-modelled by a differential equation. For Gradient descent, we have $\tilde{\mathcal{C}} = \mathcal{C}_{GD}$, that is

$$\mathcal{C}_{GD}(\theta; \mathcal{D}) = \mathcal{C}(\theta; \mathcal{D}) + \frac{\alpha}{4}\|\nabla_{\theta}\mathcal{C}(\theta; \mathcal{D})\|^2. \quad (45)$$

Geometrically, it penalizes narrower and steeper cost function minima, thus generalizing better. The contribution of the second term in (45) is shown in Figure 13.

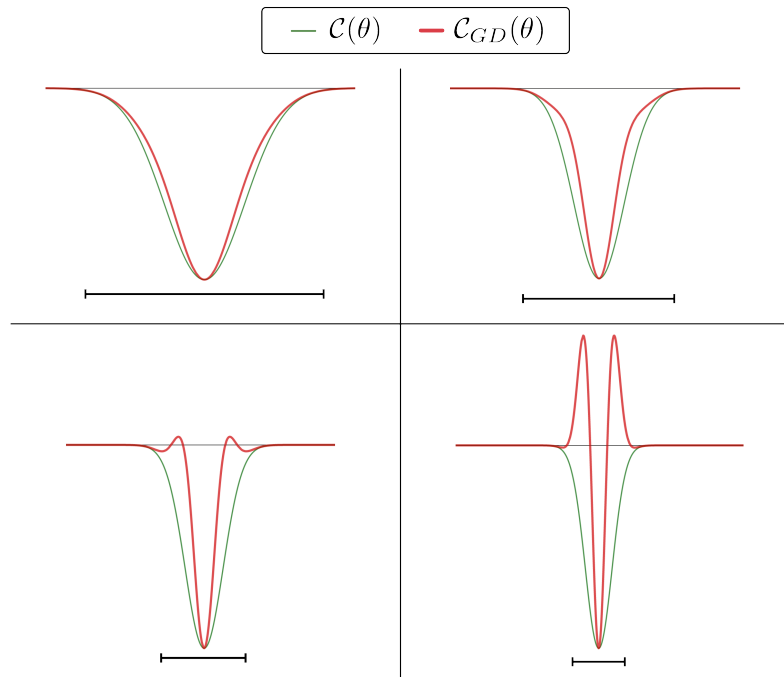


Figure 13: The tighter the minimum, i.e., the steeper its walls are, the more significant the regularization contribution in Equation 45 becomes, to the point of even reversing the sign of the cost function gradient, creating barriers around the minimum.

The second term in equation (43) is more interesting in our discussion, as it pertains specifically to Stochastic Gradient Descent and depends on the selected mini-batch size. Its importance is highest when mini-batches are small, potentially consisting of just a single example from the dataset. However, as we transition to using

the entire dataset as the mini-batch, which corresponds to standard Gradient Descent, this term gradually vanishes. This term plays a crucial role in evaluating the stability of minima within the optimization problem defined by \mathcal{C} across different mini-batches. If a minimum remains consistent and is located in the same region across all mini-batches, it is likely to reflect a genuine characteristic of the system. On the other hand, if its position in the θ space varies significantly, it suggests statistical fluctuations in the dataset and indicates a lack of generalization capability. To help in visualizing this phenomenon, refer to Figure 14: it is worth noting that even though the minimum of the cost function may remain the same in principle, its basin of attraction - i.e. the set of starting points from which the dynamics converge to the minimum - significantly narrows. In other words, the barriers prevent the Gradient Descent trajectory from stopping within the minimum.

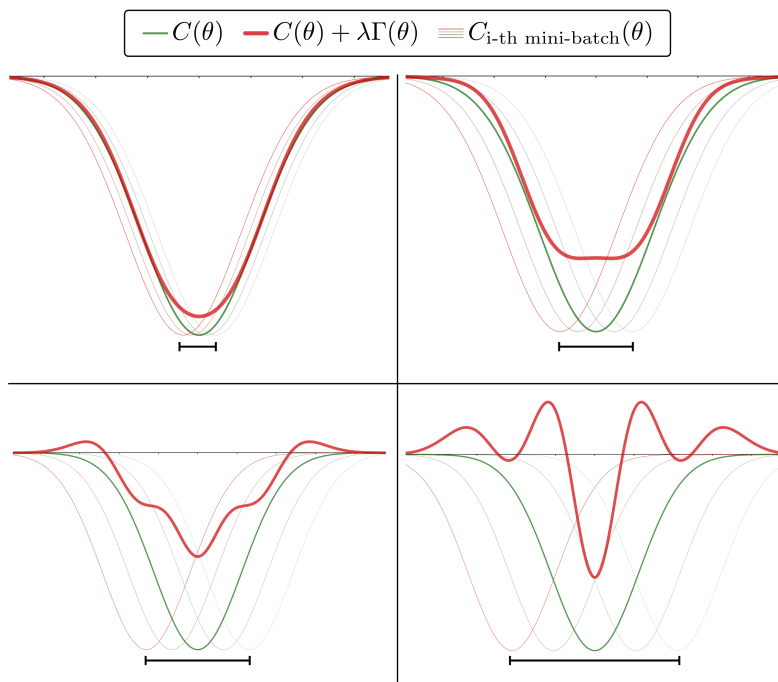


Figure 14: As the costs of different mini-batches become more spread near the minimum of \mathcal{C} , the significance of the SGD regularization term $\Gamma(\theta)$ increases.

The variation of the cost function on mini-batches reflects a variation in the instances that compose them. We can consider the mini-batches as realizations of independent datasets from the same distribution with a size that is only a fraction of the original dataset. From this perspective, mini-batches exhibit convergence properties in line with what is predicted by the bounds of statistical learning theory, one of which is (30). The bound provides an estimate of the variability of the cost function in the form of

$$|\hat{C}_i(\theta) - \hat{C}(\theta)| \approx \sqrt{k}|\hat{C}(\theta) - C(\theta)|,$$

where k is the number of parts in which the dataset is divided into mini-batches of size n/k . We observe that the term on the right-hand side is not computable explicitly, as it requires $C(\theta)$, the true distribution, thus an infinite amount of data. In this sense, this relationship can be interpreted as an inference tool. Additionally, we consider the role of model resolution with respect to the properties of the cost function and the relationship between the costs of individual mini-batches. Intuitively, reasoning in terms of coverings of ε -spheres, the higher the resolution, the greater the variability between mini-batches, for example, considering the occupation numbers of each individual sphere. Conversely, the coarser the description, the more similar the mini-batches will be to each other. In light of these considerations, the role of the second regularization term in SGD takes on a clearer meaning: if the model resolution is too high and requires an incompatible amount of data compared to what is available, the cost functions of the mini-batches will tend to be different from each other, inducing a stronger regularization as shown in the lower images of Figure 14. Conversely, if the amount of data is sufficient compared to the selected resolution, the mini-batches will exhibit more similar behaviors, not requiring the intervention of the regularization term. In other words, resolution compression and regularization are to some extent dual and address the same problem under complementary conditions.

Compression-in-resolution and locality. The role played by compression-in-resolution can involve limitations inherent in the problem at hand. Conceptually, two opposite extremes can be considered. On the one hand, inductive protocols such as Analogs Method exclusively employ local information. On the other, there are inductive procedures capable in principle of constructing high levels representations through lossy compression protocols recognizing relevant distributed patterns. While the recognized limits for Analogs Method follow directly from its local nature, a different inductive procedure could conversely employ distributed information and recognize global patterns in performing the assigned task. The information-plane framework for Deep Learning provides a high-level picture in this regard, showing how the compression procedure does not necessarily proceed by using locally concentrated information. However, a general bottleneck seems unavoidable: the more local information is relevant for the task selected, the more limitations attributable to compression-in-resolution can affect the performance. The system's properties play a role at this level, telling us the extent to which locally concentrated information is relevant and giving us the consequent applicability criteria. If, on the one hand, it may be reasonable to have a sufficiently large database for the classification of a particular type of image - as in the case of lung cancer diagnosis - contrarily, the consequences due to compression-in-resolution could be much more limiting for other tasks, such as weather forecasting. Generally, an inductive protocol might at best be able to determine a quantitative law that fits well the data available and whose recognition by a human specialist is cognitively prohibitive. This task can be interpreted as a data-compression procedure. Still, it is incorrect to conclude that the law obtained corresponds to a compressed version of all the behaviors observed [35][70]. Indeed, the exponential divergence of orbits implies the failure in compressing the time evolution effectively, also assuming that a good model is explicitly known. As

discussed above for the Analogs Method, the initial data compression corresponds to identifying a sphere of points in the phase space, forgetting some details, and thus removing a certain amount of information. Suppose it is assumed that a component of the local information can not be reconstructed by considering the one available in other regions of the phase space. In that case, its removal undermines the serviceability of any inductive protocol in that particular regime. In this sense, the limitation highlighted with the analysis of the Analogs Method becomes relevant again.

Part III

Causality

This third part aims to discuss the role of causality as a tool to manage information when some distribution shift comes into play. Some formal ways of characterizing cause-effect relations will be critically examined, and the potential connections that alternative frameworks may have will be explored to establish a more unified viewpoint. The following discussion is oriented toward justifying how *causalization* can be seen as a *compression-in-representation* procedure, thus complementing the epistemological picture already presented in the previous parts.

1 Mathematical Framework

By employing cause-effect relations, an intelligent agent elaborates a more economical representation of observational data than the purely associational one. Moreover, causality provides an inferential framework suitable for predicting the system's response in the face of external stimuli. In this regard, the language of *Bayesian Networks* offers valuable tools to formalize the two points above [82][115][170][187]. Let us delve into these matters by progressively introducing the key concepts, bearing in mind that the role of Machine Learning will be examined from two perspectives. On the one hand, automatic methods could help in approaching some typical problems about causality, such as *Causal Discovery* and *Causal Inference*. On the other hand, the possibility of enhancing inductive methods by adopting a causal framework will be explored, discussing whether causality can inform how Machine Learning should be done, at least in principle.

From a more general perspective, causality has played a central role in philosophy, originating many alternative theoretical proposals. Recently, P. Illary and F. Russo have insisted on a *pluralistic* conception of causality [72], for which the accounts developed in the literature are the tiles that can be used in building the causal mosaic needed to handle the problems of *explanation, prediction, control, inference* and *reasoning*. This approach opens up numerous metaphysical, epistemological, semantic and pragmatic questions, to some extent also dependent on the disciplinary context under consideration. On the other side - with no ontological commitment - a more restrictive viewpoint will be adopted in what follows. In our view, causality provides a compression-in-representation compatible with a notion of *embedded intelligence* that takes into account the agent's ability to intervene in the environment and generalize well, thanks to a *modular* picture of the external world. In this regard - to ensure their survival - organisms must effectively acquire information from their environment, then process and organize it into suitable representations. Moreover, the acquired information needs to be appropriately integrated with existing *background knowledge*, which is structured and ordered in a way that facilitates effective interactions with the environment by minimizing the resources required to accomplish significant tasks. In other words, processing infrastructure may respond to the environmental complexity, building up accordingly [86][163][100]. These considerations can also inspire the integration of inductive methods with causal language - where the notion of intervention has a central role - in order to achieve better-performing learning protocols [146]. Let us start with an introductory example to explore this setting step by step.

A trivial remark. Let X be a set of m binary random variables X_i . The joint probability distribution P_X can be represented through a table in which each row corresponds to a possible realization $x \in \mathcal{X}$, with the probability $p(x)$ attached. In the absence of any additional assumptions, this tabular representation requires $2^m - 1$ rows, with intractable costs from the *cognitive* and *computational* viewpoints

as m increases. However, independence relations among variables can come into play allowing us to represent P_X more compactly. In general, the probability chain rule holds for every choice in indexes in X :

$$P_X = P_{X_1} \cdot P_{X_2|X_1} \cdot \dots \cdot P_{X_n|X_1, \dots, X_{m-1}} = \prod_{i=1}^m P_{X_i|X_{j < i}}.$$

In simple cases, a preferred *total order* is available: each variable depends only on the one immediately preceding. Given the Markov chain

$$X_1 \longrightarrow X_2 \longrightarrow \dots \longrightarrow X_m,$$

the *Markov Condition* $P_{X_i|X_{j < i}} = P_{X_i|X_{i-1}}$ holds, and only $2m - 1$ rows are needed for the tabular representation. The number of parameters required to represent P_X changes from scaling exponentially to scaling linearly in m . This example is generalizable to *partial orders* by considering more complex hierarchies represented by a *Directed Acyclic Graph (DAG)* \mathcal{G} , where variables in X correspond to nodes in \mathcal{G} . This approach follows from a simple observation: the conditional independence relation satisfies the *graphoid axioms* [122][123], and the independence statement $X \perp\!\!\!\perp Y|Z$ can be read graphically as follows: all paths from node X to Y are intercepted by a subset Z of nodes in \mathcal{G} [115].

Graph representation. Let us start with some terminological elements [115]. Given a node $X_k \in \mathcal{G}$, we say that the nodes from which incoming links depart into X_k are its *parents*, denoted by $Pa_k^{\mathcal{G}}$. Moreover, nodes receiving an outgoing link from X_k are its *children*, $Ch_k^{\mathcal{G}}$. Similarly - adopting this intuitive genealogical language - we have the *ancestors* $An_k^{\mathcal{G}}$, and the *descendants* $De_k^{\mathcal{G}}$. In what follows, the superscript in \mathcal{G} may be omitted if the graph is clear from the context. On the general level, the Markov property is a well-established assumption in graphical modeling: when a distribution P is Markov with respect to \mathcal{G} , independencies are encoded in \mathcal{G} and can be exploited for efficient computation and data storage. More particularly, the graph \mathcal{G} is said *P_X -compatible* if the *Local Markov Condition* holds, assuming each variable to be independent of its ancestors given its parents:

$$P_{X_i|\neg De_i, Pa_i} = P_{X_i|Pa_i}. \quad (46)$$

The Local Markov Condition as in (46) can be equivalently expressed in terms of the *canonical factorization* with respect to \mathcal{G} , so that the joint distribution can be re-written as

$$P_X = \prod_{i=1}^m P_{X_i|Pa_i}. \quad (47)$$

Moreover, the P_X -compatibility condition - as captured by (46) and (47) - also admits a characterization in terms of *d-separation* relations [115][129]. Given three sets of nodes A, B, C in \mathcal{G} , A and B are *d-separated* by C , say $A \perp\!\!\!\perp B|C$, if - for every couple $A_i \in A$ and $B_j \in B$, and for every path γ in \mathcal{G} which connects A_i with B_j - at least one of the following conditions holds.

- There is a *chain* blocked by C , that is to say, we have $\rightarrow C_k \rightarrow$ or $\leftarrow C_k \leftarrow$ in γ , with $C_k \in C$.
- There is a *fork* blocked by C , that is to say, we have $\leftarrow C_k \rightarrow$ in γ , with $C_k \in C$;
- There is a *collider* not fixed by C , that is to say, we have $\rightarrow W \leftarrow$ in γ , with W and De_W with a null intersection with C .

Given the terminology above, it is possible to show that \mathcal{G} is P_X -compatible if and only if the following *Global Markov Condition* holds:

$$A \perp_{\mathcal{G}} B|C \Rightarrow A \perp_P B|C, \quad (48)$$

where $A \perp_P B|C$ stays for $P_{A,B|C} = P_{A|C}P_{B|C}$ [115]. Remarkably, we would an algorithm that - taking as input P_X or a dataset distributed as P_X - returns a graph \mathcal{G} capable of representing all and only the independence relations of P_X . In other words, the *Faithfulness Assumption* is typically required:

$$A \perp_{\mathcal{G}} B|C \Leftrightarrow A \perp_P B|C. \quad (49)$$

In summary, while the Markov condition (48) enables us to read off independence relations from the graph structure, the Faithfulness assumption (49) allows us to infer dependencies starting from the graph. Although the P_X -compatible DAG \mathcal{G} is a compact way to represent the set of independence relations provided by P_X , it does not necessarily correspond to the system causal structure. This point becomes clear by realizing that the set of P_X -compatible graphs is not necessarily a singleton, as discussed below.

Markov Equivalence Class. Given m variables X_1, \dots, X_m , the number of possible DAGs grows super-exponentially as m increases: the length of the numbers grows faster than any linear term in m [129]. On the other side, the independence relations provided by P_X introduce some constraints and reduces the number of DAGs of interest: the subset of P_X -compatible DAGs is called the *Markov Equivalence Class* (MEC) for P_X . This class can be completely characterized by introducing two other definitions. First - given a DAG \mathcal{G} - the *\mathcal{G} -skeleton* is the graph \mathcal{G}' obtained by replacing all directed links in \mathcal{G} with undirected ones. Second, given three nodes $X, Y, Z \in \mathcal{G}$, with $X \rightarrow Z$ and $Y \rightarrow Z$ in \mathcal{G} and no link between X and Y , the pattern $X \rightarrow Z \leftarrow Y$ is called a *v-structure* of \mathcal{G} . At this point, the following result provides the desired characterization.

Theorem 1 [185]. *Two DAGs are in the same MEC if and only if they share the same skeleton and the same set of v-structures.*

Significantly, MEC_P is graphically represented through a *Partially Directed Acyclic Graph* (PDAG) \mathcal{G}_P , obtainable via three simple steps:

- We start with only nodes without any edge.

- The skeleton of \mathcal{G}_P is the same as any DAG in MEC_P : the required undirected edges are added to the graph.
- If an edge of \mathcal{G}_P has the same direction for all the DAGs in MEC_P , this edge is directed in that direction.

The new graph \mathcal{G}_P is referred to the *Completed Partially DAG (CPDAG)* of P_X . We can think of MEC_P as the set of alternative causal structures with which the probability distribution P_X can be generated. These structures can be represented simultaneously by \mathcal{G}_P , whose oriented arrows correspond to the common cause-effect relations. Informally, every directed edge $X_i \rightarrow X_j$ in \mathcal{G} means that X_i directly influences X_j , so that intervening on X_i changes the probability distribution of X_j when all other variables are held constant. At this point, the concept of intervention - and the distribution shift that consequently occurs - requires to be formalized. In this regard, the language of *Structural Equations* provides us with a valuable tool to clarify the semantics of intervention.

Structural Equations. First of all, we distinguish the *endogenous* variables X of the system, i.e. those actually observable, from the *exogenous* ones N , that we can not observe. At this level, the *Causal Sufficiency Principle* is assumed. Each exogenous N_k variable influences one and only one observable X_k so that the following independence properties hold:

$$Pa_k \perp\!\!\!\perp N_k, \quad N_i \perp\!\!\!\perp N_j$$

for all $i \neq j$ and $i, j, k = 1, \dots, m$. The graph \mathcal{G} can be consequently extended with the unexplained nodes N_k , adding all the directed edges $N_k \rightarrow X_k$. The extended graph tells us that Pa_k and N_k produce the observable value of X_k through an *underlying mechanism*. Structural equations make explicit this mechanism by providing a deterministic function f_k such that

$$X_k := f_k(Pa_k, N_k). \tag{50}$$

It is important to emphasize that (50) must be interpreted as an *asymmetric assignment* - from right to left - and not as a standard symmetric equation. Given the distribution P_N , the functional assignments f_k allow us to compute the joint distribution P_X , which has properties inherited from the \mathcal{G} 's topology. Intuitively, we can think of N as a source of information that spreads through the graph \mathcal{G} . Given the factorization (47), each factor $P_{X_k|Pa_k}$ is interpreted as produced by the mechanism represented by the deterministic function f_k and the marginal distribution P_{N_k} . On the other side, N_k admits different interpretations.

- First of all, N_k might reflect the presence of a measurement error on X_k .
- Secondly, other variables that influence X_k besides Pa_k - not observed - might be described by N_k .

- Alternatively, N_k represents some indeterminacy in the functional relationship between X_k and Pa_k , which is about the underlying mechanism.
- Finally, also a combination of all previous interpretations can not be excluded.

In any case, the individual values of N_k are assumed not to be directly measured, while X_k and Pa_k can be - in principle - observed. In other words, causal relationships are expressed in terms of deterministic models while probabilities are introduced through some unobserved variables. This framework is compatible with Laplace's idea of natural phenomena. Nature's laws are deterministic, while randomness is a consequence of the observer's ignorance about boundary conditions. Traditionally - particularly in the natural sciences - we expect that relationships such as (50) occur over time and are governed by a set of coupled differential equations. Under certain conditions, it is possible to derive structural equations that prescribe how a system at equilibrium responds to an external stimulus [107]. However - in our following discussion - (50) is considered the primitive object for introducing the interventional semantics.

Interventional semantics. Adopting the above-mentioned mathematical framework, the external *intervention* $do(X_l = \hat{x})$ takes place when the system's structural equation $X_l := f_l(Pa_l, N_l)$ is replaced by the assignment $X_l := \hat{x}$. As far as the graph level is concerned, the correspondent \mathcal{G} is modified in $\hat{\mathcal{G}}$ by removing all the arrows incoming in X_l , and setting X_l distributed as a delta-function $\delta_{\hat{x}}(x)$ concentrated in \hat{x} . Consequently, the observational P_X shifts to a new distribution $P_{X|do(X_l=\hat{x})}$, which admits the canonical factorization with respect to the new graph $\hat{\mathcal{G}}$,

$$P_{X|do(X_l=\hat{x})} := \prod_{i=1}^m P_{X_i|Pa_i^{\hat{\mathcal{G}}}} = \delta_{\hat{x},x_l} \prod_{i \neq l}^m P_{X_i|Pa_i^{\mathcal{G}}} \quad (51)$$

The construction above can be generalized considering a more general shift as in (5). At any rate, two implicit *invariance assumptions* are required. First, the mechanism f_k is independent of Pa_k -distribution: each structural equation (50) is invariant under interventions on independent variables Pa_k . Second, a *modularity* principle is satisfied: the functions $\{f_k\}_{k=1}^m$ are independent of each other, that is to say, every f_i is invariant under changes of f_j for all $i \neq j$. Remarkably, $P_{X|do(X_l:=\hat{x})}$ is characterized by (51) only if \mathcal{G} is known and Causal Sufficiency holds, and unobserved confounders are not allowed. On a general level, determining $P_{X|do(X_l:=\hat{x})}$ can be an arduous task. In this regard, J. Pearl has proposed a *complete* set of rules [71] - known as *Do-Calculus* - to infer the interventional distribution starting from the observational one, when the causal graph also admits unobserved common causes [115][120][121]. Two comments may be valuable at this point.

