Journal of applied logics (Print) ISSN 2631-9810
Journal of applied logics (Online) ISSN 2631-9829

Journal of Applied Logics

The IfCoLog Journal of Logics and their Applications

Volume 11 ● Issue 1 ● January 2024

Special Issue on Foundations, Applications and Theory of Inductive Logic

Guest Editors
Martin Adamčík
Matthias Thimm

Available online at
www.collegepublications.co.uk/journals/ifcolog/
Free open access

Journal of Applied Logics - IfCoLog Journal of Logics and their Applications

Volume 11, Number 1

January 2024

Disclaimer

Statements of fact and opinion in the articles in Journal of Applied Logics - IfCoLog Journal of Logics and their Applications (JALs-FLAP) are those of the respective authors and contributors and not of the JALs-FLAP. Neither College Publications nor the JALs-FLAP make any representation, express or implied, in respect of the accuracy of the material in this journal and cannot accept any legal responsibility or liability for any errors or omissions that may be made. The reader should make his/her own evaluation as to the appropriateness or otherwise of any experimental technique described.

ISBN 978-1-84890-449-1
ISSN (E) 2631-9829
ISSN (P) 2631-9810

College Publications
Scientific Director: Dov Gabbay
Managing Director: Jane Spurr

http://www.collegepublications.co.uk

Editorial Board

Scope and Submissions

This journal considers submission in all areas of pure and applied logic, including:

pure logical systems
proof theory
constructive logic
categorical logic
modal and temporal logic
model theory
recursion theory
type theory
nominal theory
nonclassical logics
nonmonotonic logic
numerical and uncertainty reasoning
logic and AI
foundations of logic programming
belief change/revision
systems of knowledge and belief
logics and semantics of programming
specification and verification
agent theory
databases

dynamic logic
quantum logic
algebraic logic
logic and cognition
probabilistic logic
logic and networks
neuro-logical systems
complexity
argumentation theory
logic and computation
logic and language
logic engineering
knowledge-based systems
automated reasoning
knowledge representation
logic in hardware and VLSI
natural language
concurrent computation
planning

This journal will also consider papers on the application of logic in other subject areas: philosophy, cognitive science, physics etc. provided they have some formal content.

Submissions should be sent to Jane Spurr (jane@janespurr.net) as a pdf file, preferably compiled in LaTeX using the IFCoLog class file.

Contents

ARTICLES

Editorial Note for the Speical Issue on Foundations, Applications, and Theory of Inductive Logic 1
Martin Adamčík and Matthias Thimm

An Intuitive Introduction to Information Geometry 5
Martin Adamčík

Analogical Proportion-based Induction: From Classification to Creativity . 51
Henri Prade and Gilles Richard

Inductive Reasoning, Conditionals, and Belief Dynamics 89
Gabriele Kern-Isberner and Wolfgang Spohn

Special Issue on Foundations, Applications, and Theory of Inductive Logic

Martin Adamčík
Assumption University, Thailand
maths38@gmail.com

Matthias Thimm
Artificial Intelligence Group, University of Hagen, Germany
matthias.thimm@fernuni-hagen.de

Inductive reasoning is one of the most important reasoning techniques for humans and formalises the intuitive notion of “reasoning from experience”. It has thus influenced both theoretical work on the formalisation of rational models of thought in Philosophy as well as practical applications in the areas of Artificial Intelligence and, in particular, Machine Learning.

This special issue is a follow-up to the First International Conference on Foundations, Applications, and Theory of Inductive Logic (FATIL2022) that took place on October 12-14, 2022, in Munich, Germany.[1] It aimed at bringing together experts from all fields concerned with inductive reasoning. This included in particular the following aspects:

- Foundations of many of our best theories crucially depend on inductive logic and more widely induction. Uncertainty is ubiquitous in our lives and the philosophical problem arises to make sense of probabilities and to act sensibly in the face of uncertainties. General philosophy of science is much interested in (the reconstruction of) rational inference in general and in science, in particular, in cases with inconclusive evidence.

- Theory of inductive inference can be developed within several traditions such as pure inductive logic or inductive logic based on the maximum entropy principle.

- Applications have sprung from foundational thinking on induction in computer and data science. This includes aspects such as knowledge representation in multi-agent settings and machine learning approaches (such as inductive logic programming).