- The mathematical constructs introduced thus far enable the formalization of an initial concept of *embedded representation* in the external world. While structural equations can be regarded as a model for a data-generating process - both with and without interventions - it is essential to consider that the accurate representation of the system via structural equations should provide the correct

observational distribution P_X , and all interventional distributions $P_{X|do(X_l=\hat{x})}$ - and more generally $P_{X|do(X_l\sim P_{\hat{X}})}$ as in (5) - should correspond to distributions obtained through *randomized experiments* [46][115].

- Structural equations are more expressive than the observational and interventional plan. Indeed, it is possible to show how two systems of structural equations can agree on these two levels while showing differences on the counterfactual plane. Remarkably, counterfactual statements lack an evident correspondence in the external world, which consequently renders them not falsifiable. However, humans often think in counterfactual terms, and counterfactuals can have a role in organizing the background knowledge available and in making better decisions [115][61]. This point lies beyond this dissertation.

2 Causal Pipeline

Understanding a system via causal language must necessarily pass through progressive procedural bottlenecks, each with theoretical and practical consequences. Generally speaking, the more one intends to climb the *Ladder of Causality* as represented in Table 2 - firstly passing through the interventional level and then accessing the counterfactual one - the more critical the barriers become [117][15]. A preliminary indication of this typical scenario emerges considering the cause-effect problem in the bivariate context [129][73]. Beyond its practical applications, a detailed discussion of this elementary topic can clarify some general features of causality, its role in making practical inferences, and its possible relations with Machine Learning methods. Postponing the details to the next section, the following scenario arises.

- The relation between cause and effect is asymmetric. As opposed to mere statistical correlation, causation has directionality. Consequently, we say that an adequate theory of causality must explain this asymmetry, which we should recognize at some level when we try to extract causal information from statistical dependencies.
- However, given the observational joint distribution $P_{X,Y}$ without any additional prescriptions, the asymmetry between cause and effect can not be established. More precisely, for every $P_{X,Y}$ a functional model $\Phi_{X\rightarrow Y}$ exists, that is $Y := f(X, N)$ for some deterministic function f and a noise N such that $N \perp\!\!\!\perp X$. On the other hand, also a functional model $\Phi_{Y\rightarrow X}$ is available.
- The symmetry recognized above concerns the theoretical distribution and does not depend in any way on the amount of available data. This observation marks an important difference from the observational level, where incremental gains in accuracy are achievable by adding more training data.
- Remarkably, as we will elaborate on later, the functional models $\Phi_{X\rightarrow Y}$ and $\Phi_{Y\rightarrow X}$ generally exhibit different mathematical structures. As a result, the asymmetry between cause and effect can be established by imposing specific

conditions on the class Φ of admissible functional models, thereby incorporating relevant background knowledge regarding the bivariate system.

- Alternatively, a measure of complexity can be introduced to distinguish the two available options, and an *Occam's Razor* argument comes into play. At any rate, in each of the two options, some background knowledge or additional assumptions seem unavoidable to breaking the symmetry without using interventional data.

It is important to emphasize that other problems emerge immediately in the face of a slightly deeper analysis. In concrete scenarios, the dependency between X and Y can manifest in different ways. For instance, the observed random variables can be conditioned on unobserved variables, leading to the occurrence of a *selection bias*. Moreover, X and Y can follow a physical law, inheriting a time dependence that does not correspond to a causal relation. At any rate, problems and limitations that occur in the bivariate case can be more challenging in the *multivariate* context. The structures involved become richer and more complex, while further limitations also arise at the computational level. Therefore - before returning to consider the bivariate problem - a causal pipeline for the multivariate case will be discussed by considering well-separated modules, keeping in mind two general goals. On the one hand, this description aims at providing a broad and high-level look at the scientific practice involved in causality. On the other, it will be clarifying in characterizing the specific assumptions and bottlenecks that arise step by step.

Handling Data

The first problem to be addressed - which extends far beyond the realm of causality - concerns the limited availability of data. Although many inductive methods are *asymptotically consistent* - that is to say, better and better approximating an ideal operating condition as the number of available data increases - consistency does not always induce the expected practical advantage, and some problems due to the finite data regime are often simply unavoidable.

Independence tests. Suppose we want to investigate a system involving multiple variables, and we are interested in determining whether there is conditional independence between X and Y given the set of variables \mathbf{S} . To test this, we rely on a statistical procedure that examines a subset of the available data. Specifically, it verifies the validity of independence for each possible realization $\mathbf{S} = \mathbf{s}$. When considering binary variables for simplicity, the number of possible realizations becomes $2^{|\mathbf{S}|}$. Consequently, each test only has access to a fraction of the complete dataset, approximately $1/2^{|\mathbf{S}|}$. As a direct consequence, in order to maintain the efficacy of statistical independence tests, we would need to exponentially increase the amount of data available for analysis. This rapid increase makes the procedure practically infeasible. The reasoning can be directly extended to non-binary discrete variables. When one considers continuous data, the problem often gets worse. First of all, many of the most popular independence tests, such as *G-test*, *Fisher-test*, and *Entropy-based*

tests, involve a binning procedure, i.e., a block discretization of the continuous distribution, thus ending up with the problems analyzed before. As an additional problem, the choice of the bins' size can significantly affect the result, introducing an extra degree of uncertainty and arbitrariness [196]. As a last remark, finite data causes some algorithms-specific problems, even of theoretical nature. For example, among the algorithms that will be analyzed later, in this regime PC becomes order-dependent [40], while GES loses the guarantee of finding the absolute optimum w.r.t its score [37].

From Data to Graphs

Let's now put aside the finite data problem and shift our attention to the subsequent steps in the process. Assuming we have access to a probability distribution $P(\mathbf{O})$, our goal is to examine the observed variables in the system and uncover information about their causal relationships. However, we encounter an immediate challenge: the causal model we are trying to infer could encompass a larger set of variables, denoted as \mathbf{V} , than those currently available to us. Since \mathbf{V} is unknown, there are an infinite number of models that could potentially fit the given distribution. Each model may involve distinct "hidden" variables and establish different causal connections among the observed variables. As a result, without placing any restrictions on the types of models considered, *the scientist is unable to make meaningful assertions about the underlying structure that governs the phenomena*. Following the principle of Occam's razor, it is reasonable to rule out any theory for which we find a simpler, less elaborate theory that is equally consistent with the data. Theories that survive this selection process are called minimal [115].

However, it is important to recognize that Occam's razor serves as a heuristic principle and does not guarantee that the chosen model perfectly represents the actual underlying mechanism being studied. Therefore, its application, while reasonable, inherently involves making arbitrary assumptions. In fact, an excessive reliance on Occam's razor can sometimes lead us towards overly simplistic models rather than genuinely simple ones. For instance, let's consider the assumption of "Causal Sufficiency." By imposing $\mathbf{V} = \mathbf{O}$, we assume the simplest possible scenario by eliminating any potential latent variables. However, this assumption comes at a considerable cost. In the vast majority of cases, it yields models that are not only simplistic but also highly unrealistic, contradicting the fundamental rationale behind applying Occam's razor – that the simplest solution is likely to be the correct one.

Along the same lines, let us consider the assumption of *Causal Faithfulness*, quickly introducing a classic counterexample. Consider a DAG \mathcal{G} represented in Figure 15 and defined through the linear equations

$$\begin{aligned} X_1 &:= \varepsilon_1 \\ X_2 &:= \alpha X_1 + \varepsilon_2 \\ X_3 &:= X_1 + \beta X_2 + \varepsilon_3. \end{aligned} \tag{52}$$

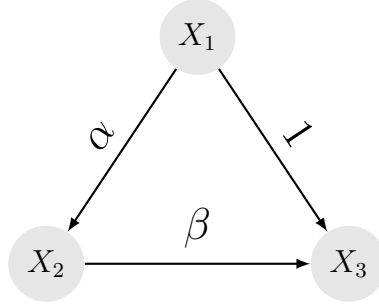


Figure 15: A simple DAG. When $\alpha\beta = -1$, faithfulness is broken. In this case X_1 's effect over X_3 is cancelled.

Setting $\alpha\beta = -1$, the two causal paths of \mathcal{G} from X_1 to X_3 cancel each other, inducing the relationship $X_1 \perp_P X_3$ over the probability P generated by the graph, even if $X_1 \not\perp_{\mathcal{G}} X_3$. Actually, the simpler DAG $X_1 \rightarrow X_2 \leftarrow X_3$ would be P-compatible.

Causal Faithfulness excludes the possibility that causal paths on the graph cancel each other out through unlikely coincidences, hiding from independence tests. In this sense, the guiding principle is analogous to Occam's razor. Typically, this assumption is considered not particularly relevant, arguing that non-faithful configurations occupy a null measure in parameter space in probabilistic terms. However, recent results show how, in the finite data regime and with a permissible range of error on statistical tests, Causal Faithfulness could be much more problematic than we think [180].

Functional Assumptions

As a final step, let us assume that we have passed the first two steps described above unscathed, obtaining the causal graph up to its MEC. To infer the direction of remaining undirected edges is necessary to narrow the class of assumptions even further by making additional assumptions. Let us consider a specific node of the graph, say X_j . Using the language of structural equations, we can represent the causal mechanism that determines X_j through the equation

$$X_j := f_j(Pa^{\mathcal{G}}(X_j), N_j). \quad (53)$$

In the previous stage, we concentrated on investigating the topological properties of \mathcal{G} , exploiting its set of independencies induced by the d-separation criterion. We did not impose any constraint on the form of f_j , except by determining its parameters $Pa(X_j)$. To further determine the DAG, we must make assumptions about f_j 's properties, restricting ourselves to a more specific set of admissible functions. If the assumed Functional Causal Model is too restrictive to be able to approximate the true data-generating process, the causal discovery results may be misleading. Therefore, if the specific knowledge about the generating mechanism is not available, to make it useful in practice, the assumed causal model should be general enough, such that it can reveal the data generating processes approximately [55].

A natural question then arises: How can we be able *a priori* to assess the reasonableness of the assumptions we make on f_j without having access to any ground truth on the model's interventional data? As an additional remark, determining the form of f would automatically give us access to the counterfactual plane, which is not even accessible *a priori*. From this point of view, albeit partially implicitly, the strength of the assumptions is shown not so much by the theoretical consequences they imply but by the abnormality of the causal information they give us access to.

In addition, causal discovery methods based on Functional Causal Models cover a broad spectrum of different hypotheses: Invertibility of f , linearity or non-linearity, Gaussian noise or not, additive, multiplicative, and many combinations of these concepts. In conclusion, while qualitative knowledge of the system can provide a great deal of information on the structure of the underlying DAG, the blind application of FCMs to models of which little or nothing is known would, in all likelihood, provide little more valid conclusions than choosing the direction of the arrows in a random fashion.

3 Cause-Effect Pairs

This section discusses a specific problem that may be considered both the most fundamental from a conceptual point of view and the least realistic in modeling concrete situations: the cause-effect problem in the *bivariate* context [73][129]. Given only two observable and *correlated* variables - X and Y - it is intended to establish in which of the two possible alternatives the eventual causal link is directed. More precisely, the scenario of interest stays in the following terms.

Given n independent and identically distributed observations drawn from some distribution $P_{X,Y}$ - say $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ - infer whether X causes Y or Y causes X , considering the premise that precisely one of these alternatives holds true.

A formal analysis of this binary classification problem is considered in some detail by expanding further the analysis briefly summarized in Table 2 and reconsidering the concepts already introduced so far. Remarkably, $X \rightarrow Y$ and $X \leftarrow Y$ belong to the same *MEC*, as guaranteed by the Theorem 1. Consequently, the required inference will require assumptions additional to the observational level. It will be a matter of discussing in what terms these assumptions can be considered general and whether a theoretical characterization is available for the correct direction. Although we can not expect to describe a realistic system as a collection of bivariate relations, the scientific literature contains many discussions about variables in pairs. Moreover - and beyond its practical applications - a detailed discussion of this elementary topic clarifies some general features of causality, its role in making practical inferences, and its possible relations with Machine Learning methods. Starting with the *Reichenbach's Principle* [135][129], a set of principles useful in framing the notion of causality is discussed. Secondly, we will discuss some conditions under which the symmetry between X and

Y is broken, and requirements helpful in identifying direct causation will be examined, excluding the existence of a not observed confounder. The preliminary formal results suggest some considerations about using the *information* and *algorithmic complexity* as technical tools in characterizing the difference between cause and effect [74][129]. The role of Occam's Razor is considered in light of this discussion, providing us with an epistemological motivation.

Causal Principles

As previously mentioned, causal representation involves the utilization of a probabilistic model but necessitates additional information beyond what the observational level provides. Therefore, learning causal structures and reasoning in causal terms are more challenging objectives compared to their statistical counterparts. This section aims to delve deeper into the relationship between causal and statistical planes, presenting a collection of general principles that will be further examined in their respective application domains. While statistical properties alone are inadequate for determining the underlying causal structure, it is plausible to hypothesize that we can infer the presence of causal connections based on statistical dependencies. This assumption is the content of *Reichenbach's Principle* [135][129].

Principle 1 *Assuming that the random variables X and Y are statistically dependent, a third variable Z causally influences both.*

As a particular case for Principle 1, Z may coincide either with X or Y . In other words, if X and Y are correlated, there are at least three elementary alternatives: X causes Y , Y causes X , or there is a not observed confounder. In graphical terms $X \rightarrow Y$, $X \leftarrow Y$ or $X \leftrightarrow Y$. Moreover, two other combinations are in principle available: the confounder case can be further equipped with an arrow between X and Y . In interventional terms, the three elementary alternatives mentioned above correspond to the following distributions:

$$\begin{aligned} X \rightarrow Y : P_{Y|do(X=\hat{x})} &= P_{Y|X=\hat{x}} & P_{X|do(Y=\hat{y})} &= P_X, \\ X \leftarrow Y : P_{Y|do(X=\hat{x})} &= P_Y & P_{X|do(Y=\hat{y})} &= P_{X|Y=\hat{y}}, \\ X \leftrightarrow Y : P_{Y|do(X=\hat{x})} &= P_Y & P_{X|do(Y=\hat{y})} &= P_X. \end{aligned}$$

On the other side, the confounder cases equipped with an arrow between X and Y do not imply differences with respect to the first two cases above. Consequently, they are not falsifiable and can be omitted in our analysis. At this point, three remarks can be valuable.

- First, if X precedes Y in time, the case $X \leftarrow Y$ can be excluded and only the other two alternatives remain available. Although time may involve simplification, only the static case is considered in this section, while the dynamic one will be examined later in some detail.
- Second, given only the joint observational distribution $P_{X,Y}$ - independently of the amount of available data - in many concrete situations it is not possible to

construct an interventional procedure that allows us to establish which of the three elementary options is the actual one.

- Third - excluding the possibility of direct interventions on the system - it is a mathematical problem to find other conditions that can be helpful in identifying the correct causal link, adopting in some sense a background knowledge about the system.

In the face of the last remark mentioned above, the language of *structural equations* provides us with a system representation that can be used to introduce additional assumptions. The following principle is then motivated.

Principle 2 *Assuming that there is a causal link from X to Y , a function f and a noise variable N exist such that $Y := f(X, N)$.*

On the one hand, the noise N in Principle 2 can interpret the influence of a not observed external environment or errors in measuring. On the other hand, the function f may be thought of as a physical mechanism that relates the variables involved. Crucially, the function f should be read as an assignment rather than as a mathematical equation. This subtle difference is essential in justifying the interventional framework: if we set Y to a fixed value \hat{y} and N takes the value \hat{n} , the assignment $Y := f(X, N)$ does not imply that X assumes a the value that the equation $Y = f(X, N)$ would prescribe. More succinctly, the interventions on X and Y involve the following relations:

$$\begin{aligned} X := \hat{x} &\mapsto Y = f(\hat{x}, N), \\ Y := \hat{y} &\mapsto X \sim P_X \end{aligned}$$

It is important to emphasize that while Principle 2 admits an ontological interpretation, Principle 1 should be read as a methodological insight. In concrete situations, the dependency between X and Y may also arise in different terms. For example, the random variables we observe are conditioned on others that we can not observe, producing a selection bias; random variables could only appear to be dependent through a failable test; our two random variables follow a simple physical law, inheriting a time dependence that does not correspond to a causal relation.

Rethinking modularity. Returning to the joint distribution $P_{X,Y}$, two possible decompositions in terms of conditional probabilities are available: $P_{Y|X}P_X$ and $P_{X|Y}P_Y$. These two alternatives in decomposing $P_{X,Y}$ can correspond to the two possible causal links. If X causes Y , then the first decomposition is preferred and can be read productively. First, it is in principle possible to perform a localized intervention on X , changing P_X without altering $P_{Y|X}$. Second, P_X and $P_{Y|X}$ correspond to two not necessarily related objects. These facts directly follow Principle 2 whenever we assume that the conditional probability $P_{Y|X}$ is the probabilistic counterpart of an underlying mechanism f that does not depend on P_X . On the other side, this propriety is not obvious: we can imagine a system where selecting a particular x -range produces a correspondent regime in the Y 's response so that the underlying

mechanism required by Principle 2 may depend on X . This discussion motivates the introduction of a modularity principle as a simplification in representation.

Principle 3 *Every causal bivariate process is composed of two independent modules - represented by probabilities P_{cause} and $P_{effect|cause}$ - that do not influence each other. Furthermore, it is in principle possible to intervene on a particular module without perturbing the other.*

The above principle means that while one module's output may influence another module's input, the modules themselves are independent of each other. The notion of independence invoked has no statistical content but can be explained as follows. In principle, we can transport the module $P_{effect|cause}$ in a context where the input is provided by a module \hat{P}_{cause} different from P_{cause} . In other words, this modularity principle allows localized interventions and corresponds to the belief that there is no *meta-mechanisms* connecting different modules. Similarly, assuming this principle, we can expect that every module tends to remain invariant for changes stemming outside the system.

Causality as compression-in-representation. Principle 3 also admits an informational interpretation - not in statistical terms - based on the *Kolmogorov Complexity* K [89]. Given a *Universal Turing Machine* T and a binary string s , the Kolmogorov Complexity of s with respect to T is defined as

$$K_T(s) := |s^*|,$$

where $|s^*|$ denotes the length of the shortest program s^* for which T outputs s and stops. Remarkably, s^* can be regarded as the most concise compression of s that encompasses all the necessary information for executing the decompression process and producing the output s . Moreover, in what follows, subscripts in T will be omitted, and the Turing Machine T will be considered fixed once and for all. Similar to what was done for information in Shannon's sense, it is possible to define the conditional quantity $K(s|t)$ to be interpreted as the length of the shortest program that generates the output s from the input t and then stops. Therefore, *algorithmic mutual information* is defined as

$$I(s : t) := K(s) - K(s|t^*).$$

Moreover, it is possible to show that

$$I(s : t) \stackrel{\pm}{=} K(s) + K(t) - K(s, t),$$

where $\stackrel{\pm}{=}$ indicates that equation only holds up to a constant. Having defined these objects, let us now consider the variables X and Y assuming for a moment that $X \rightarrow Y$ is the correct causal link. Intuitively, X can be interpreted as a program that produces the output x . Similarly, the noise N selects the mechanism f_N provided by Principle 2, which corresponds to a program that takes x as input and produces

outputs y . If Principle 3 holds, then the program corresponding to X contains no information about the program that produces Y given the input x . In other terms, we can consider $p(x)$ and $p(y|x)$ as two programs that produce the correspondent probability distribution for every input x and y , to be considered as finite length strings. This interpretation can be formalized in terms of *algorithmic independence* [74], that is,

$$I(P_{\text{cause}} : P_{\text{effect}|\text{cause}}) \stackrel{\pm}{=} 0. \quad (54)$$

In other words - whenever (54) holds - the description of P_{cause} does not get shorter when $P_{\text{effect}|\text{cause}}$ is known, and vice versa. Moreover, condition (54) implies [129]

$$\begin{aligned} K(P_{\text{cause}}) + K(P_{\text{effect}|\text{cause}}) &\stackrel{\pm}{=} K(P_{\text{cause}|\text{effect}}) \\ &\stackrel{+}{\leq} K(P_{\text{effect}}) + K(P_{\text{cause}|\text{effect}}), \end{aligned}$$

and this relation says that the *causal factorization of P is a more compressed representation than the anti-causal one*. Remarkably, condition (54) considers the sum of marginal and conditional complexities and can *not* be substituted by a relation that compares only the conditional distribution, i.e.

$$K(P_{\text{effect}|\text{cause}}) \stackrel{+}{\leq} K(P_{\text{cause}|\text{effect}}),$$

that in general does *not* hold. In other words, Occam's Razor-type principle intervenes at the level of the available representation and not at the level of the mechanism generating the joint distribution, where the complexity of the mechanism for the conditional distribution could be offset for the mechanism generating the marginal one. At this point, two further remarks are valuable to complete the picture.

- Principle 3 holds also under a more weakened hypothesis. Let us add a node M that represents one part of the mechanism that connects X and Y , for which modularity could not hold. Graphically, we have three nodes with $X \rightarrow Y$ and the chain $X \rightarrow M \rightarrow Y$. However, Principle 3 applies to the extended graph. This remark makes it clear that Principle 3 has a dual nature. First, it encapsulates a physical statement about the relations between the observable variables. Second, it has a methodological content forcing the selection of variables and mechanisms that make possible localized interventions.
- Principle 3 is also applicable in considering a dynamic system. Let s be the initial state, $\phi^t s$ its evolution after a time t . Typically, we assume that the initial state s has no interactions with the underlying dynamics ϕ^t . This assumption is coherent with Principle 3 whenever we consider a fully concentrated measure in s and the time evolution as a functional mechanism, taking for granted that the system is isolated. If the system was not isolated, the external environment could produce a dependence between s and ϕ^t .