[1] http://fatil2022.krportal.org

The special issue welcomed contributions in all areas dealing with inductive reasoning. We specifically welcomed extended versions of works presented at FATIL2022, but the call was open for further works as well. Topics of interest included, but were not limited to:

- Foundational works about inductive reasoning, inductive logic, and induction, in particular critical examinations of existing principles.
- Computational approaches to inductive reasoning, in particular non-monotonic and other non-classical logics.
- Computational approaches to reasoning under uncertainty.
- Machine learning approaches taking inductive reasoning into account such as inductive logic programming.

This special issues features three contributions from areas described above.

The first contribution "An intuitive introduction to information geometry" by Martin Adamčík introduces traditional information geometry, and in particular the convergence of the alternating minimisation procedure, from a position of an inductive logician. It discusses geometrical principles that one would find reasonable when asked to merge several conflicting beliefs of rational agents or information sources. This is done in an entirely abstract setting without references to heavy theory, and it should be accessible to anyone in the target audience of the Journal of Applied Logics. Traditional and more technical Euclidean and Hilbertian settings are presented towards the end of the paper.

The second contribution is "Analogical proportion-based induction: From classification to creativity" by Henri Prade and Gilles Richard. Here, analogical inference is considered as a specific form of inductive reasoning, and similarities of and differences between these modes of reasoning are discussed. Analogical proportions are used as a technical framework to implement analogical reasoning and the paper further analyses analogical proportions-based classification as a specific application problem. Further, matters of creativity are discussed by considering analogical proportions as an instance of logical properties and the paper discusses the latter in depth. This could perhaps draw attention of the wider analogical reasoning community towards logic, while it provides an interesting topic for the readers of this journal.

The final contribution is "Inductive Reasoning, Conditionals, and Belief Dynamics" by Gabriele Kern-Isberner and Wolfgang Spohn. The authors consider inductive reasoning as a special case of belief revision with epistemic states, the latter being implemented by both probability distributions as well as ranking functions. The framework allows for reasoning with conditional information and the incorporation of background knowledge. In their presentation, the authors take perspectives of both Artificial Intelligence and Philosophy into account.

The editors are grateful to all the authors, and equally to the reviewers, for their contribution. Special thanks go to Jürgen Landes for making the conference, and thus this special issue, possible.

Received January 2024

An Intuitive Introduction to Information Geometry

Martin Adamčík*
Assumption University of Thailand
maths38@gmail.com

Abstract

In this paper, we recover some traditional results in the geometry of probability distributions, and in particular the convergence of the alternating minimisation procedure, without actually referring to probability distributions. We will do this by discussing a new general concept of two types of points: admissible and agreeable, inspired by multi–agent uncertain reasoning and belief merging. On the one hand, this presents a unique opportunity to make traditional results accessible to a wider audience as no prior knowledge of the topic is required. On the other hand, it allows us to contemplate how a group of rational agents would seek an agreement given their beliefs without necessarily expressing it in terms of probability distributions, focusing instead on logical properties. Finally, we recover Euclidean and Hilbertian settings of discrete and continuous probability distributions.

Keywords: Information geometry, Divergence, Uncertain reasoning, Belief merging, Fixed point, Alternating minimisation procedure, Bregman divergence, L^2 space

1 Introduction

The inspiration for the information geometry presented in this paper is the problem of merging several conflicting beliefs of rational agents or information sources. The key application that will be further elaborated once we develop the necessary theory is one of combining several medical studies. Combining more information sources presents an advantage over basing medical recommendations on a single study but

*Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) — 432308570. Thanks also go to anonymous referees who helped to improve the paper, and to Jürgen Landes and Matthias Thimm who organised the First International Conference on Foundations, Applications, and Theory of Inductive Logic, where this work was first presented.

presents a challenge of dealing with conflicting information and complex knowledge; rarely are two studies set exactly in the same way. Once complex aspects that are related but go beyond the original scope of the study are considered, each individual study's findings become constraints on what is admissible. We may then assign a mathematical object to represent the complex knowledge of all the studies. The practical examples, however, demonstrated that it would be unlikely that say a fixed probability distribution of patients (across some mutually exclusive and exhaustive categories) is found admissible by all of the studies.

Nevertheless, we would hope that by developing some geometry of these mathematical objects, we establish a procedure to find a point that could be called an agreement. Not in an arbitrary way, as there are a plethora of random choices that we can make, but in an intuitive way where the geometry is built from principles that we find reasonable. This is pretty much how the traditional axiomatic geometry was developed because only the intuitive approach gives us hope that we build a theory that resembles our world.