At a less fundamental level, given the framework provided by Principle 2, an occurrence of N can be interpreted as a selection of a deterministic mechanism that

produces the Y 's outcome starting from the input X . If X and N were dependent, the occurrence of N would contain a piece of information about the occurrence of X . Consequently, Principle 3 would be violated. This remark motivates the introduction of a principle to make Principle 2 compatible with 3:

Principle 4 *Given the functional mechanism provided by Principle 2, the noise variable N and the observational variable X must be independent.*

The above discussion can be adapted for a multivariate system. The general framework is intuitive if we start from a graph-driven approach with the above principles opportunely extended. Let X be a set of observables modeled as random variables associated with the nodes of a directed acyclic graph \mathcal{G} . The directed edge from X_i to X_j in \mathcal{G} means that X_i influences X_j directly: intervening on X_i changes the distribution of X_j if we consider the other variables fixed. We can extend the causal \mathcal{G} with the unexplained nodes N_k , adding all the directed edges from N_k to X_k . Let PA_k be the set of all variables in \mathcal{G} for which there exists an arrow to X_k . The extended graph tells us that PA_k and N_k produce the observable value of X_k through an underlying mechanism. Structural equations make explicit this mechanism by providing a deterministic function f_k , such that $X_k := f_k(PA_k, N_k)$ under the assumptions for which PA_k and N_k are statistically independent. Reconsidering Principle 3 and denoting with \mathcal{G} the underlying causal graph, the \mathcal{G} -factorization satisfies the following condition:

$$K[\mathcal{G}] := K\left(\prod_{i=1}^m P_{X_i|PA_i^{\mathcal{G}}}\right) \stackrel{+}{\leq} K\left(\prod_{i=1}^m P_{X_i|PA_i^{\mathcal{G}'}}\right), \quad \text{for every } \mathcal{G}' \neq \mathcal{G}.$$

Finally, an additional principle is necessary:

Principle 5 *Given the functional mechanism provided by Principle 2, the noise variables N_i and N_j must be independent.*

In other words, we assume that no hidden variables influence more than one of our observables in X . Given the distribution of N , the functional assignments allow us to compute the joint distribution of X , which has properties inherited from the \mathcal{G} 's topology: the density function of P_X admits a canonical factorization compatible with the modularity principle and a causal interpretation. More precisely, this construction has two assumptions of independence: each mechanism f_k is independent of PA_k distribution; secondly, an additional modularity principle is satisfied, that is, each f_k is independent of the others.

Causal Direction

Causal modeling techniques raise several methodological issues that are of interest both to philosophers and pragmatic users. Generally speaking, identifying and quantifying causal relations could be among the most relevant scientific goals since causality provides the theoretic foundation for operative knowledge in many contexts, also making possible reasonable human decisions. In what follows, we will deal with two

particular questions, both of interest from an epistemological perspective. Can causal conclusions be derived just from statistical information? If not, what additional background knowledge is required? In other words, given a bivariate distribution $P_{X,Y}$, we will discuss in some detail what formal conditions must be satisfied to have a well-defined causal relationship between X and Y .

A Symmetry result. It is commonly held that the relation between cause and effect is asymmetric. As opposed to mere statistical correlation, causation has directionality. Consequently, we say that an adequate theory of causality must explain this asymmetry, which we should recognize at some level when we try to extract causal information from statistical dependencies. In what follows, we will consider a formal argument to prove that the previous definition in terms of the functional model is not sufficient to recognize this fundamental asymmetry, as discussed in [127]. Let X and Y be two random variables with a joint probability distribution $P_{X,Y}$. Given $x \in \mathcal{X}$, we consider the cumulative distribution function of Y conditioned by $X = x$,

$$F_x(y) := \text{Prob}(Y \leq y | X = x).$$

F_x is a monotone function from \mathcal{Y} to $[0, 1]$. If we take the inf-value on each interval where $F'_x \equiv 0$, the inverse function is well-defined. We define $f(x, n) := F_x^{-1}(n)$ and $Z := f(X, N)$, where $n \in [0, 1]$ and N is an undefined random variable on $[0, 1]$. Fixed $y \in \mathcal{Y}$, we have

$$\begin{aligned} \text{Prob}(Z \leq y) &= \text{Prob}(f(X, N) \leq y) \\ &= \int \text{Prob}(f(X, N) \leq y | X = x) p_X(x) dx \\ &= \int \text{Prob}(f(x, N) \leq y) p_X(x) dx \\ &= \int \text{Prob}(F_x^{-1}(N) \leq y) p_X(x) dx \\ &= \int \text{Prob}(N \leq F_x(y)) p_X(x) dx. \end{aligned}$$

If we choose $N \sim U[0, 1]$, then $\text{Prob}(N \leq F_x(y)) = F_x(y)$ and we have

$$\begin{aligned} \text{Prob}(Z \leq y) &= \int F_x(y) p_X(x) dx = \int \text{Prob}(Y \leq y | X = x) p_X(x) dx \\ &= \int \text{Prob}(Y \leq y, X = x) dx \\ &= \text{Prob}(Y \leq y), \end{aligned}$$

concluding that $Y = f(X, N)$, with $X \perp\!\!\!\perp N$, i.e. exists a functional model $\Phi_{X \rightarrow Y}$. This construction can be repeated by swapping X and Y , with a deterministic function g and noise M such that $X = g(Y, M)$ and $Y \perp\!\!\!\perp M$, i.e. exists a functional model $\Phi_{Y \rightarrow X}$. In other words, we have the following symmetry result:

Theorem 2 *Let $P_{X,Y}$ be the observational joint distribution for X and Y . Without any additional prescriptions, we can not establish the asymmetry between cause and effect.*

It is well known that if X is a continuous random variable with a cumulative distribution function F_X , then $F_X(X)$ has a uniform distribution on $[0, 1]$. Consequently, given a noise N and its cumulative distribution function F_N , we can construct a functional model $Y = \hat{f}(X, N)$ using the argument presented above, by defining $\hat{f}(X, N) := f(X, F_N^{-1}(N))$ with $f(x, n) := F_X^{-1}(n)$ and $n \in [0, 1]$. In other words, the statement of Theorem 2 is true also when the noise distribution is a priori fixed.

Example 1 *Let $P_{X,Y}$ be a Gaussian bivariate distribution given by the functional model $Y = \alpha X + N$, where $\alpha \neq 0$, $P_X \sim N(0, \sigma_X^2)$ and $P_N \sim N(0, \sigma_N^2)$. We can construct the backward model $X = \beta Y + M$. We define*

$$M := X - \beta Y$$

and determine $\beta \in \mathbb{R}$ such that $Y \perp\!\!\!\perp M$. Stochastic independence and zero correlation are equivalent conditions for Gaussian variables, so we impose $\text{cov}(Y, M) = 0$. Then we have

$$\begin{aligned} 0 &= \text{cov}(Y, M) = \text{cov}(Y, X - \beta Y) = \text{cov}(Y, X) - \beta \sigma_Y^2 \\ &= \text{cov}(\alpha X + N, X) - \beta \sigma_Y^2 = \alpha \sigma_X^2 - \beta \sigma_Y^2 \\ &= \alpha \sigma_X^2 - \beta(\alpha^2 \sigma_X^2 + \sigma_N^2). \end{aligned}$$

Then we have the backward model $X = \frac{\alpha \sigma_X^2}{\alpha^2 \sigma_X^2 + \sigma_N^2} Y + M$.

The Example 1 is coherent with the symmetry result presented above, although it tells us something more. The functional models $\Phi_{X \rightarrow Y}$ and $\Phi_{Y \rightarrow X}$ have the same structure, linear with additive Gaussian noise. Let's try to remove the Gaussian hypothesis on $P_{X,Y}$, only assuming that exist the functional linear models $\Phi_{X \rightarrow Y}$ and $\Phi_{Y \rightarrow X}$:

$$\begin{aligned} Y &= \alpha X + N, & X &\perp\!\!\!\perp N, & \alpha &\neq 0, \\ X &= \beta Y + M, & Y &\perp\!\!\!\perp M, & \beta &\neq 0, \end{aligned}$$

so that

$$\begin{cases} Y = \alpha X + N, \\ M = (1 - \alpha\beta)X - \beta N, \\ Y \perp\!\!\!\perp M. \end{cases} \quad (55)$$

For $\alpha\beta = 1$ we conclude that $\alpha X + N \perp\!\!\!\perp N$, which is absurd if $\sigma_N^2 \neq 0$.¹ So $1 - \alpha\beta \neq 0$ and from the relation (55) we conclude that X and N are normal distributed². In other words, we have shown the following lemma:

¹If $X + Y \perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp Y$, then $\text{cov}(X + Y, Y) = \text{cov}(Y, Y) = \sigma_Y^2$, while zero-covariance is a necessary condition for independence.

²The following lemma is well known. *Let Z_1 and Z_2 be two independent variables such that $Z_1 = a_1 X + b_1 Y$ and $Z_2 = a_2 X + b_2 Y$, with $a_1, a_2, b_1, b_2 \neq 0, Z_1 \perp\!\!\!\perp Z_2$. Then X and Y are normally distributed.*

Lemma 1 *Let $P_{X,Y}$ the observational distribution and let $\Phi_{X \rightarrow Y}$ be a functional linear model for $P_{X,Y}$, with additive noise. Exists $\Phi_{Y \rightarrow X}$ linear with additive noise if and only if $P_{X,Y}$ is Gaussian.*

This Lemma suggests that by assigning some structural prescriptions on the functional model Φ , we could establish the cause-effect asymmetry. In the next section, we will explore this idea by studying two Φ -classes, with additive and multiplicative noise respectively.

Breaking the Symmetry

Let $P_{X,Y}$ be the observational joint distribution for which density $p(x, y)$ generally admits two factorizations, $p(y|x)p(x)$ and $p(x|y)p(y)$. Without additional requirements, $p(y|x)$ and $p(x|y)$ correspond to two different functional relationships, $Y = f(X, N)$ and $X = g(Y, M)$ respectively, so that f and g can not admit a causal interpretation. However, if our background knowledge about the underlying mechanisms introduces some formal constraints on f and g , then the symmetry between X and Y results be broken.

Additive Noise

Let A_Φ be the class of functional models with *additive noise*. Given the observational distribution $P_{X,Y}$, we assume that $\Phi_{X \rightarrow Y} \in A_\Phi$, that is to say exists (f, N) such that $Y = f(X) + N$, equipped with the independence condition $X \perp\!\!\!\perp N$. Under the additivity hypothesis, a fitting procedure is enough. Given

$$\hat{f}_Y(x) := \mathbb{E}[Y|X = x],$$

we can conclude that $\Phi_{X \rightarrow Y} \in A_\Phi$ is the correct causal model whenever the regression residual $Y - \hat{f}_Y(X)$ is independent of X , provided that the regression residual in the opposite direction is not independent of Y . This circumstance can be verified empirically by plotting the pairs (x_i, n_i) - where (x_i, y_i) are the observed data with $n_i = y_i - \hat{f}_Y(x_i)$ - to check for correlation. Theoretically - this time in informational terms - additive noise allows the following characterization to be formulated:

$$H[X] - H[Y - \hat{f}_Y(X)] \leq H[Y] - H[X - \hat{f}_X(Y)]$$

Reconsidering the level of structural equations, an asymmetry result can be obtained via the following reasoning. First of all, additivity implies

$$p(x, y) = p_X(x)p_N(y - f(x)).$$

If $p(x, y)$ is strictly positive, we define the self-information $h(x, y) := \ln p(x, y)$ and compute its derivatives:

$$\begin{aligned} h(x, y) &= \ln p_X(x) + \ln p_N(y - f(x)) \\ &=: \varphi(x) + \gamma(y - f(x)), \end{aligned}$$

$$\partial_y^2 h = \gamma'', \quad \partial_x \partial_y h = -f' \gamma''.$$

Given $\Omega := \{(x, y) : f' \gamma'' \neq 0\}$, h satisfies the following equation on Ω :

$$\partial_y \left\{ \frac{\partial_y^2 h}{\partial_x \partial_y h} \right\} = 0. \quad (56)$$

Assuming that exists $\Phi_{Y \rightarrow X} \in A_\Phi$, we have

$$X = g(Y) + M, \quad Y \perp M$$

and

$$p(x, y) = p_Y(y) p_M(x - g(y)).$$

Consequently, $h = \psi + \delta$ with $\psi(y) := \ln p_Y(y)$, $\delta(x - g(y)) := \ln p_M(x - g(y))$. Replacing $h = \psi + \delta$ in (56) and assuming $\delta'' g' \neq 0$, we obtain the differential equation

$$\begin{aligned} & \{\psi''' - \delta'''(g')^3 + 3\delta'' g' g'' - \delta' g'''\} \{\delta'' g'\} + \\ & + \{-\delta'''(g')^2 + \delta'' g''\} \{\psi'' + \delta''(g')^2 - \delta' g''\} = 0, \end{aligned}$$

which could admit local solution around y_0 , fixed $g(y_0)$, $g'(y_0)$, $g''(y_0)$ and given p_Y and p_M . This fact could follow directly from the Cauchy theorem for differential equations like $g''' = G(g, g', g'')$, with some regularity assumption for G . However, our construction requires a global solution g , well defined on the real axis because of the positivity condition on p_Y . Generically, a global solution g does not exist. To prove this fact, we can reason as follows. Using the argument just presented and starting from $h = \psi + \delta$ we obtain the (56)-type condition

$$\partial_x \left\{ \frac{\partial_x^2 h}{\partial_x \partial_y h} \right\} = 0.$$

We use $h = \varphi + \gamma$ directly in the previous relationship and obtain the following differential equation on Ω , where y is a parameter:

$$\varphi'''(x) = A(x, y) \varphi''(x) + B(x, y), \quad (57)$$

with

$$\begin{aligned} A(x, y) &:= \frac{\gamma'''(f')^2 - \gamma'' f''}{\gamma'' f'}, \\ B(x, y) &:= \frac{\{\gamma'''(f')^2 - \gamma'' f''\} \{\gamma''(f')^2 - \gamma' f'''\} +}{\gamma'' f'} \\ &+ \gamma'''(f')^3 - 3\gamma'' f' f'' + \gamma' f'''. \end{aligned}$$

In other words, the linear differential equation (57) produces a strong constraint on φ . Under the assumption that exists the backward functional model in A_Φ , P_X necessarily lives in the 3-dimensional flat space

$$S_+ := \left\{ p_X > 0 : \int p_X dx = 1, \quad \varphi''' = A\varphi'' + B \quad \text{on} \quad \Omega \right\},$$

while the observational distribution $P_{X,Y}$ has in general the density $p(x, y) = p_X(x)p_N(y - f(x))$ with $p_X(x)$ in some infinite dimensional space, for example $L^1(\mathbb{R}) \cap L^2(\mathbb{R})$. In other words, we have shown the following theorem:

Theorem 3 *Let $P_{X,Y}$ be the observational joint distribution, with strictly positive density. If exists a functional model $\Phi_{X \rightarrow Y} \in A_\Phi$, generically does not exist a backward model $\Phi_{Y \rightarrow X} \in A_\Phi$.*

Example 2 *Let $P_{X,Y}$ be the joint observational distribution, with $Y = \alpha X + N$, $X \perp N$ and N Gaussian. Then $f'' = 0$, $p_N(n) = C_1 \exp\{C_2 n^2\}$, $\gamma(n) = \ln C_1 + C_2 n^2$, for some constants C_1 and C_2 . The equation (57) becomes $\varphi''' = 0$, so that p_X is required to be Gaussian coherently with Lemma 1.*

Example 3 *If $f(x) = \alpha x^2$ and N is Gaussian, then the equation (57) becomes $\varphi''' = C_1 \frac{1}{x} \varphi'' + C_2 \frac{1}{x} + C_3 x$, for some constants C_1, C_2 and C_3 . If it was $C_1 = C_2 = C_3 = 1$, then $\varphi(x) = \frac{x^4}{12} + Ax^3 - \frac{x^2}{2} + Bx^2 + C$ and $p_X(x) = \ln \varphi(x)$ would not be a probability distribution because the coefficient of x^4 is positive. In this particular circumstance, no p_x would be compatible with the existence of the backward model.*

Remark 1 *Given p_x and p_N , the equation (57) provides a constraint on f so that the backward model exists. If X and N are normal distributed, then $\varphi''' \equiv \gamma''' \equiv 0$, so that the equation (57) becomes*

$$C_1 f' f'' + C_2 \frac{f''}{f'} - \gamma' \left\{ f''' + \frac{(f'')^2}{f'} \right\} = 0,$$

where C_1, C_2 are some constants and $f' \neq 0$. Obviously, $f'' \equiv 0$ is a solution, so linearity is compatible with the backward model's existence. We can prove that the null function is the only solution well-defined on the real axis. Let Ω be the open set $\{x : f'' \neq 0\} \times \mathbb{R}$, so that the previous equation becomes

$$C_1 f' + C_2 \frac{1}{f'} - \gamma' \left\{ \frac{f'''}{f''} + \frac{f''}{f'} \right\} = 0$$

on Ω . Defined $\Omega_1 := \{(x, y) : y = f(x)\}$, $\gamma' \equiv 0$ on Ω_1 so that $f' = C_3$ on $\Omega \cap \Omega_1$, which is absurd. In conclusion, the backward model's existence with p_X and p_N normal distributed is compatible only with linear models.

Extending the Asymmetry Class

We are in the following situation. If we only consider the additive class A_Φ we have an asymmetry result; on the other hand, no prescriptions on Φ mean a complete symmetry between X and Y . We can gradually extend the asymmetry class A_Φ by considering other functional prescriptions. Let M_Φ be the class of functional models with multiplicative noise, that is to say, $\Phi_{X \rightarrow Y} \in M_\Phi$ if and only if has the structure

$Y = f(X) \cdot N$, with $X \perp N$. Let $P_{X,Y}$ a observational distribution such that admits $\Phi_{X \rightarrow Y}, \Phi_{Y \rightarrow X} \in M_{\Phi}$. Consequently:

$$\begin{aligned} X &= g(Y) \cdot M, \quad Y \perp M, \quad g \neq 0, \\ p(x, y) &= p_Y(y) p_M\left(\frac{x}{g(y)}\right), \\ h(x, y) &= \ln p(x, y) = \ln p_Y(y) + \ln p_M\left(\frac{x}{g(y)}\right) \\ &=: \psi(y) + \delta\left(\frac{x}{g(y)}\right) \\ \partial_x^2 h &= \delta'' \frac{1}{g^2}, \quad \partial_x \partial_y h = -\delta'' x \frac{g'}{g^3} - \delta' \frac{g'}{g^2}. \end{aligned}$$

Similarly to what was done in the previous section, we obtain the following condition on h

$$\partial_x \left\{ \frac{\partial_x (x \partial_x h)}{\partial_x \partial_y h} \right\} = 0. \quad (58)$$

Given $\varphi(x) := \ln p_x(x)$, condition (58) produces a new differential equation $\varphi''' = F(\varphi, \varphi'', \varphi''', x; y)$, by proceeding as follows. The distribution $P_{X,Y}$ admits $\Phi_{X \rightarrow Y} \in M_{\Phi}$, so that

$$\begin{aligned} Y &= f(X) \cdot N, \quad X \perp N, \quad f \neq 0, \\ h(x, y) &= \ln p_X(x) + \ln p_N\left(\frac{y}{f(x)}\right) =: \varphi(x) + \gamma\left(\frac{y}{f(x)}\right). \end{aligned}$$

Given $h = \varphi + \gamma$ and $k := f'/f^2$, we have:

$$\begin{aligned} \partial_x h &= \varphi' - yk\gamma', \\ \partial_x^2 h &= \varphi'' - yk'\gamma' + y^2 k^2 \gamma'', \\ \partial_x \partial_y h &= -k\gamma' - yk \frac{1}{f} \gamma''. \end{aligned}$$

If we consider the equation (58) on the open set

$$\Omega := \{(x, y) : x \neq 0, \quad k\gamma' + yk \frac{1}{f} \gamma'' \neq 0\},$$

we obtain

$$\begin{aligned} &x\varphi''' + 2\varphi'' + x(y^3 k^3 \gamma''' - y^2 k k' \gamma'' - y k'' \gamma') = \\ &= \{x\varphi'' + \varphi' - x(y^2 k^2 \gamma'' + yk'\gamma' + \frac{1}{x} k \gamma')\} \left\{ \frac{k'}{k} - ky \frac{y \frac{1}{f} \gamma''' + 2\gamma''}{y \frac{1}{f} \gamma'' + \gamma'} \right\}. \end{aligned}$$

The last equation can be written in a normal form for φ , that is to say,

$$\varphi'''(x) = A(x, y)\varphi''(x) + B(x, y)\varphi'(x) + C(x, y)$$

for some functions A, B and C . This linear differential equation produces a strong constraint on φ : under the assumption that exists the backward functional model in M_Φ , P_X necessarily lives in the 3-dimensional flat space

$$S_\times := \left\{ p_X > 0 : \int p_X dx = 1, \quad \varphi''' = A\varphi'' + B\varphi' + C \quad \text{on } \Omega \right\},$$

while the distribution $P_{X,Y}$ has in general the density $p(x, y) = p_X(x)p_N(\frac{y}{f(x)})$ with $p_X(x)$ in some infinite dimensional space, for example $L^1(\mathbb{R}) \cap L^2(\mathbb{R})$. In other words, we have shown the following theorem:

Theorem 4 *Let $P_{X,Y}$ be the observational joint distribution, with strictly positive density. If exists a functional model $\Phi_{X \rightarrow Y} \in M_\Phi$, generically does not exist a backward model $\Phi_{Y \rightarrow X} \in M_\Phi$.*

With the same technique, we can also discuss mixed cases. For example, assuming that $\Phi_{X \rightarrow Y} \in A_\Phi$ we ask ourselves whether exists $\Phi_{Y \rightarrow X} \in M_\Phi$. If existed $\Phi_{Y \rightarrow X} \in M_\Phi$, then the equation (58) would hold. Because of the existence of $\Phi_{X \rightarrow Y} \in A_\Phi$, we could substitute $h(x, y) = \varphi(x) + \gamma(y - f(x))$ in (58) obtaining a constraint for φ as in the previous sections. In other words, given $N_\Phi := A_\Phi \cup M_\Phi$ we can prove that if exists a functional model $\Phi_{X \rightarrow Y} \in N_\Phi$, generically does not exist a backward model $\Phi_{Y \rightarrow X} \in N_\Phi$.

4 State of the Art

The first part of this section is devoted to *Causal Discovery*, and three significant algorithms from this research area will be examined below. On the other side, the last subsection will consider IDA, a *Causal Inference* protocol that starts from observational data. These four algorithms do not exhaust the fields of study in which they arise, and providing a comprehensive overview in this regard would be prohibitive here. However, they represent the simplest and most archetypal examples of the majority of the approaches adopted, providing a fairly comprehensive overall understanding.

Before going into more detail, some general considerations may be useful. First, Causal Discovery aims to extract causal relations starting from the statistical properties of observational data, typically by providing us with the DAG \mathcal{G} that best agrees with the underlying distribution P . Second, in this regard, an adequacy criterion is to be specified: the most used methods can be categorized into two sub-classes, *score-based* and *constraint-based*, so that both approaches will be considered, trying to clarify working hypotheses and relevant differences. Third, Causal Inference methods typically assume prior knowledge about the system's causal structure. From this point of view, IDA tries to combine the estimation of the equivalence class of DAGs with causal inference methods that can be used when the DAG is known. This combination tries to respond to a need that has already emerged in previous sections: it is often unrealistic to assume that the graph structure among the variables of interest is a priori known.

The PC Algorithm

Let us start with the first and simplest causal discovery algorithm, PC, proposed by P. Spirtes and C. Glymour [169]. The main merit of this method lies in exhibiting an efficient strategy in employing independence relations among variables. To make this point clear, suppose we have an oracle that can provide us with a “yes” or a “no” w.r.t. any conditional (in)dependence question. Even assuming we have this oracle, the total number of possible independence questions grows exponentially with the number of variables. Consequently, a naive brute-force approach is to be ruled out, and a more intelligent procedure in asking questions of the oracle must necessarily be adopted. The following procedure is computationally efficient under assumptions that will become explicit shortly.

First phase - Finding the skeleton.

- Construct a complete undirected graph \mathcal{G} between all nodes in \mathbf{X} .
- Given A, B connected in \mathcal{G} , delete $A-B$ if it exists $\mathbf{C} \subseteq Adj(A)$ or $\mathbf{C} \subseteq Adj(B)$ with **size** \mathbf{S} (initially equal to 0) s.t. $A \perp\!\!\!\perp B | \mathbf{C}$. When found, memorize $\mathbf{C}(A, B)$ ³.
- Add 1 to \mathbf{S} and repeat step 2 until $\mathbf{S} > MaxDegree(\mathcal{G})$.

Second phase - Exploiting the v-structures.

- Given A, B not connected in \mathcal{G} , define $\mathbf{V} := (Adj(A) \cap Adj(B)) - \mathbf{C}(A, B)$.
- For every node $V \in \mathbf{V}$, direct $A \rightarrow V \leftarrow B$.

Third phase - Apply *Orientation Rules* as in Figure 16.

Two important remarks should be noted. First, the computational complexity of the algorithm primarily depends on the maximum degree, denoted as $MaxDegree(\mathcal{G})$, of the underlying graph \mathcal{G} . In the worst-case scenario, i.e., for a complete DAG, it becomes necessary to conduct all possible independence tests. The efficiency of the algorithm is ensured when $MaxDegree$ grows at most logarithmically with the number of variables, denoted as N . Second, it is crucial to understand that not all edges will be reliably directed at the conclusion of the procedure. The output of the algorithm will generally be a completed partially directed acyclic graph (CPDAG), representing a specific Markov equivalence class (MEC). This implies that the resulting graph captures the essential conditional independence relationships among variables but may still have undirected edges, indicating unresolvable ambiguities in the underlying causal structure.

³It is easy to show that in any DAG, given any two non-adjacent nodes A and B , there is always a set \mathbf{C} contained either in $Adj(A)$ or in $Adj(B)$ d-separating them. A demonstration sketch follows: Consider the partial order of nodes induced by the DAG, where $X > Y$ if and only if there exists a direct path from X to Y . If neither $A > B$ nor $A < B$, it is sufficient to take the set $\mathbf{C} = \{C \in Adj(A) | C > A\}$. If on the other hand $A > B$, $\mathbf{C} = \{C \in Adj(A) | C > A \vee A > C > B\}$.

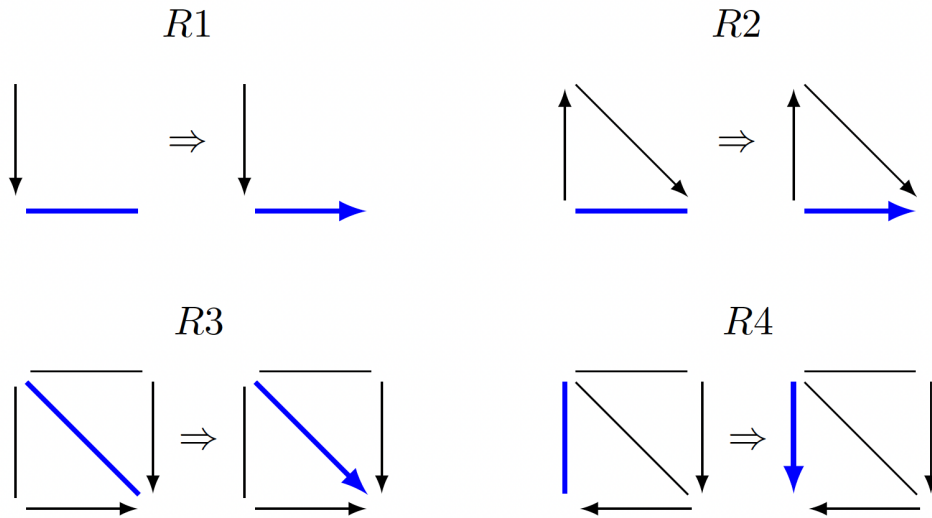


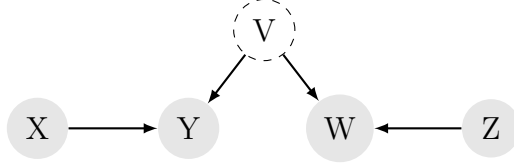
Figure 16: Orientation Rules defined in [102].

It is important to specify how PC can only be applied under the assumption of *causal sufficiency*, and why it is not able to detect the presence and/or identify latent confounders. Applying PC to a system in which latent confounders actually exist, one could come across a situation in which, due to the orientation of certain arcs, these predict contradictory results with respect to the set of possible independence tests on the set of variables. In short, there may not even be a DAG defined on the observed variables alone capable of perfectly representing the conditional independences. To solve this problem, it is then necessary to extend the space of graphical representations by introducing new semantics on the arcs.

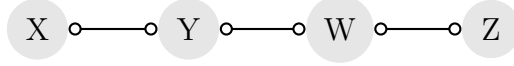
The FCI Algorithm

The assumption of Causal Sufficiency is often exaggerated and should be interpreted more as a form of approximation than as a true attainable condition. To think that one is in possession of literally *all* the variables of the system, in a nutshell, is always a stretch. This approximation is not always acceptable and, particularly when one is interested in the interventional and counterfactual levels, it may be necessary to do without it [144]. FCI [171, 170] departs from PC in exactly this respect, extending the expressive capacity of the graph representation of the system and adding the possibility of detecting latent confounders.

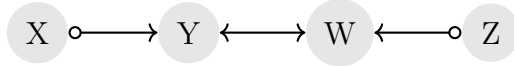
An edge between A and B in CPDAGs resulting from PC could be of three types: $A \rightarrow B$ (A "causes" B), $A \leftarrow B$ (B "causes" A), $A - B$ (A "causes" B XOR B "causes" A). It was precisely Causal Sufficiency that ruled out the absence of a direct causal relationship, always excluding the confounder between A and B . FCI extends the representation by adding circles to the extremes of the edge, which stays for the uncertainty about the two possible concrete states, i.e. the presence or the absence



The underlying DAG to be found. Note that V is hidden.



The initial DAG skeleton found using FCI.



The result of FCI after the orienting phase.

Figure 17: Intuition of how FCI works. $Y \leftrightarrow W$ is produced by "testing" both the v -structures $X \rightarrow Y \leftarrow W$ and $Y \rightarrow W \leftarrow Z$. PC would only find one of them, not testing the other one.

of the arrow. For example, $A \circ \circ B$ indicates a totally agnostic condition and may saturate in any of the PC states ($A \rightarrow B$, $A \leftarrow B$, $A - B$), but also in the new states $A \leftrightarrow B$ (there is a confounder between A and B). Similarly, the partially saturated states $A \circ \rightarrow B$ is agnostic with respect to the choice between the identified arrow $A \rightarrow B$ and the confounder $A \leftrightarrow B$. By means of this extended set of symbols, we can hope to extract additional information. Figure 17 shows intuitively how the algorithm works.

FCI follows a more complex procedure than PC and only partially shares its techniques. In general, it is not sufficient to modify PC by replacing the normal $-$ edges with $\circ \circ$ in the skeleton, proceeding with similar orientation rules. One of the most important problems when one has to consider latent confounders is the following: Consider a DAG $\mathcal{G} = \{\mathbf{V}, E\}$ of which only a subset of the nodes $\mathbf{O} \subset \mathbf{V}$ is observed. \mathcal{G} is faithful to the probability $P(\mathbf{V})$, however the accessible one is its marginal

$$P(\mathbf{O}) = \int_{\mathbf{L}} P(\mathbf{v}) d\mathbf{l} = \int_{\mathbf{L}} P(\mathbf{O}|\mathbf{l})P(\mathbf{l})d\mathbf{l}, \quad \mathbf{L} := \mathbf{V} - \mathbf{O}. \quad (59)$$

Considering the graph \mathcal{G}' obtained by applying PC to $P(\mathbf{O})$, it is possible that new dependencies are introduced with respect to the original ones in \mathcal{G} . In general, it holds that if $\mathbf{O} \subset \mathbf{V}$ is not causally sufficient, then **it is not the case** that, given any two nodes $X, Y \in \mathbf{O}$, conditional dependence on every subset of $\mathbf{O} - \{X, Y\}$ implies:

1. X is a direct cause of Y , or

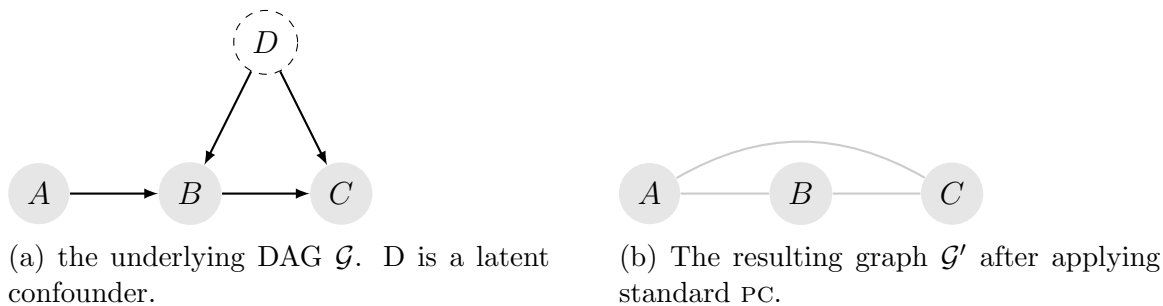


Figure 18: d-separation relations are not preserved marginalizing over hidden variables. Figure inspired by [170].

2. Y is a direct cause of X , or
3. there is a common cause $Z \in \mathbf{V}$ s.t. $X \leftarrow Z \rightarrow Y$.

To see this fact, consider Figure 18: the presence of a single latent confounder causes all tests of conditional independence between A and C to fail, making the two nodes adjacent in the graph obtained by applying PC. Note that in the original graph, A and C are not adjacent and have no common causes.

The FCI algorithm can be described through the following steps:

- Finding the skeleton, following the same procedure as PC.
- Orienting all edges as $\circ-\circ$.
- Exploiting the v-structures (as in PC).
- For every edge in the skeleton from A to B , finding the sets $PossibleDsep(A, B)$ and $PossibleDsep(B, A)$ (not necessarily parents of either A or B as in PC) and removing edges for which a d-separation set is found [170].
- Exploiting the v-structures again.
- Applying *Orientation Rules*.

In general, FCI has a higher computational complexity compared to PC and is capable of recovering less causal information about the system, in line with the fact that its underlying assumptions are weaker. In causal discovery, the tradeoff between the strength of assumptions and the inferential capacity of the results is particularly clear, remaining constant across the spectrum of existing methods.

The GES Algorithm

GES [37] (Greedy Equivalence Search) is a score-based algorithm that implements a greedy search in the space of MECs. The algorithm is based on certain theoretical results, which we will quickly describe.

Let \mathcal{G} and \mathcal{H} be two DAGs. \mathcal{H} is an independence map of \mathcal{G} iff any independence implied by the structure of \mathcal{H} is also implied by the structure of \mathcal{G} . We use $\mathcal{G} \leq \mathcal{H}$ to denote that \mathcal{H} is an independence map of \mathcal{G} . The symbol “ \leq ” is meant to express the fact that if $\mathcal{G} \leq \mathcal{H}$ then \mathcal{H} contains more edges than \mathcal{G} . The following result holds.

Theorem 5 ([36]) *Let \mathcal{G} and \mathcal{H} be two DAGs s.t. $\mathcal{G} \leq \mathcal{H}$. Then there exists a finite sequence of edge additions and edge reversals that can be applied to \mathcal{G} with the following properties:*

- *After each edge change, \mathcal{G} is a DAG and \mathcal{H} remains an independence map of \mathcal{G} .*
- *After all edge changes $\mathcal{G} = \mathcal{H}$.*

A special case is when \mathcal{G} is the graph with no edge, for which always holds $\mathcal{G} \leq \mathcal{H}$. Starting from the “empty” graph, it is always possible to reach any DAG through a finite sequence of elementary steps. Moreover, all intermediate DAGs \mathcal{G}' on the path satisfy $\mathcal{G}' \leq \mathcal{H}$.

The idea of GES is to reach a specific \mathcal{H} in the space of DAGs, which is intuitively *the most simple DAG model capable of reproducing the input probability distribution*. Identifying \mathcal{H} can be done by greedy searching for a DAG model that fits a set of observed data \mathcal{D} well, maximizing some scoring criterion $S(\mathcal{G}; \mathcal{D})$. Let \mathcal{D} be a collection of m i.i.d. samples from the probability distribution P . A scoring criterion $S(\mathcal{G}; \mathcal{D})$ is consistent if, in the limit as m goes to infinity, the following two properties hold:

- If \mathcal{H} can reproduce P exactly while \mathcal{G} can't, then $S(\mathcal{G}; \mathcal{D}) < S(\mathcal{H}; \mathcal{D})$.
- If both \mathcal{G} and \mathcal{H} can reproduce P , and \mathcal{H} contains more parameters than \mathcal{G} , then $S(\mathcal{G}; \mathcal{D}) > S(\mathcal{H}; \mathcal{D})$.

The original algorithm chooses a specific scoring criterion, BIC (Bayesian Information Criterion), defined through the formula

$$S_B(\mathcal{G}; \mathcal{D}) = \log P(\mathcal{G}) + \log P(\mathcal{D}|\mathcal{G}),$$

and well approximated using Laplace's method in the $m \gg 1$ regime by

$$\hat{S}_B(\mathcal{G}; \mathcal{D}) = \log P(\mathcal{D}|\mathcal{G}, \hat{\boldsymbol{\theta}}) - \frac{|\boldsymbol{\theta}|}{2} \log m,$$

where $\hat{\boldsymbol{\theta}}$ denotes the maximum-likelihood values for the model parameters, $|\boldsymbol{\theta}|$ denotes the number of free parameters of \mathcal{G} , and m is the number records in \mathcal{D} .

Moreover, BIC is decomposable into independent modules, one for each conditional probability $P(X_i|Pa(X_i))$. Exploiting this property, BIC is also *locally consistent*, i.e. Given two DAGs \mathcal{G} and \mathcal{G}' resulting from adding a single edge $X_i \rightarrow X_j$ to \mathcal{G} ,

- if $X_i \perp\!\!\!\perp X_j | Pa^{\mathcal{G}}(X_j)$, then $S_B(\mathcal{G}; \mathcal{D}) > S_B(\mathcal{G}'; \mathcal{D})$,
- else $S_B(\mathcal{G}; \mathcal{D}) < S_B(\mathcal{G}'; \mathcal{D})$.

These properties are sufficient to prove the optimality of GES, characterized by two successive phases. Starting from the graph with no edge, the single arrow that locally maximizes \mathcal{S}_B is added step by step. When there are no more arrows to be added that can increase the overall score, the first phase is terminated. Exploiting the definition of consistency, it is possible to show that the graph obtained after the first stage is definitely able to reproduce the probability P , given as input to the algorithm. Starting from the graph obtained, arrows that once removed increase \mathcal{S}_B are eliminated one by one. Having completed this second stage, the graph preserves the ability to reproduce P and, at the same time, is “as simple as possible”.

It is useful to add a few observations. First, GES, like the Causal Discovery algorithms seen above, is not able to discern between graphs of the same MEC. In fact, it is easy to show that, given any two $\mathcal{G} \equiv \mathcal{H}$, $S_B(\mathcal{G}; \mathcal{D}) = S_B(\mathcal{H}; \mathcal{D})$. The DAG obtained by executing GES is only one representative of its MEC, “selected” randomly from those in the class. Second, the optimality of the algorithm *is only guaranteed* assuming perfect knowledge of the probability P , i.e. in the regime of infinite data. GES is therefore *asymptotically consistent*. As a final point, GES assumes causal sufficiency.

The IDA Algorithm

IDA (Intervention-calculus when the DAG is Absent) [97, 96, 77] is a method that allows the estimation of causal effects under the condition of *partial uncertainty*, knowing the MEC but not the “true” underlying DAG. Its relevance is evident, considering that this condition is the typical one encountered after applying common Causal Discovery methods. IDA requires causal sufficiency i.e., the absence of latent confounders, taking as input the CPDAGs resulting from, for example, PC.

To describe how the method works, it is best to observe it in practice: Consider the CPDAG in figure 19. PC identified a v-structure in the graph, saturating the directions $X_4 \rightarrow Y$ and $X_3 \rightarrow Y$. The three undirected edges can be oriented in either direction, admitting in principle $2^3 = 8$ different DAGs belonging to the MEC. Some of these, however, would introduce new v-structures, creating colliders on X_2 or X_1 , and must therefore be discarded. The DAGs actually in the MEC are four in total, depicted in figure 20. Each of these induces a different set of causal relationships between variables. The basic idea of IDA is to estimate the causal effects for each DAG and report each result in a tensor Θ , in which the element Θ_{ij}^k can be described as *the causal effect of the i -th variable on the j -th, in the k -th DAG of the MEC*⁴.

⁴In [97], the authors assume that the system’s variables are jointly gaussian, which guarantees the independence of the average causal effect between two variables w.r.t. the specific value they

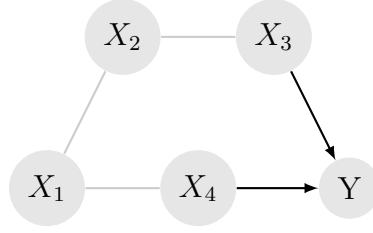


Figure 19: The initial CPDAG.

It is useful at this point to introduce the pivotal tool of the method, namely the use of the *valid adjustment set*.

Valid Adjustment Set. Consider a DAG \mathcal{G} and two variables $X, Y \in N(\mathcal{G})$. We say that a set $\mathbf{Z} \subseteq N(\mathcal{G})$ is a valid adjustment set relative to (X, Y) if it holds that

$$P(y|do(x)) = \int_{\mathbf{z}} P(y|x, \mathbf{z})P(\mathbf{z})d\mathbf{z}. \quad (60)$$

There can generally be several valid adjustment sets. In particular, IDA uses $Pa(X)$, since it also possesses the desirable property of being extremely local. Given a specific DAG, once identified its $Pa(X)$, it is therefore immediate to calculate the causal effect of X on Y , by applying (60). In general, the number of different causal effects obtainable is determined by the different ways of choosing $Pa(X)$ in the MEC. It is reasonable to expect therefore, especially for graphs with many variables, many 'repeated' causal effects.

Let us focus now on the causal effect of X_1 over Y evaluating Θ_{X_1Y} . We can immediately see that among the four DAGs of figure 20, two of them have the same set of parents for X_1 , thus resulting in the exact same value for $\Theta_{X_1,Y}$.

The first version of the proposed method in [97], called *Global IDA* or *Population Version*, lists each DAG in the MEC, computing its causal effect and storing the result obtained in the **multiset**

$$\Theta_{X_1,Y} = \{\Theta_{X_1,Y}^1, \dots, \Theta_{X_1,Y}^k, \dots, \Theta_{X_1,Y}^D\},$$

where D stands for the total number of DAGs in the MEC. It is possible to consider equivalently **a set of unique causal effects**, defined as $\hat{\Theta}_{X_1,Y}$, associating each causal effect $\hat{\Theta}_{X_1,Y}^k$ with its *multiplicity* $M(\hat{\Theta}_{X_1,Y}^k)$ i.e., the number of DAGs in which it occurs. Formally, we map

$$\Theta \mapsto \{\hat{\Theta}, M(\hat{\Theta})\}.$$

assume. In a nutshell, if in the general case Θ_{ij}^k would be a function of x_i , with this assumption we are reduced to a single scalar. This step is not obligatory and only serves as a function of conceptual and computational simplification. For the purposes of understanding the method, we will keep this clarification in the background, without elaborating on it.

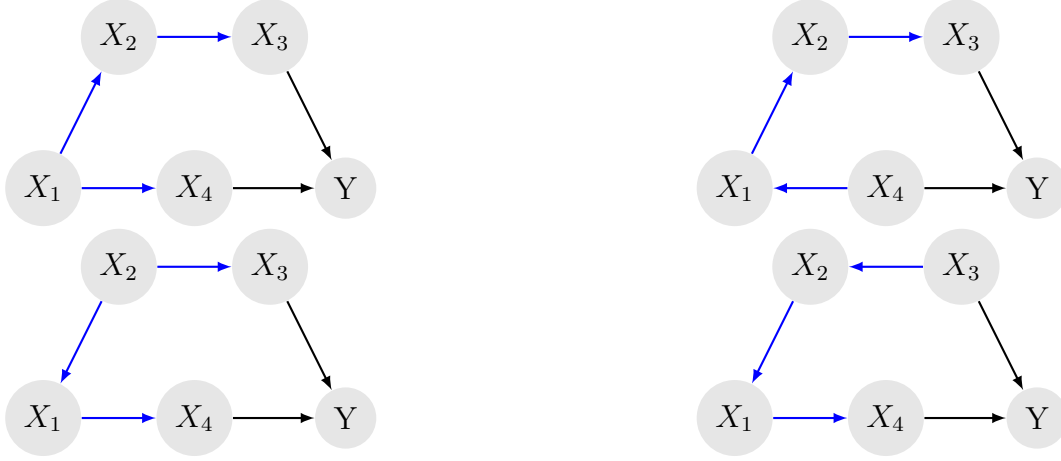


Figure 20: The four different DAGs of the MEC.

The main problem of this approach lies in the exponentially growth of the MEC’s cardinality w.r.t. the number of covariates. Hence, according to [97], the method becomes computationally unfeasible when the number of covariates is greater than twelve.

To overcome this limitation, the second version of the algorithm, *Local IDA*, bypasses the DAG enumeration phase through the following result:

Theorem 6 *Consider the CPDAG G and let $\mathbf{S} \subseteq \text{Sib}(X)$ in \mathcal{G} . $G_{\mathbf{S} \rightarrow X}$ is locally valid (i.e., it has no new v -structure with X as collider) if and only if there is a DAG \mathcal{G}^k in the equivalence class of G such that $\text{Pa}(X)_{\mathcal{G}} = \text{Pa}(X)_{\mathcal{G}^k} \cup \mathbf{S}$.*

Exploiting the theorem, the search for all the unique causal effects in $\hat{\Theta}_{X_1, Y}$ is greatly simplified. It is sufficient to list all the possible $\mathbf{S} \subseteq \text{Sib}(X_1)$ and check which ones are locally valid. As a major drawback, the information concerning the multiplicity $M(\hat{\Theta}_{X_1, Y})$ cannot be retrieved via this route and is completely lost. *Local IDA* can therefore be useful for estimating lower and upper bounds of the causal effect of interest, but is not suitable in the study of quantities that require knowledge of its statistics, such as its mean and variance.

5 Causality with time

With this section we consider a change of perspective, delving into the link between causality and predictability depicted in Figure 2. When considering a system in evolution, it is natural to expect that identifying causal links can improve the ability to exhibit robust predictions. Moreover, causal reasoning involving variables across different time instances is comparatively more straightforward than causal reasoning lacking a temporal structure whenever it is assumed that the cause must precede the effect in time [129]. Yet, the problem of giving a formal definition for causality that

is sufficiently comprehensive and suitable to be tested empirically remains open. In what follows, we propose a preliminary discussion of the Dynamical System context by considering three distinct approaches: *Granger Causality*, *Transfer Entropy*, and *Fluctuation-Dissipation* relations. Equivalent characterizations for the existence of causal relations in the linear case are obtained formally, leaving the epistemological frame of reference in the background. A detailed analysis of the concept of information flow will be presented. Although informational quantities are usually understood at the observational level, we will show how causal semantics can also be employed in the case of Transfer Entropy, at least by considering the linear case.

Granger Causality

From the point of view of time series analysis, the first significant contribution is due to C. Granger [60]. Assuming that a cause must temporally precede the corresponding effect and assuming that the former contains information about the latter, it is stated that the process Y causes the process X if the future values of X can be better predicted provided the past values of X and Y are known, than if only the past values of X are known. This idea can be formalized for a linear regression model,

$$X_t = a_0 + \sum_{k=1}^{l_X} a_k^X X_{t-k} + \sum_{k=1}^{l_Y} a_k^Y Y_{t-k} + \mu_t, \quad (61)$$

where the hypothesis in which Y does not cause X corresponds to the requirement of $a_k^Y = 0$ for $k = 1, 2, \dots, l_Y$, giving us the equation

$$X_t = a_0 + \sum_{k=1}^{l_X} a_k^X X_{t-k} + \nu_t. \quad (62)$$

Having made a choice for l_X and l_Y , the parameters a_0 , a_k^X , and a_k^Y are selected to minimize the error. Given the covariance matrices for the noises μ_t and ν_t respectively Σ_μ and Σ_ν , we can quantify causality in the Granger sense via

$$G_{Y \rightarrow X} := \log \frac{|\Sigma_\nu|}{|\Sigma_\mu|}, \quad (63)$$

where $|\cdot|$ denotes the determinant for the corresponding matrix. We interpret the eventuality that $G_{Y \rightarrow X} \neq 0$ as evidence that knowledge of Y improves the predictive ability about X , delivering us a reduction in error, being $|\Sigma_\mu| < |\Sigma_\nu|$. We will then say that in the hypothesis that $G_{Y \rightarrow X} \neq 0$ exists a causal link $Y \rightarrow X$, quantified precisely by the value assumed by $G_{Y \rightarrow X}$. Other theoretical proposals set in the context of information theory followed. Among these, *Transfer Entropy* offers an equivalent characterization for the case of Gaussian processes [14]. Reconsidering the discussion carried out for the static case, Granger Causality can be characterized by the following separation relation:

$$Y_t \not\perp\!\!\!\perp X_{\text{past}(t)} | Y_{\text{past}(t)}.$$

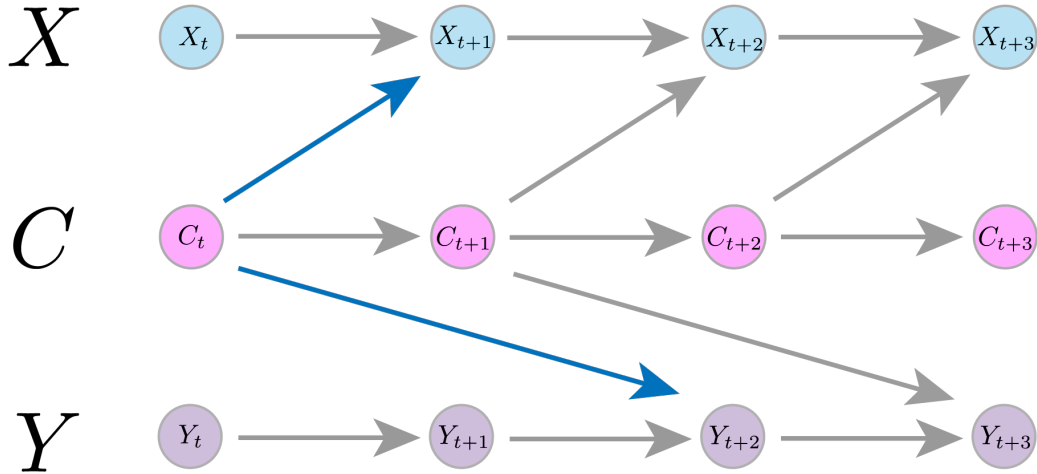


Figure 21: *Violation of Causal Sufficiency.* Due to the hidden common cause C , Granger Causality erroneously infers a causal link from X to Y .

Remarkably, Figure 21 shows how any purely informational index can infer incorrect causation in the case where a confounder is present. In what follows, we will present a more detailed analysis of the concept of *information flow* by readapting the Transfer Entropy index via causal semantics. It will be done by considering the simplest possible case - the linear case - establishing a link with the Granger index and linear response.

Causality in models: an elementary case

From a different point of view, the problem arises of defining causality within the framework offered by the dynamic systems theory, assuming an explicit model is available. To illustrate this point, we begin by considering an elementary case. The *paradox of chocolate* [104] effectively illustrates the difference between correlation and causation. The available data gives us the following situation: there is a strong correlation between the consumption of chocolate in a state, quantified by the variable y , and the number z of Nobel laureates of the corresponding nationality. To explain this fact, one can resort to a third variable x , capable of quantifying the well-being of a community and thus justifying both the consumption of chocolate and the level of education of its citizens. In other words, we are in a situation of the type $y \leftarrow x \rightarrow z$, where the correlation between y and z is a consequence of the causal links $x \rightarrow y$ and $x \rightarrow z$, without a causal relation between y and z . We can formalize this circumstance via a minimal model,

$$\begin{cases} x_t = a_x x_{t-1} + \sigma_1 w_x \\ y_t = b_x x_{t-1} + b_y y_{t-1} + \sigma_2 w_y \\ z_t = c_x x_{t-1} + c_z z_{t-1} + \sigma_3 w_z \end{cases}, \quad (64)$$

where w_x, w_y, w_z are normalised and mutually independent Gaussian variables, x_t, y_t, z_t are Gaussian processes with zero mean, a_x, b_x, c_x, b_y, c_z are appropriate positive constants. Considering the invariant distribution for the process, obtained for $t \rightarrow \infty$, we obtain the correlation $\langle yz \rangle \neq 0$. It is natural to expect that the presence of a causal link for the system (64) is interpreted by the coefficients b_x and c_z , respectively, for the causal relations $x \rightarrow y$ e $x \rightarrow z$. Similarly, modifying the system (64) via a feedback of y on x , i.e. writing

$$\begin{cases} x_t = a_x x_{t-1} + \epsilon y_t + \sigma_1 w_x \\ y_t = b_x x_{t-1} + b_y y_{t-1} + \sigma_2 w_y \\ z_t = c_x x_{t-1} + c_z z_{t-1} + \sigma_3 w_z \end{cases}, \quad (65)$$

we expect the positive parameter ϵ to interpret the relation $y \rightarrow x$, reflecting on the evolution of z and producing a causal link of the type $y \rightarrow x$. From here on, we have in mind the toy models (64) and (65).

Transfer

In order to define a reasonable notion of causality, time dependence must be introduced for the variables involved, assuming that causes must temporally precede effects. Let two discrete-time stochastic processes be given, $X := X_t$ and $Y := Y_t$. Under the assumption that X and Y are of Markovian type, we admit that an occurrence of X_t is somehow determined by the history of X , i.e. by the occurrence of X_{t-1} , as well as by the action of Y on X , e.g. via the past state Y_{t-1} alone. These assumptions can be generalized, considering a history of length k for X , say $(X_{t-1}, X_{t-2}, \dots, X_{t-k})$, as well as an action at l times exerted by Y , via $(Y_{t-1}, Y_{t-2}, \dots, Y_{t-l})$. We start with the simplest case, with $k = l = 1$. In order to capture a causal link of the type $Y \rightarrow X$, we employ conditioning to remove the information due to the history of X , isolating the contribution of the action of Y_{t-1} on X_t . Following T. Schreiber [149], we recall the definition of *transfer entropy* $T_{Y \rightarrow X}$ at time t :

Definition 1 *Let $X = \{X_t\}$ and $Y = \{Y_t\}$ be markovian stochastic processes. We define the transfer entropy from Y to X at time t*

$$T_{Y \rightarrow X}(t-1, t) := I(X_t : Y_{t-1} | X_{t-1}). \quad (66)$$

The Definition 1 can be made more transparent by writing

$$H[X_t | X_{t-1}] = H[X_t | X_{t-1}, Y_{t-1}] + T_{Y \rightarrow X}(t-1, t).$$

In other words, the information associated with the occurrence of X_t , given the history X_{t-1} , is equal to the sum of the contribution of the process X alone, $H[X_t | X_{t-1}, Y_{t-1}]$, with the contribution transferred from Y to X . We note that the conditioning introduces an asymmetry, absent for the mutual information $I(X_t : Y_{t-1})$ alone, allowing us to distinguish $T_{Y \rightarrow X}$ from $T_{X \rightarrow Y}$. Assuming that $T_{Y \rightarrow X} \neq 0$, we will say that there is a causal link $Y \rightarrow X$, quantified precisely by the value assumed by $T_{Y \rightarrow X}$.

Remark 2 We can highlight the role played by the probability, joint and conditional distributions by making the definition more explicit. We obtain

$$\begin{aligned} T_{Y \rightarrow X}(t-1, t) &= \sum_{x_{t-1} \in R_{X_{t-1}}} \sum_{x_t \in R_{X_t}, y_{t-1} \in R_{Y_{t-1}}} p(x_{t-1}) p(x_t, y_{t-1} | x_{t-1}) \cdot \\ &\quad \cdot \log \frac{p(x_t, y_{t-1} | x_{t-1})}{p(x_t | x_{t-1}) p(y_{t-1} | x_{t-1})} \\ &= \sum_{x_{t-1} \in R_{X_{t-1}}} \sum_{x_t \in R_{X_t}, y_{t-1} \in R_{Y_{t-1}}} p(x_t, y_{t-1}, x_{t-1}) \log \frac{p(x_t | x_{t-1}, y_{t-1})}{p(x_t | x_{t-1})}. \end{aligned}$$

Clearly, if $p(x_t, y_{t-1} | x_{t-1}) = p(x_t | x_{t-1}) p(y_{t-1} | x_{t-1})$, i.e., if X_t e Y_{t-1} are conditionally independent w.r.t. X_{t-1} , then $T_{Y \rightarrow X}(t-1, t) = 0$.

Remark 3 In general, we are interested in the case of systems with dimensionality $d \geq 2$, i.e., the interdependence between different degrees of freedom, each corresponding to a stochastic process X^k , with $k = 1, 2, \dots, d$. To characterize the information transfer for $d > 2$, we start by giving a more transparent interpretation for $d = 2$, via a simple manipulation of the definition of $T_{X^2 \rightarrow X^1}$. Adding $H[X_t^1]$ into both members of (66), it is straightforward to derive the equality

$$H[X_t^1] = A_{X^1}(t) + T_{X^2 \rightarrow X^1}(t-1, t) + H[X_t^1 | X_{t-1}^1, X_{t-1}^2], \quad (67)$$

where $A_{X^1}(t) := H[X_t^1] - H[X_t^1 | X_{t-1}^1]$. Put differently, the information associated with X_t^1 is equal to the sum of three components: the self-information A_{X^1} contained in the process $X^{(1)}$, the information transferred from X^2 to X^1 , $T_{X^2 \rightarrow X^1}$, a higher-order term taking into account the conditioning on both processes, $H[X_t^1 | X_{t-1}^1, X_{t-1}^2]$. If we consider a third process X^3 , from (67) we immediately obtain

$$\begin{aligned} H[X_t^1] &= A_{X^1}(t) + T_{X^2 \rightarrow X^1}(t-1, t) + T_{X^3 \rightarrow X^1 | X^2}(t-1, t) + \\ &\quad + H[X_t^1 | X_{t-1}^1, X_{t-1}^2, X_{t-1}^3], \end{aligned}$$

where

$$T_{X^3 \rightarrow X^1 | X^2}(t-1, t) := H[X_t^1 | X_{t-1}^1, X_{t-1}^2] - H[X_t^1 | X_{t-1}^1, X_{t-1}^2, X_{t-1}^3],$$

i.e. the information transferred by X^3 minus the contribution already allocated by X^2 . Retracing the steps just performed, taking care to reverse the roles of X^2 and X^3 , it is immediate to verify that

$$T_{X^2 \rightarrow X^1} + T_{X^3 \rightarrow X^1 | X^2} = T_{X^3 \rightarrow X^1} + T_{X^2 \rightarrow X^1 | X^3},$$

so that the definition of $T_{X^3, X^2 \rightarrow X^1}$ is well placed. The reasoning can be reiterated to include other variables, defining at each step k the transfer $T_{X^k, \dots, X^2 \rightarrow X^1}$ and gradually reducing the size of the residual information term $H[X_t^1 | X_{t-1}^1, X_{t-1}^2, \dots, X_{t-1}^k]$, so that the result is

$$\begin{cases} H[X_t^1] = A_{X^1}(t) + T_{X^2, X^3, \dots, X^k \rightarrow X^1}(t-1, t) + H[X_t^1 | X_{t-1}^1, X_{t-1}^2, \dots, X_{t-1}^k] \\ T_{X^2, X^3, \dots, X^k \rightarrow X^1} := T_{X^2 \rightarrow X^1} + T_{X^3 \rightarrow X^1 | X^2} + \dots + T_{X^k \rightarrow X^1 | X^1, \dots, X^{k-1}} \end{cases}$$

Remark 4 For the case $k, l \neq 0$, we introduce the notation $X_{t-1}^{(k)} := (X_{t-1}, X_{t-2}, \dots, X_{t-k})$, $Y_{t-1}^{(l)} := (Y_{t-1}, Y_{t-2}, \dots, Y_{t-l})$, so that we can give a definition for the transfer entropy of order (k, l) ,

$$T_{Y \rightarrow X}^{(k,l)}(t-1, t) := I(X_t : Y_{t-1}^{(l)} | X_{t-1}^{(k)}),$$

which can be written explicitly, in full analogy with what has already been done in Observation 2.1. In particular:

$$\begin{aligned} T_{Y \rightarrow X}^{(k,l)}(t-1, t) &= \sum_{x_t, x_{t-1}^{(k)}, y_{t-1}^{(l)}} p(x_{t-1}^{(k)}) p(x_t, y_{t-1}^{(l)} | x_{t-1}^{(k)}) \\ &\quad \cdot \log \frac{p(x_t, y_{t-1}^{(l)} | x_{t-1}^{(k)})}{p(x_t | x_{t-1}^{(k)}) p(y_{t-1}^{(l)} | x_{t-1}^{(k)})} \\ &= \sum_{x_t, x_{t-1}^{(k)}, y_{t-1}^{(l)}} p(x_t, y_{t-1}^{(l)}, x_{t-1}^{(k)}) \log \frac{p(x_t | x_{t-1}^{(k)}, y_{t-1}^{(l)})}{p(x_t | x_{t-1}^{(k)})}. \end{aligned}$$

In concrete applications, we have the time series x_t^k and y_t^l , to be thought of as realizations for appropriate stochastic processes. At least in principle, making appropriate assumptions about the memory of the processes involved and the time of influence, it is possible to estimate from the available data the probabilities involved in the definition of transfer, allowing us to offer a quantification for the causal links.

Dynamical Systems

We can think of x_t^k as a set of values assumed by the observables of an underlying dynamical system, which we assume to be known. In this scenario, we propose a definition of transfer inspired by the one discussed above, explicitly interpreting conditioning. We consider the case of a discrete-time dynamical system via an equation of the type

$$z_t = F(z_{t-1}) + \sigma w, \quad (68)$$

where $z_t = (z_{1t}, z_{2t}, \dots, z_{nt}) \in \mathbb{R}^n$, $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$, $\sigma = (\sigma_1, \sigma_2, \dots, \sigma_n) \in \mathbb{R}_+^n$, $w = (w_1, w_2, \dots, w_n)$ a vector of independent gaussian variables $N(0, 1)$, so that $\sigma w := (\sigma_1 w_1, \sigma_2 w_2, \dots, \sigma_n w_n)$, interpreting the presence of noise for the system. With the aim of identifying causal links between the different $\{z_i\}_{i=1, \dots, n}$, equation (68) can be rewritten by decomposing z into the tuple $(x, y) \in \mathbb{R}^p \times \mathbb{R}^q$, with $p + q = n$. We obtain

$$\begin{cases} x_t = F_p(x_{t-1}, y_{t-1}) + \sigma_p w_p, \\ y_t = F_q(x_{t-1}, y_{t-1}) + \sigma_q w_q, \end{cases} \quad (69)$$

where $F_p : \mathbb{R}^n \rightarrow \mathbb{R}^p$ and $F_q : \mathbb{R}^n \rightarrow \mathbb{R}^q$ are respectively the projections of F on the subspaces of $x = (z_1, z_2, \dots, z_p)$ and of $y = (z_{p+1}, z_{p+2}, \dots, z_n)$, while Given $\sigma_p = (\sigma_1, \sigma_2, \dots, \sigma_p)$, $\sigma_q = (\sigma_{p+1}, \sigma_{p+2}, \dots, \sigma_n)$, $w_p = (w_1, w_2, \dots, w_p)$, $w_q = (w_{p+1}, w_{p+2}, \dots, w_n)$, the terms $\sigma_p w_p$ and $\sigma_q w_q$ are defined as above. The chocolate paradox models fall

into this class. Assigned the initial data z_0 via a certain probability distribution ρ_{z_0} , this will evolve due to the dynamics defined by (69), delivering us the distributions ρ_{z_t} as t varies, as well as the marginal distributions ρ_{x_t} and ρ_{y_t} . Therefore, it is possible to calculate the corresponding entropies, henceforth denoted $H[\rho]$, where ρ is the probability distribution considered.

The basic idea is simple: to characterize a causal link of the type $y \rightarrow x$, we introduce a constraint in the evolution of y , and then quantify the resulting information gap. Let us go into some details. Instead of the system (69), we consider the *constrained system*

$$\begin{cases} x_t = F_p(x_{t-1}, \hat{y}) + \sigma_p w_p, \\ y_t = y_{t-1} = \hat{y}, \end{cases} \quad (70)$$

where the evolution of y_t is fixed at the constant value \hat{y} . Given the same distribution of initial data ρ_{z_0} already introduced for system (69), the dynamics of system (70) will give us a distribution $\hat{\rho}_{z_t}$ as t varies, as well as the marginal $\hat{\rho}_{x_t}$ and $\hat{\rho}_{y_t}$, so that we can compute the corresponding entropies $H[\hat{\rho}]$.

The dynamical system (68) gives us a Markovian type process, whereby the state at time t is completely determined by that at step $t - 1$. By analogy with the case of time series, we can quantify the information transfer $y \rightarrow x$ by conditioning x w.r.t. its history to isolate the information contribution of y alone. It is therefore natural to consider the conditional distributions $\rho_{x_t|x_{t-1}}$ and $\hat{\rho}_{x_t|x_{t-1}}$, establishing the following correspondence:

$$\begin{cases} H[X_t] \longleftrightarrow H[\rho_{x_t}] \\ H[X_t|X_{t-1}] \longleftrightarrow H[\rho_{x_t|x_{t-1}}] \\ H[X_t|X_{t-1}, Y_{t-1}] \longleftrightarrow H[\hat{\rho}_{x_t|x_{t-1}}] \end{cases} .$$

Having said all this, we can reformulate the definition of transfer, adapting it to the case of a dynamic system.

Definition 2 *Given a dynamic system of type (69), the information transfer $T_{y \rightarrow x}$ at time t is defined as*

$$T_{y \rightarrow x}(t - 1, t) := H[\rho_{x_t|x_{t-1}}] - H[\hat{\rho}_{x_t|x_{t-1}}],$$

where $\rho_{x_t|x_{t-1}}$ and $\hat{\rho}_{x_t|x_{t-1}}$ are respectively the conditional distributions for x_t , in the case where y_t is free or constrained.

As before, the Definition 2 can be made more transparent by writing $H[\rho_{x_t|x_{t-1}}] = H[\hat{\rho}_{x_t|x_{t-1}}] + T_{y \rightarrow x}(t - 1, t)$. The information associated with the distribution $\rho_{x_t|x_{t-1}}$, is equal to the sum of two contributions: that due to x alone, isolated by constraining y , with that transferred from y to x . Again in analogy with what we discussed for the time series case, some simple formal properties for the transfer hold:

Lemma 2 *Given a dynamical system of type (69) and defined the transfer entropy with $T_{y \rightarrow x}(t - 1, t)$ as in Definition 2, then:*

- i. If $\rho_{x_t|x_{t-1}} = \hat{\rho}_{x_t|x_{t-1}}$, i.e. if the dynamics of y do not condition the dynamics of x , then $T_{y \rightarrow x}(t-1, t) = 0$.
- ii. Let $A[\rho_{x_t}] := H[\rho_{x_t}] - H[\rho_{x_t|x_{t-1}}]$ be the information associated with the evolution of ρ_{x_t} alone, it results that

$$H[\rho_{x_t}] = A[\rho_{x_t}] + T_{y \rightarrow x}(t-1, t) + H[\hat{\rho}_{x_t|x_{t-1}}],$$

so that the information associated with the distribution ρ_{x_t} is equal to the sum of three contributions: the self-information $A[\rho_{x_t}]$, the component in the absence of y and the information transferred from y to x .

- iii. In general, it holds that

$$T_{y_1, y_2, \dots, y_q \rightarrow x} = T_{y_1 \rightarrow x} + T_{y_2 \rightarrow x|y_1} + \dots + T_{y_q \rightarrow x|y_1, \dots, y_{q-1}},$$

where conditioning with respect to y_i is interpreted by fixing its value.

Linear Systems

To test Definition 2, let us consider the linear case in detail. In place of (68), we consider the system

$$z_t = Az_{t-1} + \sigma w, \quad (71)$$

$$z_t = (x_t, y_t) \in \mathbb{R}^l \times \mathbb{R}^n,$$

$$A = \begin{pmatrix} A_p & A_{pq} \\ A_{qp} & A_q \end{pmatrix} \in M_n(\mathbb{R}),$$

$$\begin{cases} x_t = A_p x_{t-1} + A_{pq} y_{t-1} + \sigma_p w_p \\ y_t = A_{qp} x_{t-1} + A_q y_{t-1} + \sigma_q w_q \end{cases}, \quad (72)$$

where $A_p \in M_p(\mathbb{R})$, $A_q \in M_q(\mathbb{R})$, $A_{pq} \in M_{p,q}(\mathbb{R})$, $A_{qp} \in M_{q,p}(\mathbb{R})$. The initial datum z_0 is assigned via a Gaussian distribution ρ_{z_0} , with covariance matrix Σ_0 and null mean,

$$\rho_{z_0}(z) = \frac{1}{\pi^{n/2} |\Sigma_0|^{1/2}} \exp \left\{ -\frac{1}{2} z \cdot \Sigma_0^{-1} z \right\}.$$

With the aim of determining the entropies $H[\rho_{x_t|x_{t-1}}]$ and $H[\hat{\rho}_{x_t|x_{t-1}}]$, we recall some elementary facts for Gaussian processes, collected in the following lemmas.

Lemma 3 Under the action of the linear system (71), the Gaussian $\rho_{z_{t-1}}$ with covariance matrix Σ_{t-1} gives us a Gaussian ρ_{z_t} , with covariance matrix

$$\Sigma_t = A \Sigma_{t-1} A^T + \sigma^2 I.$$

Lemma 4 Given a Gaussian distribution ρ with a covariance matrix Σ , the entropy $H[\rho]$ depends on the Σ alone, via the relation

$$H[\rho] = \frac{1}{2} \ln \{ (2\pi e)^n |\Sigma| \}.$$

Lemma 5 *If $z = (x, y) \in \mathbb{R}^p \times \mathbb{R}^q$, with $p+q = n$, is a gaussian vector with covariance matrix Σ ,*

$$\Sigma = \begin{pmatrix} \Sigma_p & \Sigma_{pq} \\ \Sigma_{qp} & \Sigma_q \end{pmatrix} \in M_n(\mathbb{R}), \quad (73)$$

where $\Sigma_p = \langle x_i x_j \rangle \in M_p(\mathbb{R})$, $\Sigma_q = \langle y_i y_j \rangle \in M_q(\mathbb{R})$, $\Sigma_{qp} = \Sigma_{pq}^T = \langle x_i y_j \rangle \in M_{p,q}(\mathbb{R})$. Then x is a Gaussian vector with covariance matrix Σ_p and y is a Gaussian vector with covariance matrix Σ_q .

Lemma 6 *If $z = (x, y) \in \mathbb{R}^p \times \mathbb{R}^q$, with $p + q = n$, is a gaussian vector with covariance matrix Σ decomposed as in 73, then conditioning $\rho_{x|y}$ is a Gaussian vector with covariance matrix $\Sigma_{p|q}$ given by the relation*

$$\Sigma_{p|q} = \Sigma_p - \Sigma_{pq} \Sigma_q^{-1} \Sigma_{qp} \quad (74)$$

At this point, we have all the elements to calculate the covariance matrices associated with the distributions $\rho_{x_t|x_{t-1}}$ and $\hat{\rho}_{x_t|x_{t-1}}$. We consider the Gaussian vector $(x_t, x_{t-1}) \in \mathbb{R}^p \times \mathbb{R}^p$, with covariance matrix

$$\Sigma := \begin{pmatrix} \Sigma_{p'} & \Sigma_{p'p} \\ \Sigma_{pp'} & \Sigma_p \end{pmatrix},$$

$$\Sigma_p := \langle x_{i,t-1} x_{j,t-1} \rangle,$$

$$\Sigma_{pp'} = \Sigma_{p'p}^T := \langle x_{i,t-1} x_{j,t} \rangle,$$

$$\Sigma_{p'} := \langle x_{i,t} x_{j,t} \rangle.$$

In general, we adopt a simple convention: primed indices correspond to objects defined at time t , and non-primed indices to objects defined at time $t - 1$. For the sake of simplicity, without compromising the generality of the argument, let us set all parameters σ_i to σ . Applying the Lemma 6, it turns out that $\rho_{x_t|x_{t-1}}$ has a covariance matrix

$$\Sigma_{p'|p} = \Sigma_{p'} - \Sigma_{p'p} \Sigma_p^{-1} \Sigma_{pp'}, \quad (75)$$

for which it is necessary to calculate the matrices $\Sigma_{p'}$ and $\Sigma_{pp'}$, remembering that $x_t = A_p x_{t-1} + A_{pq} y_{t-1} + \sigma_p w_p$. In matrix notation, this results in:⁵

$$\begin{cases} \Sigma_{p'} = A_p \Sigma_p A_p^T + A_p \Sigma_{pq} A_{pq}^T + A_{pq} \Sigma_{qp} A_{pq}^T + A_{pq} \Sigma_q A_{pq}^T + \sigma^2 I_p \\ \Sigma_{pp'} = A_{pq} \Sigma_{qp} + A_p \Sigma_p \end{cases} \quad (76)$$

so by replacing 76 in 75, we obtain

$$\Sigma_{p'|p} = A_{pq} (\Sigma_q - \Sigma_{qp} \Sigma_p^{-1} \Sigma_{pq}) A_{pq}^T + \sigma^2 I_p.$$

From Lemma 4, we obtain $H[\rho_{x_t|x_{t-1}}]$. Similarly, if we repeat the account just performed, this time for the constrained system,

$$\begin{cases} x_t = A_p x_{t-1} + A_{pq} \hat{y} + \sigma_p w_p \\ y_t = y_{t-1} = \hat{y} \end{cases},$$

we derive that the associated covariance matrix $\hat{\Sigma}_{p'|p}$. It is immediate to observe that x_{t-1} and \hat{y} , do not affect the writing for the covariance matrix, so that

$$\hat{\Sigma}_{p'|p} = \sigma^2 I_p$$

From Lemma 4, we obtain $H[\hat{\rho}_{x_t|x_{t-1}}]$. We then proved the following result.

Theorem 7 *Given a linear system of the type (72), the transfer entropy $T_{y \rightarrow x}$ is of the form*

$$T_{y \rightarrow x}(t-1, t) = \frac{1}{2} \ln |A_{pq} (\Sigma_q - \Sigma_{qp} \Sigma_p^{-1} \Sigma_{pq}) A_{pq}^T + \sigma^2 I_p| - \frac{1}{2} \ln |\sigma^2 I_p|.$$

Corollary 1 *Under the same assumptions as the 7, it turns out that $T_{y \rightarrow x}(t-1, t) = 0$ if $A_{pq} = 0$, i.e. if in the equations for x the coefficients for y are all null.*

⁵It is sufficient to explicitly write down the quantities in question:

$$\begin{aligned} (\Sigma_{p'})_{ij} &:= \langle x'_i x'_j \rangle = \\ &= \langle \{ (A_p)_i^k x_k + (A_{pq})_i^l y_l + \sigma_{pi} w_{pi} \} \cdot \{ (A_p)_j^s x_s + (A_{pq})_j^t y_t + \sigma_{pj} w_{pj} \} \rangle = \\ &\quad (A_p)_i^k \langle x_k x_s \rangle (A_p)_j^s + (A_p)_i^k \langle x_k y_t \rangle (A_{pq})_j^t + \\ &\quad + (A_{pq})_i^l \langle y_l x_s \rangle (A_p)_j^s + (A_{pq})_i^l \langle y_l y_t \rangle (A_{pq})_j^t + \sigma^2 \delta_{ij}, \end{aligned}$$

and similarly

$$\begin{aligned} (\Sigma_{p'p})_{ij} &:= \langle x'_i x_j \rangle = \\ &= \langle \{ (A_p)_i^k x_k + (A_{pq})_i^l y_l + \sigma^i w_i \} x_j \rangle = \\ &= (A_p)_i^k \langle x_k x_j \rangle + (A_{pq})_i^l \langle x_l x_i \rangle. \end{aligned}$$

Remark 5 We observe that from the proposed definitions and the Lemma 4, the result due to L. Bennett et al., concerning the Granger-causality-transfer equivalence for the case of Gaussian processes, immediately follows. Reconsidering the notation proposed for the linear regression model, it is easy to realise that, for example in the case of a single time step, $\Sigma_{x_t|x_{t-1}}$ and $\Sigma_{mu} = \Sigma_{x_t|x_{t-1}, y_{t-1}}$, so that:

$$T_{y \rightarrow x} = \frac{1}{2} G_{y \rightarrow x}.$$

Remark 6 We can specify the reasoning made for the case where we want to quantify the transfer from a subset of the variables y , say (y_1, \dots, y_c) , to a subset of x , say (x_1, \dots, x_a) . With a little abuse of notation, we decompose the vector z as (x_a, x_b, y_c, y_d) , where $a + b = p$ and $c + d = q$, denoting by x_a, x_b, y_c, y_d also the corresponding vectors. We rewrite the linear system:

$$\begin{cases} x_{a,t} = A_a x_{a,t-1} + A_{ab} x_{b,t-1} + A_{ac} y_{c,t-1} + A_{ad} y_{d,t-1} + \sigma_a w_a \\ x_{b,t} = A_{ba} x_{a,t-1} + A_b x_{b,t-1} + A_{bc} y_{c,t-1} + A_{bd} y_{d,t-1} + \sigma_b w_b \\ y_{c,t} = A_{ca} x_{a,t-1} + A_{cb} x_{b,t-1} + A_c y_{c,t-1} + A_{cd} y_{d,t-1} + \sigma_c w_c \\ y_{d,t} = A_{da} x_{a,t-1} + A_{db} x_{b,t-1} + A_{dc} y_{c,t-1} + A_d y_{d,t-1} + \sigma_d w_d \end{cases}, \quad (77)$$

where

$$A = \begin{pmatrix} A_a & A_{ab} & A_{ac} & A_{ad} \\ A_{ba} & A_b & A_{bc} & A_{bd} \\ A_{ca} & A_{cb} & A_c & A_{cd} \\ A_{da} & A_{db} & A_{dc} & A_d \end{pmatrix},$$

Adopting the same conventions as before, let us consider the gaussian vector $(x_{a,t}, x_{a,t-1}) \in \mathbb{R}^a$, so that $\rho_{x_{a,t}|x_{a,t-1}}$ is a gaussian with covariance matrix

$$\Sigma_{a'|a} = \Sigma_a - \Sigma_{a'a} \Sigma_a^{-1} \Sigma_{aa'},$$

to be calculated explicitly using the relations in 77. It results:

$$\begin{aligned} \Sigma_{a'} &= A_a \Sigma_a A_a^T + A_a \Sigma_{ab} A_{ab}^T + A_a \Sigma_{ac} A_{ac}^T + A_a \Sigma_{ad} A_{ad}^T + \\ &+ A_{ab} \Sigma_{ba} A_{ab}^T + A_{ab} \Sigma_b A_{ab}^T + A_{ab} \Sigma_{bc} A_{bc}^T + A_{ab} \Sigma_{bd} A_{bd}^T + \\ &+ A_{ac} \Sigma_{ca} A_{ac}^T + A_{ac} \Sigma_{cb} A_{cb}^T + A_{ac} \Sigma_c A_{ac}^T + A_{ac} \Sigma_{cd} A_{cd}^T + \\ &+ A_{ad} \Sigma_{da} A_{ad}^T + A_{ad} \Sigma_{db} A_{db}^T + A_{ad} \Sigma_{dc} A_{dc}^T + A_{ad} \Sigma_d A_{ad}^T + \\ &+ \sigma^2 I_a, \end{aligned}$$

$$\Sigma_{a'a} = A_a \Sigma_a + A_{ab} \Sigma_{ba} + A_{ac} \Sigma_{ca} + A_{ad} \Sigma_{da},$$

so that we obtain

$$\begin{aligned} \Sigma_{a'|a} &= A_{ab} (\Sigma_b - \Sigma_{ba} \Sigma_a^{-1} \Sigma_{ab}) A_{ab}^T + \\ &+ A_{ac} (\Sigma_c - \Sigma_{ca} \Sigma_a^{-1} \Sigma_{ac}) A_{ac}^T + \\ &+ A_{ad} (\Sigma_d - \Sigma_{da} \Sigma_a^{-1} \Sigma_{ad}) A_{ad}^T + \end{aligned}$$

$$\begin{aligned}
& +A_{ab} (\Sigma_{bc} - \Sigma_{ba}\Sigma_a^{-1}\Sigma_{ac}) A_{ac}^T + \\
& +A_{ac} (\Sigma_c - \Sigma_{ca}\Sigma_a^{-1}\Sigma_{ac}) A_{ac}^T + \\
& +A_{ad} (\Sigma_{dc} - \Sigma_{da}\Sigma_a^{-1}\Sigma_{ac}) A_{ac}^T + \\
& +A_{ab} (\Sigma_{bd} - \Sigma_{ba}\Sigma_a^{-1}\Sigma_{ad}) A_{ad}^T + \\
& +A_{ac} (\Sigma_{cd} - \Sigma_{ca}\Sigma_a^{-1}\Sigma_{ad}) A_{ad}^T + \\
& +A_{ad} (\Sigma_d - \Sigma_{da}\Sigma_a^{-1}\Sigma_{ad}) A_{ad}^T + \\
& +\sigma^2 I_a.
\end{aligned}$$

Given $q := (b, c, d)$, we can rewrite the conditional covariance matrix in compact form:

$$\Sigma_{a'|a} = A_{aq} (\Sigma_q - \Sigma_{qa}\Sigma_a^{-1}\Sigma_{aq}) A_{aq}^T + \sigma^2 I_a.$$

If, instead of the 77, we consider the corresponding constrained system, fixing $y_c = \text{haty}$, we obtain

$$\hat{\Sigma}_{a'|a} = A_{a\hat{q}} (\Sigma_{\hat{q}} - \Sigma_{\hat{q}a}\Sigma_a^{-1}\Sigma_{a\hat{q}}) A_{a\hat{q}}^T + \sigma^2 I_a,$$

where $\hat{q} := (b, d)$. We have thus proved the following theorem and the consequent corollaries.

Theorem 8 Given a linear system of type 77, the transfer entropy $T_{y_c \rightarrow x_a}$ is of the form

$$\begin{aligned}
T_{y_c \rightarrow x_a}(t-1, t) &= \frac{1}{2} \ln |A_{aq} (\Sigma_q - \Sigma_{qa}\Sigma_a^{-1}\Sigma_{aq}) A_{aq}^T + \sigma^2 I_a| + \\
&\quad - \frac{1}{2} \ln |A_{a\hat{q}} (\Sigma_{\hat{q}} - \Sigma_{\hat{q}a}\Sigma_a^{-1}\Sigma_{a\hat{q}}) A_{a\hat{q}}^T + \sigma^2 I_a|,
\end{aligned}$$

where we defined $q := (b, c, d)$ e $\hat{q} = (b, d)$.

Corollary 2 Under the same assumptions as in Theorem 8, it turns out that $T_{y_c \rightarrow x_a}(t-1, t) = 0$ if $A_{ac} = 0$, i.e. if in the equations for x_a the coefficients in front of y_c are all null.

Corollary 3 With the same hypothesis of 8, fixed $a = c = 1$, $x_a = z_j$ and $y_c = x_i$, it follows that $T_{z_i \rightarrow z_j}(t-1, t) = 0$ if and only if $A_{ji} = 0$.

Remark 7 If we reconsider the minimal model for the chocolate paradox 64, the Corollary 3 assures us that $T_{x \rightarrow y}(t-1, t) \neq 0$, $T_{x \rightarrow z}(t-1, t) \neq 0$, $T_{y \rightarrow z}(t-1, t) = T_{z \rightarrow y}(t-1, t) = 0$, in agreement with our expectations. The definition of transfer entropy, at least when applied to the particular case of linear systems with a gaussian distribution, captures the existence of causal connections. To quantify the connections $x \rightarrow y$ and $x \rightarrow z$, we compute $T_{x \rightarrow y}(0, 1)$ and $T_{x \rightarrow z}(0, 1)$ under the assumption that

the gaussian distribution for the initial datum has covariance matrix $\Sigma_0 = \sigma_0^2 I$. It holds that⁶:

$$\begin{aligned} A_{2q} &= (b_x, 0), & q &:= (1, 3), & A_{2\hat{q}} &= 0, & \hat{q} &:= (3) \\ \Sigma_{q2} &= (0, 0)^T, & \Sigma_q^* &= \Sigma_q - \Sigma_{q2}\Sigma_2^{-1}\Sigma_{2q} = \sigma_0^2 I, \\ A_{3q} &= (c_x, 0), & q &:= (1, 2), & A_{3\hat{q}} &= 0, & \hat{q} &:= (1) \\ \Sigma_{q3} &= (0, 0)^T, & \Sigma_q^* &= \Sigma_q - \Sigma_{q3}\Sigma_2^{-1}\Sigma_{3q} = \sigma_0^2 I, \end{aligned}$$

so that it is immediate to obtain

$$\begin{aligned} T_{x \rightarrow y}(0, 1) &= \frac{1}{2} \ln \left(1 + b_x^2 \frac{\sigma_0^2}{\sigma^2} \right), \\ T_{x \rightarrow z}(0, 1) &= \frac{1}{2} \ln \left(1 + c_x^2 \frac{\sigma_0^2}{\sigma^2} \right). \end{aligned}$$

We observe that the transfer does not depend on the sign of the coefficient b_x and that it is negligible if $\frac{\sigma_0}{\sigma} < 1$, i.e. if the noise is significantly larger than the uncertainty on the initial datum.

Remark 8 If we consider the modified model 65, the condition $T_{y \rightarrow z}(t-1, t) = 0$ tells us that the transfer is null, despite the feedback $y \rightarrow x$, with $\epsilon \neq 0$. We expect a non-zero transfer $y \rightarrow z$ provided we consider two time steps. It is therefore natural to propose the following generalisation of Definition 2, for k time steps:

Definition 3 Given a dynamic system of type 69, the information transfer $T_{y \rightarrow x}$, at time t and after k instants, is defined as

$$T_{y \rightarrow x}(t-k, t) := H[\rho_{x_t|x_{t-k}}] - H[\hat{\rho}_{x_t|x_{t-k}}],$$

where $\rho_{x_t|x_{t-k}}$ and $\hat{\rho}_{x_t|x_{t-k}}$ are the conditional distributions for x_t , respectively, in the case where y_t is free or constrained.

If we consider a linear system as in 71, it is immediate to write z_t in dependence on z_{t-k} , i.e.

$$z_t = A^k z_{t-k} + \sum_{l=0}^{k-1} A^l \sigma w = A^k z_{t-k} + \sigma' w.$$

Going over the above arguments, we can obtain the same results as for the one-time entropy transfer by replacing A with its power A^k :

⁶To perform the calculation easily, it suffices to recall that, in the proposed notation, A_{2q} must be understood as the object made up of the second row of A and all columns included in the index q . Similarly Σ_{q2} is the object made by the second column of Σ and all the rows comprised in the index q .

Theorem 9 Given a linear system of the type 72, the transfer entropy $T_{y \rightarrow x}$, at time t and after k instants, is of the form

$$T_{y \rightarrow x}(t - k, t) = \frac{1}{2} \ln |A_{pq}^k (\Sigma_q - \Sigma_{qp} \Sigma_p^{-1} \Sigma_{pq}) A_{pq}^k{}^T + \sigma'^2 I_p| - \frac{1}{2} \ln |\sigma'^2 I_p|.$$

It turns out that $T_{y \rightarrow x}(t - k, t) = 0$ if $A_{pq}^k = 0$, i.e. if in the equations for x the coefficients for y are all zero.

Theorem 10 Given a linear system of the type 77, the transfer entropy $T_{y_c \rightarrow x_a}$, at time t and after k instants, is of the form

$$T_{y_c \rightarrow x_a}(t - k, t) = \frac{1}{2} \ln |A_{aq}^k (\Sigma_q - \Sigma_{qa} \Sigma_a^{-1} \Sigma_{aq}) A_{aq}^k{}^T + \sigma'^2 I_a| + \\ - \frac{1}{2} \ln |A_{a\hat{q}}^k (\Sigma_{\hat{q}} - \Sigma_{\hat{q}a} \Sigma_a^{-1} \Sigma_{a\hat{q}}) A_{a\hat{q}}^k{}^T + \sigma'^2 I_a|,$$

where we have defined $q := (b, c, d)$ and $\hat{q} = (b, d)$. In particular, it turns out that $T_{y_c \rightarrow x_a} = 0$ if $A_{ac}^k = 0$, i.e. if in the equations for x_a the coefficients in front of y_c are all zero.

Corollary 4 Under the same assumptions of Theorem 10, given $a = c = 1$, $x_a = x_j$ e $y_c = x_i$, it results that $T_{z_i \rightarrow z_j}(t - k, t) = 0$ iff $A_{ji}^k = 0$.

Remark 9 We can compute the transfers at k times for the model 64 and for the model 65, again assuming that the initial datum is assigned with covariance matrix $\sigma_0^2 I$. Let us start with the model 64.⁷

$$T_{x \rightarrow y}(0, k) = \frac{1}{2} \ln \left(1 + \hat{b}_x^2 \frac{\sigma_0^2}{\sigma'^2} \right), \quad \hat{b}_x := b_x \sum_{l=0}^{k-1} a_x^l b_y^{k-1-l},$$

$$T_{x \rightarrow z}(0, k) = \frac{1}{2} \ln \left(1 + \hat{c}_x^2 \frac{\sigma_0^2}{\sigma'^2} \right), \quad \hat{c}_x := c_x \sum_{l=0}^{k-1} a_x^l c_z^{k-1-l},$$

$$T_{y \rightarrow z}(0, k) = 0.$$

For the 65 model, let us simply consider the two-stage transfer, starting with the structure of the matrix

$$A^2 = \begin{pmatrix} a_x^2 + \epsilon b_x & \epsilon a_x + \epsilon b_y & \\ a_x b_x + b_x b_y & \epsilon b_x + b_y^2 & \\ a_x c_x + c_x c_z & \epsilon c_x & c_z^2 \end{pmatrix}.$$

⁷It is sufficient to observe that the matrix A^k has the same structure as the matrix A :

$$A^k = \begin{pmatrix} a_x^k & & \\ b_x \sum_{l=0}^{k-1} a_x^l b_y^{k-1-l} & b_y^k & \\ c_x \sum_{l=0}^{k-1} a_x^l c_z^{k-1-l} & & c_z^k \end{pmatrix} \quad (78)$$

Wanting to quantify the causal link $y \rightarrow x$, null for the model 64, the result is

$$A_{3q} = (a_x c_x + c_x c_z, \epsilon c_x), \quad q := (1, 2), \quad A_{3\hat{q}} = (a_x c_x + c_x c_z), \quad \hat{q} := (1),$$

$$\Sigma_{q3} = (0, 0)^T, \quad \Sigma_q^* = \Sigma_q - \Sigma_{q3} \Sigma_3^{-1} \Sigma_{3q} = \sigma_0^2 I, \quad \Sigma_{\hat{q}}^* = \sigma_0^2,$$

so that it is immediate to obtain

$$T_{y \rightarrow z}(0, 2) = \frac{1}{2} \ln \left(1 + \epsilon^2 \frac{\sigma_0^2 c_x^2}{\sigma_0^2 c_x^2 (a_x + c_z)^2 + \sigma'^2} \right),$$

obtaining a similar expression for the transfer to those found previously, this time depending on the y size of the feedback y on x .

Remark 10 *Having defined the quantities $T_{z_i \rightarrow z_j}(t - k, t)$ as k varies, we can introduce an index that measures causal links cumulatively over time. Imagining that we assign an initial condition for $t = 0$, via a covariance matrix $\Sigma_0 = \sigma_0^2 I$, to compute the transfers $(t - k, k)$ we must take into account the evolution of the matrix according to the Lemma 3. Imagining to perform these calculations, we could consider*

$$T_{z_i \rightarrow z_j}(t) := \sum_{k=1}^t T_{z_i \rightarrow z_j}(t - k, t), \quad (79)$$

$$T_{z_i \rightarrow z_j} := \lim_{t \rightarrow \infty} T_{z_i \rightarrow z_j}(t). \quad (80)$$

We expect that if the spectral radius of A is less than 1, then $T_{z_i \rightarrow z_j}$ is a finite quantity. In the next section, we will introduce the cumulative index $D_{z_i \rightarrow z_j}$, so that the problem of establishing a comparison between the two will arise.

Remark 11 *We can generalise the results obtained to the case where the state of the system at time t , z_t , depends on the s preceding states, $z_{t-1}, z_{t-2}, \dots, z_{t-s}$. Reconsidering the notation adopted earlier, we write*

$$z_t = F(z_{t-1}, z_{t-2}, \dots, z_{t-s}) + \sigma w,$$

$$\begin{cases} x_t = F_p(x_{t-1}, y_{t-1}, x_{t-2}, y_{t-2}, \dots, x_{t-s}, y_{t-s}) + \sigma_p w_p \\ y_t = F_q(x_{t-1}, y_{t-1}, x_{t-2}, y_{t-2}, \dots, x_{t-s}, y_{t-s}) + \sigma_q w_q \end{cases},$$

considering transfer entropy

$$T_{y \rightarrow x}^s(t - 1, t) := H[\rho_{x_t | x_{t-1} \dots x_{t-s}}] - H[\hat{\rho}_{x_t | x_{t-1} \dots x_{t-s}}],$$

where $\hat{\rho}$ is the distribution for the constrained system, obtained by posing $y_t = y_{t-1} = \dots = y_{t-s} = \hat{y}$. For the linear case, we have the system

$$\begin{cases} x_t = A_{p_1} x_{t-1} + A_{p_1 q_1} y_{t-1} + \dots + A_{p_s} x_{t-s} + A_{p_s q_s} y_{t-s} \\ y_t = A_{q_1 p_1} x_{t-1} + A_{q_1} y_{t-1} + \dots + A_{q_s p_s} x_{t-s} + A_{q_s} y_{t-s} \end{cases}. \quad (81)$$

To calculate the transfer entropy, the probability distribution must be determined for the free and the constrained system. It is therefore sufficient to proceed, taking care to reorder the notation correctly. Let us consider the Gaussian vector $(x_t, x_{t-1}, \dots, x_{t-s})$, adopting the following convention: the prime indices refer to the objects at time t , those with the subscript k to the objects at time $t - k$. By adopting the objects

$$\begin{aligned}\Sigma &= \begin{pmatrix} \Sigma_{p'} & \Sigma_{p'p_1} & \dots & \Sigma_{p'p_s} \\ \Sigma_{p_1p'} & \Sigma_{p_1} & \dots & \Sigma_{p_1p_s} \\ \dots & \dots & \dots & \dots \\ \Sigma_{p_s p'} & \Sigma_{p_s p_1} & \dots & \Sigma_{p_s} \end{pmatrix}, \\ \Sigma_{p'|p} &= \Sigma_{p'} - \Sigma_{p'p} \Sigma_p^{-1} \Sigma_{pp'}, \\ \Sigma_{p'p} &= (\Sigma_{p'p_1} \quad \Sigma_{p'p_2} \quad \dots \quad \Sigma_{p'p_s}) \\ \Sigma_p &= \begin{pmatrix} \Sigma_{p_1} & \Sigma_{p_1p_2} & \dots & \Sigma_{p_1p_s} \\ \Sigma_{p_2p_1} & \Sigma_{p_2} & \dots & \Sigma_{p_2p_s} \\ \dots & \dots & \dots & \dots \\ \Sigma_{p_s p_1} & \Sigma_{p_s p_2} & \dots & \Sigma_{p_s} \end{pmatrix}, \\ A_{pq} &= (A_{p_1q_1} \quad A_{p_2q_2} \quad \dots \quad A_{p_s q_s}), \\ \Sigma_q &= \begin{pmatrix} \Sigma_{q_1} & \Sigma_{q_1q_2} & \dots & \Sigma_{q_1q_s} \\ \Sigma_{q_2q_1} & \Sigma_{q_2} & \dots & \Sigma_{q_2q_s} \\ \dots & \dots & \dots & \dots \\ \Sigma_{q_s q_1} & \Sigma_{q_s q_2} & \dots & \Sigma_{q_s} \end{pmatrix}, \\ \Sigma_{pq} &= \begin{pmatrix} \Sigma_{p_1q_1} & \Sigma_{p_1q_2} & \dots & \Sigma_{p_1q_s} \\ \Sigma_{p_2q_1} & \Sigma_{p_2q_2} & \dots & \Sigma_{p_2q_s} \\ \dots & \dots & \dots & \dots \\ \Sigma_{p_s q_1} & \Sigma_{p_s q_2} & \dots & \Sigma_{p_s q_s} \end{pmatrix},\end{aligned}$$

we obtain the following theorem:

Theorem 11 *Given a linear system of the type 81, the transfer entropy $T_{y \rightarrow x}$ is of the form*

$$T_{y \rightarrow x}^s(t-1, t) = \frac{1}{2} \ln |A_{pq}(\Sigma_q - \Sigma_{qp} \Sigma_p^{-1} \Sigma_{pq}) A_{pq}^T + \sigma^2 I_p| - \frac{1}{2} \ln |\sigma^2 I_p|.$$

It turns out that $T_{y \rightarrow x}(t-1, t) = 0$ if $A_{pq} = 0$, i.e. if in the equations for x the coefficients for y are all null.

Revisiting the analysis discussed above, the corresponding adaptations also follow for this case, and it is possible to consider the transfer in k steps.

Remark 12 *Let us reconsider the conditioning problem, with the aim of explicitly writing down the expression of $T_{y_d \rightarrow x_a | y_c}$. It suffices to recall what is stated in Lemma 2 in iii., a direct consequence of the proposed definition of transfer:*

$$T_{y_c, y_d \rightarrow x} = T_{y_c \rightarrow x} + T_{y_d \rightarrow x | y_c}.$$

Given $p = (a, b)$, $q = (c, d)$, $\hat{q} = (d)$, as well as reconsidering the theorems 7 and 8, it is immediate to verify that

$$T_{y_d \rightarrow x_p | y_c} = T_{y_q \rightarrow x_p} - T_{y_c \rightarrow x_p} = \frac{1}{2} \ln \frac{|A_{p\hat{q}} \Sigma_{\hat{q}}^* A_{p\hat{q}}^T + \sigma^2 I_p|}{|\sigma^2 I_p|}.$$

Similarly, given $q = (b, c, d)$, $\hat{q} = (b, d)$, $\tilde{q} = (b)$, it holds

$$T_{y_d \rightarrow x_a | y_c} = T_{y_q \rightarrow x_a} - T_{y_c \rightarrow x_a} = \frac{1}{2} \ln \frac{|A_{a\hat{q}} \Sigma_{\hat{q}}^* A_{a\hat{q}}^T + \sigma^2 I_a|}{|A_{a\tilde{q}} \Sigma_{\tilde{q}}^* A_{a\tilde{q}}^T + \sigma^2 I_a|}.$$

If, for example, we consider the 64 model, it is easy to verify, in line with our expectations, that

$$\begin{aligned} T_{x \rightarrow z | y}(0, 1) &= T_{x, y \rightarrow z}(0, 1) = T_{x \rightarrow z}(0, 1), \\ T_{y \rightarrow z | x}(0, 1) &= 0. \end{aligned}$$

A similar check can be made for the model 65.

Linear Response

Let us consider [8], where causal links are characterized in terms of fluctuation-dissipation relations. In the linear case, this approach to causality gives us indications consistent with what was discussed in the previous section.

In the field of non-equilibrium statistical physics, response theory provides a relationship linking the spontaneous fluctuations of a system with the response following a perturbation. In general, we consider a n -component Markovian process, $z_t = (z_t^1, z_t^2, \dots, z_t^n)$, under the assumption that it admits an invariant distribution ρ_s , non-zero everywhere and regular. Given an instantaneous perturbation at time $t = 0$ on component z^i , say

$$z_0^i \rightarrow z_0^i + \delta z_0^i$$

, we have a probability distribution ρ'_s , for which the relation $\rho'(z_0) = \rho(z_0 - \delta z_0)$ holds. We can then consider the mean for an observable F , for the unperturbed and the perturbed case respectively, under the assumption that the dynamics of the system is governed by some law $z_t = \phi^t z_0$, deterministic or stochastic. We write

$$\langle F(z_t) \rangle = \int F(z_t) \rho_s(z_t, z_0) dz_t dz_0,$$

$$\langle F(z_t) \rangle' = \int F(z_t) \rho'_s(z_t, z_0) dz_t dz_0,$$

where $\rho_s(z_t, z_0) = p_{z_0 \rightarrow z_t} \rho_s(z_0)$, $\rho'_s(z_t, z_0) = p_{z_0 \rightarrow z_t} \rho'_s(z_0)$ and $p_{z_0 \rightarrow z_t}$ is the transition probability for the system from the state z_0 to z_t . It is easy to evaluate the difference between the two averages:

$$\overline{\delta F(z_t)} := \langle F(z_t) \rangle' - \langle F(z_t) \rangle = \int F(z_t) p_{z_0 \rightarrow z_t} \{ \rho'_s(z_0) - \rho_s(z_0) \} dz_t dz_0 =$$

$$\begin{aligned}
&= \int F(z_t) \frac{\rho'_s(z_0) - \rho_s(z_0)}{\rho_s(z_0)} p_{z_0 \rightarrow z_t} \rho_s(z_0) dz_t dz_0 = \\
&= - \int F(z_t) \sum_{k=1}^n \frac{\partial \ln \rho_s(z)}{\partial z^k} \Big|_{z_0} \delta z_0^k p_{z_0 \rightarrow z_t} \rho_s(z_0) dz_t dz_0 + O(|\delta z_0|^2).
\end{aligned}$$

Assuming that we have only perturbed the component z^i and neglecting higher-order terms, we obtain the relationship:

$$\frac{\overline{\delta F(z_t)}}{\delta z_0^i} = - \int F(z_t) \frac{\partial \ln \rho_s(z)}{\partial z^i} \Big|_{z_0} \rho_s(z_t, z_0) dz_0 dz_t. \quad (82)$$

If, instead of the generic function F , we consider the projection onto the component z^j , we define the linear response of z^j at time t , to the perturbation on z^i at time 0:

$$R_{z_i \rightarrow z_j}(0, t) := \frac{\overline{\delta z_t^j}}{\delta z_0^i} = - \int z_t^j \frac{\partial \ln \rho_s(z)}{\partial z^i} \Big|_{z_0} \rho_s(z_t, z_0) dz_0 dz_t. \quad (83)$$

We can then say that z^i influences z^j after t time steps if a small perturbation of z^i at time $t = 0$, say δz_0^i , produces a change on average for z_t^j . This change is quantified by the response $R_{z_i \rightarrow z_j}(0, t)$. We interpret the possibility that $R_{z_i \rightarrow z_j}(0, t) \neq 0$ as evidence that there is a causal connection from z_i to z_j , quantified by the value assumed by $R_{z_i \rightarrow z_j}(0, t)$.

Remark 13 *The assumption of regularity of the invariant distribution is essential to proceed with the first-order expansion and write the relation 82. This condition is not satisfied if the system has a nonsmooth attractor, a typical situation in the presence of a nonlinear term in the model equations. However, this limitation does not arise for the calculation of the transfer, where the concrete difficulty lies in determining explicitly the evolution of the distribution for the initial data.*

Linear Case

If we consider a linear system of the form $x_t = Ax_{t-1} + \sigma w$, it is easy to calculate the response, taking into account that ρ_s goes to zero at infinity and proceeding with some simple integration by parts.

$$\begin{aligned}
R_{z_i \rightarrow z_j}(0, t) &= - \int z_t^j \frac{\partial \ln \rho_s(z)}{\partial z^i} \Big|_{z_0} \rho_s(z_t, z_0) dz_0 dz_t = \\
&= - \int z_t^j dz_t \int \frac{\partial \rho_s(z_0)}{\partial z_0^i} p_{z_0 \rightarrow z_t} dz_0 = \int z_t^j dz_t \int \rho_s(z_0) \frac{\partial p_{z_0 \rightarrow z_t}}{\partial z_0^i} dz_0 = \\
&= \int z_t^j dz_t \int \rho_s(z_0) \sum_{k=1}^n A_{ki}^t \frac{\partial}{\partial z_t^k} p_{z_0 \rightarrow z_t} dz_0 = \\
&= - \sum_{k=1}^n \int A_{ki}^t \frac{\partial z_t^j}{\partial z_t^k} \rho_s(z_0) p_{z_0 \rightarrow z_t} dz_0 dz_t =
\end{aligned}$$

$$= -A_{ji}^t \int \rho_s(z_0) p_{z_0 \rightarrow z_t} dz_0 dz_t = -A_{ji}^t.$$

We observe that the relationship just established directly follows from the linear nature of the system, without the need to make assumptions about the distribution of the initial data or the shape of the invariant distribution.

Considering the discussion so far, we have proven the following result:

Theorem 12 *Given a linear system of the type 71, assuming that the initial data is distributed as a Gaussian, the transfer $T_{z_i \rightarrow z_j}(0, t)$ is zero if and only if the response $R_{z_i \rightarrow z_j}(0, t)$ is zero.*

We can therefore say that, under the aforementioned assumptions, we have a causal relationship in the sense of transfer if and only if we have it in the sense of response. Assuming that the initial data is distributed as a Gaussian with covariance matrix $\Sigma_0 = \sigma_0^2 I$, we can write a simple relationship between transfer and response.

$$T_{z_i \rightarrow z_j}(0, 1) = \frac{1}{2} \ln \left(1 + \{R_{z_i \rightarrow z_j}(0, 1)\}^2 \frac{\sigma_0^2}{\sigma_0^2 \sum_{l \neq i, j}^n a_{jl}^2 + \sigma^2} \right), \quad (84)$$

which can generalize to the case at time k ,

$$T_{z_i \rightarrow z_j}(0, k) = \frac{1}{2} \ln \left(1 + \{R_{z_i \rightarrow z_j}(0, k)\}^2 \frac{\sigma_0^2}{\sigma_0^2 \sum_{l \neq i, j}^n \hat{a}_{jl}^2 + \sigma'^2} \right), \quad (85)$$

Where \hat{a}_{jl} is the (j, l) -th element of the matrix A^k and σ' is the noise redefined at step k . These relationships can be generalized to the case of an arbitrary matrix Σ_0 , taking care to rewrite the involved coefficients accordingly. The problem arises of determining how the introduction of a nonlinear term affects the validity of these relationships.

Remark 14 *Given that the perturbation for z^i is instantaneous and at time $t = 0$, we can define a cumulative index in time for causality by summing the contributions given by the response for z^j at the various subsequent instants:*

$$D_{z_i \rightarrow z_j}(t) := \sum_{k=1}^t R_{z_i \rightarrow z_j}(0, k), \quad (86)$$

$$D_{z_i \rightarrow z_j} := \lim_{t \rightarrow \infty} D_{z_i \rightarrow z_j}(t). \quad (87)$$

These indices are of a different nature compared to those proposed for transfer, which take into account the information transferred for all the pairs $(t - k, t)$.

Revisiting the definition of transfer entropy in the framework of dynamical systems theory, we obtained an exact expression for linear models in the presence of Gaussian initial data and noise. This led to a necessary and sufficient condition for the existence

of a causal relationship between two variables in the system. This characterization can also be obtained in the context of response theory, which offers a different conceptual framework from that of information theory. The next step could involve discussing the robustness of this correspondence for nonlinear systems, starting from some specific cases.

References

- [1] S. Acid and L. M. De Campos. Searching for bayesian network structures in the space of restricted acyclic partially directed graphs. *Journal of Artificial Intelligence Research*, 18:445–490, 2003.
- [2] Mohammad Ali Alomrani. A critical review of information bottleneck theory and its applications to deep learning. *ArXiv*, abs/2105.04405, 2021.
- [3] Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, and Vince I Madai. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making*, 20(1):1–9, 2020.
- [4] Daniel J. Amit. *Modeling Brain Function—the World of Attractor Neural Networks*. Cambridge University Press, USA, 1989.
- [5] C. Anderson. The end of the theory: the Data Deluge Makes the Scientific Method Obsolete. *Wired*, 2008.
- [6] Susan Athey and Guido W. Imbens. The state of applied econometrics: Causality and policy evaluation. *The Journal of Economic Perspectives*, 31(2):3–32, 2017.
- [7] P. Baldi. *Deep Learning in Science*. Cambridge University Press, 2021.
- [8] Marco Baldovin, Fabio Cecconi, and Angelo Vulpiani. Understanding causation via correlations and linear response theory. *Phys. Rev. Res.*, 2:043436, Dec 2020.
- [9] Randall Balestriero, Jerome Pesenti, and Yann LeCun. Learning in high dimension always amounts to extrapolation. 10 2021.
- [10] Elias Bareinboim, Juan Correa, Duligur Ibeling, and Thomas Icard. On pearl’s hierarchy and the foundations of causal inference (1st edition). In Hector Geffner, Rina Dechter, and Joseph Y. Halpern, editors, *Probabilistic and Causal Inference: the Works of Judea Pearl*, pages 507–556. ACM Books, 2022.
- [11] Elias Bareinboim, Juan D Correa, Duligur Ibeling, and Thomas Icard. On pearl’s hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 507–556. 2022.
- [12] Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.
- [13] Trevor J. Barnes and Matthew W. Wilson. Big data, social physics, and spatial analysis: The early years. *Big Data & Society*, 1, 2014.
- [14] Lionel Barnett, Adam Barrett, and Anil Seth. Granger causality and transfer entropy are equivalent for gaussian variables. *Physical review letters*, 103:238701, 12 2009.

- [15] Solon Barocas, Andrew D Selbst, and Manish Raghavan. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 80–89, 2020.
- [16] David Barrett and Benoit Dherin. Implicit gradient regularization. In *International Conference on Learning Representations*.
- [17] J.M. Beggs. *The Cortex and the Critical Point: Understanding the Power of Emergence*. MIT Press, 2022.
- [18] Y. Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35:1798–1828, 08 2013.
- [19] William Bialek, Ilya Nemenman, and Naftali Tishby. Predictability, complexity, and learning. *Neural Computation*, 13:2409–2463, 2000.
- [20] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, 2007.
- [21] J. Mark Bishop. Artificial intelligence is stupid and causal reasoning will not fix it. *Frontiers in Psychology*, 11, 2021.
- [22] G. Boffetta, M. Cencini, M. Falcioni, and A. Vulpiani. Predictability: a way to characterize complexity. *Physics Reports*, 356(6):367–474, 2002.
- [23] Guido Boffetta, Giovanni Paladin, and Angelo Vulpiani. Strong chaos without the butterfly effect in dynamical systems with feedback. 29(10):2291, may 1996.
- [24] G. Bontempi, S. Ben Taieb, and Y. Le Borgne. *Machine Learning Strategies for Time Series Forecasting*, volume 138. 01 2013.
- [25] T. R. J. Bossomaier, Lionel C. Barnett, Michael Harr, and Joseph T. Lizier. An introduction to transfer entropy: Information flow in complex systems. 2016.
- [26] C. Calude and G. Longo. The deluge of spurious correlations in big data. *Foundations of Science*, 22, 09 2017.
- [27] Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová. Machine learning and the physical sciences. 2019.
- [28] N. Cartwright. *How the Laws of Physics Lie*. Clarendon paperbacks. Clarendon Press, 1983.
- [29] Nancy Cartwright. Causal laws and effective strategies. *Noûs*, 13:419, 1979.
- [30] P. Castiglione, M. Falcioni, A. Lesne, and A. Vulpiani. *Chaos and Coarse Graining in Statistical Mechanics*. Cambridge University Press, 2008.

- [31] F. Cecconi, M. Cencini, M Falcioni, and A. Vulpiani. Predicting the future from the past: An old problem from a modern perspective. *American Journal of Physics*, 80, 10 2012.
- [32] F. Cecconi, M. Cencini, and A. Vulpiani. *Chaos: From Simple Models to Complex Systems*. World Scientific, Singapore, 2013.
- [33] M. Chantry, H. Christensen, P. Dueben, and T. Palmer. Opportunities and challenges for machine learning in weather and climate modelling: hard, medium and soft ai. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194), 2021.
- [34] Gal Chechik, Amir Globerson, Naftali Tishby, and Yair Weiss. Information bottleneck for gaussian variables. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003.
- [35] S. Chibbaro and A. Vulpiani. Compressibility, laws of nature, initial conditions and complexity. *Foundations of Physics*, 47:1–19, 10 2017.
- [36] David Maxwell Chickering. Learning equivalence classes of bayesian-network structures. *The Journal of Machine Learning Research*, 2:445–498, 2002.
- [37] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.
- [38] P. Churchland. Mind-brain reduction: New light from the philosophy of science. *Neuroscience*, 7:1041–7, 06 1982.
- [39] P. Churchland. The impact of neuroscience on philosophy. *Neuron*, 60:409–11, 12 2008.
- [40] Diego Colombo, Marloes H Maathuis, et al. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, 15(1):3741–3782, 2014.
- [41] Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pages 294–321, 2012.
- [42] A Feder Cooper, Emanuel Moss, Benjamin Laufer, and Helen Nissenbaum. Accountability in an algorithmic society: relationality, responsibility, and robustness in machine learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 864–876, 2022.
- [43] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, 2012.
- [44] Felix Creutzig, Amir Globerson, and Naftali Tishby. Past-future information bottleneck in dynamical systems. *Physical review. E, Statistical, nonlinear, and soft matter physics*, 79:041925, 05 2009.

- [45] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Math. Control Signal Systems*, 2:303–314, 1989.
- [46] R.A. Fisher. *Statistical methods for research workers*. Edinburgh Oliver & Boyd, 1925.
- [47] L. Floridi. *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*. Oxford University Press, 2023.
- [48] Mathias Frisch. *Causal Reasoning in Physics*. Cambridge University Press, 2014.
- [49] Bernhard C. Geiger. On information plane analyses of neural network classifiers—a review. *IEEE Transactions on Neural Networks and Learning Systems*, 33:7039–7051, 2020.
- [50] AmirEmad Ghassami, Saber Salehkaleybar, Negar Kiyavash, and Kun Zhang. Counting and sampling from markov equivalent dags using clique trees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3664–3671, 2019.
- [51] Virginia Ghiara. Taking the russo-williamson thesis seriously in the social sciences. *Synthese*, 200(6), 2022.
- [52] Ran Gilad-Bachrach, Amir Navot, and Naftali Tishby. An information theoretic tradeoff between complexity and accuracy. In *Annual Conference Computational Learning Theory*, 2003.
- [53] Donald Gillies. The russo-williamson thesis and the question of whether smoking causes heart disease. In Phyllis McKay Illari, Federica Russo, and Jon Williamson, editors, *Causality in the Sciences*, pages 110–125. Oxford University Press, 2011.
- [54] Donald Gillies. *Causality, Probability, and Medicine*. New York: Routledge, 2017.
- [55] Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.
- [56] Ziv Goldfeld and Yury Polyanskiy. The information bottleneck problem and its applications in machine learning. *IEEE Journal on Selected Areas in Information Theory*, 1:19–38, 2020.
- [57] Ziv Goldfeld, Ewout Van Den Berg, Kristjan Greenewald, Igor Melnyk, Nam Nguyen, Brian Kingsbury, and Yury Polyanskiy. Estimating information flow in deep neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2299–2308. PMLR, 09–15 Jun 2019.

- [58] Joaquín Goñi, Bernat Corominas-Murtra, Ricard V Solé, and Carlos Rodríguez-Caso. Exploring the randomness of directed acyclic networks. *Physical Review E*, 82(6):066115, 2010.
- [59] I. J. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, Cambridge, MA, USA, 2016.
- [60] Clive W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, Aug 1969.
- [61] Joseph Y. Halpern. *Actual Causality*. MIT Press, Cambridge, MA, 2016.
- [62] James Douglas Hamilton. *Time Series Analysis*. Princeton University Press, 1994.
- [63] Peter Harremoës and Naftali Tishby. The information bottleneck revisited or how to choose a good distortion measure. In *2007 IEEE International Symposium on Information Theory*, pages 566–570, 2007.
- [64] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.
- [65] Kevin D. Hoover. Reductionism in economics: Intentionality and eschatological justification in the microfoundations of macroeconomics. *Philosophy of Science*, 82(4):689–711, 2015.
- [66] Kevin D. Hoover and Kevin D. Autor Hoover. *Causality in Macroeconomics*. Cambridge University Press, 2001.
- [67] K Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2):251–257, 1991.
- [68] H. Hosni and A. Vulpiani. Data science and the art of modelling. *Lettera Matematica*, 6:1–9, 05 2018.
- [69] H. Hosni and A. Vulpiani. Forecasting in light of big data. *Philosophy and Technology*, 31, 12 2018.
- [70] H. Hosni and A. Vulpiani. Random thoughts about complexity, data and models. 04 2020.
- [71] Yimin Huang and Marco Valtorta. Pearl’s calculus of intervention is complete. *ArXiv*, abs/1206.6831, 2006.
- [72] P.M.K. Illari and F. Russo. *Causality: Philosophical Theory Meets Scientific Practice*. Oxford University Press, 2014.

- [73] Dominik Janzing. The cause-effect problem: Motivation, ideas, and popular misconceptions. In *Cause Effect Pairs in Machine Learning*, 2019.
- [74] Dominik Janzing and Bernhard Schölkopf. Causal inference using the algorithmic markov condition. *IEEE Transactions on Information Theory*, 56:5168–5194, 2008.
- [75] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.
- [76] M. Kac. On the notion of recurrence in discrete stochastic processes. *Bullettin of the American Mathematical Society*, 53:1002–1010, 1947.
- [77] Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H Maathuis, and Peter Bühlmann. Causal inference using graphical models with the r package pcalg. *Journal of statistical software*, 47:1–26, 2012.
- [78] George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6), 5 2021.
- [79] Brian Karrer and Mark EJ Newman. Random graph models for directed acyclic networks. *Physical Review E*, 80(4):046110, 2009.
- [80] A. Katok and B. Hasselblatt. *Introduction to the Modern Theory of Dynamical Systems*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1997.
- [81] A.I.A. Khinchin. *Mathematical Foundations of Information Theory*. Dover Books on Mathematics. Dover Publications, 1957.
- [82] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. Adaptive computation and machine learning. MIT Press, 2009.
- [83] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.
- [84] Shane Legg and Marcus Hutter. A collection of definitions of intelligence. In *Artificial General Intelligence*, 2007.
- [85] Shane Legg and Marcus Hutter. Universal intelligence: A definition of machine intelligence. *Minds and Machines*, 17(4):391–444, 2007.

- [86] Maria Leggio, Laura Mandolesi, Francesca Federico, Francesca Spirito, Benedetta Ricci, Francesca Gelfo, and Laura Petrosini. Environmental enrichment promotes improved spatial abilities and enhanced dendritic growth in rat. *Behavioural brain research*, 163:78–90, 09 2005.
- [87] Sabina Leonelli. *Data-Centric Biology: A Philosophical Study*. London: University of Chicago Press, 2016.
- [88] R. Lguensat, P. Tandeo, P. Ailliot, M. Pulido, and R. Fablet. The analog data assimilation. *Monthly Weather Review*, 145(10):4093 – 4107, 2017.
- [89] Ming Li and Paul Vitanyi. *An introduction to Kolmogorov complexity and its applications*. Springer Verlag, London, 1997.
- [90] B. Lim and S. Zohren. Time-series forecasting with deep learning: a survey. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194), 2021.
- [91] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2020.
- [92] E. N. Lorenz. Deterministic nonperiodic flow. *J. Atmos. Sci.*, 20:130–141, 1963.
- [93] E. N. Lorenz. Atmospheric predictability as revealed by naturally occurring analogues. *Journal of Atmospheric Sciences*, 26(4):636 – 646, 1969.
- [94] E. N. Lorenz. Three approaches to atmospheric predictability. *Bull. Amer. Meteor. Soc.*, 50, 1969.
- [95] E. N. Lorenz. Predictability. a problem partly solved. *Proc. Seminar on Predictability (ECMWF, Reading, UK, 1996)*., page 1–18, 1996.
- [96] Marloes H Maathuis, Diego Colombo, Markus Kalisch, and Peter Bühlmann. Predicting causal effects in large-scale systems from observational data. *Nature methods*, 7(4):247–248, 2010.
- [97] Marloes H Maathuis, Markus Kalisch, and Peter Bühlmann. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164, 2009.
- [98] D.J.C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.
- [99] Sara Magliacane, Tom Claassen, and Joris M Mooij. Ancestral causal inference. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2016.

- [100] Sarah Marzen and Simon DeDeo. The evolution of lossy compression. *Journal of The Royal Society Interface*, 14, 06 2015.
- [101] Warren Mcculloch and Walter Pitts. A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5:127–147, 1943.
- [102] Christopher Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, UAI'95*, page 403–410, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.
- [103] P. Mehta, C. Wang, A.G.R Day, Richardson C., M. Bukov, C. K. Fisher, and Schwab D. J. A high-bias, low-variance introduction to machine learning for physicists. *Physics Reports*, 810:1–124, 2019.
- [104] Franz H Messerli. Chocolate Consumption, Cognitive Function, and Nobel Laureates. *New England Journal of Medicine*, 367(16):1562–1564, 2012.
- [105] Marvin Minsky and Seymour Papert. *Perceptrons: An Introduction to Computational Geometry*. MIT Press, Cambridge, MA, USA, 1969.
- [106] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. Adaptive Computation and Machine Learning series. MIT Press, 2012.
- [107] Joris Mooij, Dominik Janzing, and Bernhard Schölkopf. From ordinary differential equations to structural causal models: the deterministic case. *Uncertainty in Artificial Intelligence - Proceedings of the 29th Conference, UAI 2013*, 04 2013.
- [108] K. P. Murphy. *Machine learning. A Probabilistic Perspective*. MIT Press, 2013.
- [109] E. Nagel. *The Structure of Science: Problems in the Logic of Scientific Explanation*. Donald F. Koch American Philosophy Collection. Harcourt, Brace & World, 1961.
- [110] A. Ng. *Machine Learning Yearning*. 2017.
- [111] C. Nicolis. Atmospheric analogs and recurrence time statistics: Toward a dynamical formulation. *Journal of the Atmospheric Sciences*, 55, 3 1998.
- [112] Claudio Novelli, Mariarosaria Taddeo, and Luciano Floridi. Accountability in artificial intelligence: what it is and how it works. *AI & SOCIETY*, pages 1–12, 2023.
- [113] E. Ott. *Chaos in Dynamical Systems*. Cambridge University Press, 1993.
- [114] J. Pathak, Z. Lu, B. Hunt, M. Girvan, and E. Ott. Using machine learning to replicate chaotic attractors and calculate lyapunov exponents from data. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 27, 10 2017.

- [115] J. Pearl. *Causality*. Cambridge University Press, 2 edition, 2009.
- [116] J. Pearl, M. Glymour, and N.P. Jewell. *Causal Inference in Statistics: A Primer*. Wiley, 2016.
- [117] J. Pearl and D. Mackenzie. *The Book of Why*. Basic Books, 2018.
- [118] Judea Pearl. *Probabilistic reasoning in intelligent systems : networks of plausible inference*. Morgan Kaufmann, San Francisco, Calif., 2009.
- [119] Judea Pearl. Bayesian networks. 2011.
- [120] Judea Pearl. The do-calculus revisited. In Nando de Freitas and Kevin P. Murphy, editors, *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA, August 14-18, 2012*, pages 3–11. AUAI Press, 2012.
- [121] Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Commun. ACM*, 62(3):54–60, feb 2019.
- [122] Judea Pearl and Azaria Paz. Graphoids: Graph-based logic for reasoning about relevance relations or when would x tell you more about y if you already know z? In Benedict du Boulay, David C. Hogg, and Luc Steels, editors, *Advances in Artificial Intelligence II, Seventh European Conference on Artificial Intelligence, ECAI 1986, Brighton, UK, July 20-25, 1986, Proceedings*, pages 357–363. North-Holland.
- [123] Judea Pearl and Thomas Verma. The logic of representing dependencies by directed graphs. In *Proceedings of the Sixth National Conference on Artificial Intelligence - Volume 1, AAAI’87*, page 374–379. AAAI Press, 1987.
- [124] Emilija Perkovic. Identifying causal effects in maximally oriented partially directed acyclic graphs. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 530–539. PMLR, 2020.
- [125] Emilija Perkovic. Identifying causal effects in maximally oriented partially directed acyclic graphs. In *Conference on Uncertainty in Artificial Intelligence*, pages 530–539. PMLR, 2020.
- [126] Emilija Perković, Markus Kalisch, and Maloes H Maathuis. Interpreting and using cpdags with background knowledge. *arXiv preprint arXiv:1707.02171*, 2017.
- [127] J.M. Peters, P.L. Bühlmann, and B. Schölkopf. *Restricted Structural Equation Models for Causal Inference*. Dissertationen, PhD Thesis. ETH, 2012.
- [128] Jonas Peters and Peter Bühlmann. Identifiability of gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

- [129] Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2017.
- [130] W. Pietsch. *Big Data*. Elements in the Philosophy of Science. Cambridge University Press, 2021.
- [131] W. Pietsch. *On the Epistemology of Data Science: Conceptual Tools for a New Inductivism*. Springer Verlag, 2022.
- [132] H. Poincaré. Sur le problème des trois corps et les équations de la dynamique. *Acta Mathematica*, 13(1,2), 1980.
- [133] G. Popkin. A twisted path to equation-free prediction. *Quanta Magazine*, 10 2015.
- [134] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.
- [135] Hans Reichenbach. *The Direction of Time*. Mineola, N.Y.: Dover Publications, 1956.
- [136] Thomas Richardson and Peter Spirtes. Ancestral graph Markov models. *The Annals of Statistics*, 30(4):962 – 1030, 2002.
- [137] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.
- [138] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.
- [139] Bertrand Russell. On the notion of cause. *Proceedings of the Aristotelian Society*, 13(1):1–26, 07 2015.
- [140] Federica Russo. *Causality and Causal Modelling in the Social Sciences*. Springer, Dordrecht, 2009.
- [141] Federica Russo and Jon Williamson. Interpreting causality in the health sciences. *International Studies in the Philosophy of Science*, 21:157 – 170, 2007.
- [142] Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J Anders, and Klaus-Robert Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278, 2021.
- [143] Iqbal H Sarker. Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, 2(3):160, 2021.

- [144] Richard Scheines. An introduction to causal inference. 1997.
- [145] Richard Scheines, Peter Spirtes, Clark Glymour, Christopher Meek, and Thomas Richardson. The tetrad project: Constraint based aids to causal model specification. *Multivariate Behavioral Research*, 33(1):65–117, 1998.
- [146] B. Schölkopf, F. Locatello, S. Bauer, N.R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Towards causal representation learning. *CoRR*, 2021.
- [147] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Conference on International Conference on Machine Learning*, ICML’12, page 459–466, Madison, WI, USA, 2012. Omnipress.
- [148] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.
- [149] Thomas Schreiber. Measuring information transfer. *Phys. Rev. Lett.*, 85:461–464, Jul 2000.
- [150] M. G. Schultz, C. Betancourt, B. Gong, F. Kleinert, M. Langguth, L. H. Leufen, A. Mozaffari, and S. Stadtler. Can deep learning beat numerical weather prediction? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194), 2021.
- [151] B. Schölkopf. Causality for machine learning. 2019.
- [152] B. Schölkopf and J. von Kügelgen. From statistical to causal learning. 04 2022.
- [153] Angelo Vulpiani Sergio Chibbaro. *Reductionism, Emergence and Levels of Reality: The Importance of Being Borderline*. Springer, 2014.
- [154] James P. Sethna. *Statistical Mechanics: Entropy, Order Parameters and Complexity*. Oxford University Press, 2006.
- [155] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, USA, 2014.
- [156] Ohad Shamir, Sivan Sabato, and Naftali Tishby. Learning and generalization with the information bottleneck. *Theoretical Computer Science*, 411(29):2696–2711, 2010. Algorithmic Learning Theory (ALT 2008).
- [157] C. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, 1948.
- [158] R. Shwartz-Ziv and N. Tishby. Opening the black box of deep neural networks via information. 03 2017.

- [159] Ravid Shwartz-Ziv. Information flow in deep neural networks. *Ph.D Thesis*, 02 2022.
- [160] Jenni Sidey-Gibbons and Chris Sidey-Gibbons. Machine learning in medicine: a practical introduction. *BMC Medical Research Methodology*, 19, 03 2019.
- [161] D. Silver, A. Huang, C. Maddison, A. Guez, L. Sifre, G. Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489, 01 2016.
- [162] E. Simpson. The interpretation of interaction in contingency tables. *Journal of the royal statistical society series b-methodological*, 13:238–241, 1951.
- [163] Joy Simpson and John Kelly. The impact of environmental enrichment in laboratory rats—behavioural and neurochemical aspects. *Behavioural brain research*, 222:246–64, 09 2011.
- [164] L. Sklar. *Physics and Chance: Philosophical Issues in the Foundations of Statistical Mechanics*. Philosophical Issues in the Foundations of Statistical Mecha. Cambridge University Press, 1993.
- [165] Noam Slonim. The information bottleneck: Theory and applications. *Ph.D Thesis*, 01 2002.
- [166] Noam Slonim and Naftali Tishby. Agglomerative information bottleneck. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999.
- [167] Samuel L. Smith, Benoit Dherin, David G. T. Barrett, and Soham De. On the origin of implicit regularization in stochastic gradient descent. *CoRR*, abs/2101.12176, 2021.
- [168] Payal Soni, Yogya Tewari, and Deepa Krishnan. Machine learning approaches in stock price prediction: A systematic review. *Journal of Physics: Conference Series*, 2161:012065, 01 2022.
- [169] Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):62–72, 1991.
- [170] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.
- [171] Peter Spirtes, Christopher Meek, and Thomas Richardson. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, UAI’95*, page 499–506, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

- [172] Dietrich Stauffer. Introduction to statistical physics outside physics. *Physica A: Statistical Mechanics and its Applications*, 336(1):1–5, 2004. Proceedings of the XVIII Max Born Symposium ”Statistical Physics outside Physics”.
- [173] Dietrich Stauffer. A biased review of sociophysics. *Journal of Statistical Physics*, 151:9–20, 2012.
- [174] John Q. Stewart. Suggested principles of ”social physics”. *Science*, 106(2748):179–180, 1947.
- [175] John Q. Stewart. Demographic gravitation: Evidence and applications. *Sociometry*, 11(1/2):31–58, 1948.
- [176] F. Takens. Detecting strange attractors in turbulence. *Lecture Notes in Mathematics, Berlin Springer Verlag*, 1981.
- [177] Stefan Thurner, Peter Klimek, and Rudolf Hanel. *Introduction to the Theory of Complex Systems*. Oxford University Press, 09 2018.
- [178] N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. *Proceedings of the 37th Allerton Conference on Communication, Control and Computation*, 49, 07 2001.
- [179] N. Tishby and N. Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5, 2015.
- [180] Caroline Uhler, Garvesh Raskutti, Peter Bühlmann, and Bin Yu. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, pages 436–463, 2013.
- [181] Leslie G. Valiant. A theory of the learnable. In *Symposium on the Theory of Computing*, 1984.
- [182] H. M. Van Den Dool. Searching for analogues, how long must we wait? *Tellus A*, 46(3):314–324, 1994.
- [183] V. Vapnik. *The Nature of Statistical Learning Theory*. Information Science and Statistics. Springer New York, 2013.
- [184] V.N. Vapnik and V.N. Vapnik. *Statistical Learning Theory*. A Wiley-Interscience publication. Wiley, 1998.
- [185] Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, UAI ’90, page 255–270, USA, 1990. Elsevier Science Inc.
- [186] Eugene Wigner. The Unreasonable Effectiveness of Mathematics in the Natural Sciences. *Communications in Pure and Applied Mathematics*, 13, 1960.

- [187] Jon Williamson. *Bayesian Nets and Causality: Philosophical and Computational Foundations*. Oxford, England: Oxford University Press, 2004.
- [188] Jon Williamson. Establishing causal claims in medicine. *International Studies in the Philosophy of Science*, 32(1):33–61, 2019.
- [189] Janine Witte, Leonard Henckel, Marloes H Maathuis, and Vanessa Didelez. On efficient adjustment in causal graphs. *Journal of Machine Learning Research*, 21:246, 2020.
- [190] James Woodward. Causation with a human face. In Huw Price and Richard Corry, editors, *Causation, Physics, and the Constitution of Reality: Russell’s Republic Revisited*. Oxford University Press, 2007.
- [191] Tailin Wu. *Intelligence, physics and information – the tradeoff between accuracy and simplicity in machine learning*. PhD thesis, 01 2020.
- [192] Stefano Zamagni. Economic reductionism as a hindrance to the analysis of structural change: scattered notes. *Structural Change and Economic Dynamics*, 11:197–208, 2000.
- [193] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. 2016.
- [194] Jie M Zhang, Mark Harman, Lei Ma, and Yang Liu. Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering*, 48(1):1–36, 2020.
- [195] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 804–813, 2011.
- [196] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI’11, page 804–813, Arlington, Virginia, USA, 2011. AUAI Press.
- [197] George K. Zipf. The unity of nature, least-action, and natural social science. *Sociometry*, 5, 1942.
- [198] George K. Zipf. *Human Behaviour and the Principle of Least Effort*. Addison-Wesley, 1949.