In fact, there are axiomatic frameworks in information geometry that fully capture the notions of cross–entropy (Shore and Johnson, [27]) and entropy (Paris and Vencovská, [25]), and intuitive principles that postulate how we should merge conflicting beliefs of several rational agents or information sources in a propositional setting (Konieczny and Pino–Pérez, [22]) and a probabilistic setting (Wilmers, [30]). The main feature that distinguishes this paper from the previous axiomatic frameworks of information geometry is the absence of any reference to probability distributions and that we also deal with conflicting information. On the other hand, our emphasis on information geometry is what differentiates our properties from those mentioned in the earlier approaches to inconsistent belief merging.

Yet the geometry that we will develop is not just geometry, but it is the information geometry of alternating minimisation procedure due to Csiszár [12, 13], which has been generalised in literature many times [8] and that also stands behind the motivating application of combining several real medical studies from [4].

Our approach, where the usual Euclidean and Hilbertian discrete and continuous probability distributions are introduced only after the main results are derived, provides a simple and captivating introduction to information geometry that will only progressively get more challenging. Section 2 can be read by anyone regardless of mathematical specialisation, while Section 5 will present results where familiarity with function spaces might be required.

More specifically, in Section 2 we will introduce information geometry on an entirely abstract concept of admissible and agreeable points. No reference will be made to usual Euclidean or Hilbertian space settings: to discrete or continuous probability distributions. This presents a unique point of view, inspired by multi–agent uncertain reasoning and belief merging, that does not seem to have yet appeared in the literature and which promises to be a fun introduction to the subject.

In Section 3 we will increase the difficulty by adding a metric topology to introduce a key procedure of information geometry: the alternating minimisation procedure. This procedure kept appearing in various forms before and after Csiszár [12, 13], who was perhaps the first to state the most general form of it and at the same time, write a completely correct proof for it. The procedure presented here is likewise not more general; it is merely a different presentation of the famous result.

Section 4 will give us an opportunity to explain how the geometry works in the Euclidean setting of discrete probability distributions and reference some literature. While a traditional paper would have this section right at the beginning, we did not want to discourage a potential reader from fields of logic and philosophy with unnecessary terminology when presenting the results of previous sections. There, we wished to present information geometry as driven by reason and logical principles as opposed to something hidden in mathematical formalism. The section, however, contains a further explanation of the real practical application of information geometry and the alternating minimisation procedure that we have mentioned and should not be omitted by anyone interested in applying information geometry.

Finally, Section 5 contains a Hilbertian space setting, where all results are proven and not referenced as in Section 4, and these proofs can also be easily modified to supply proofs omitted in Section 4. While general information geometric approaches focus on a whole Hilbertian space setting [10] or a general function space setting [14], we will work here in the L^2 space setting, which is the intersection of the two. This will simplify things, going well with the overall spirit of the paper, and the author is not aware that the provided proofs have yet appeared in this form elsewhere.

2 Information geometry

2.1 Intuition

> "A point is that which has no part."
>
> Euclid of Alexandria, [20]

Whenever we build a mathematical theory, we need to consult our intuition. Should we not do it, we may end up building a theory that little resembles the world we are living in and which is equally inapplicable. In this section, we will start building an intuitive framework that deals with information. We will need to confer with our intuition in the form of our experience on how information is used and how conflicting statements are dealt with.

Our first notion will indeed be the *point*. As in the Euclidean definition that starts this section, it is a building block that is further indivisible. Our point is, however, introduced to represent information rather than the position in a three–dimensional world. We think of several different beliefs on a particular matter; each different belief can be represented as a point. We are not concerned with what further constitutes the belief and we disregard any knowledge concerning the origins of the belief; it is simply an indivisible entity to us.

The points, which we have just introduced, can have any of the two following properties in this paper:

1. They can represent an admissible (i.e., consistent with the knowledge) collective point of view of a given group of rational agents or a collection of information sources, shortly called simply an *admissible point*,

2. and to represent an agreement of the group (resolving potentially conflicting nature of knowledge possessed by distinct agents or information sources), shortly called an *agreeable point*.

Now, intuitively, an admissible point is meant only to represent the state of collective knowledge, individual members of the group could well disagree and there could be no, what we call, agreeable point. An example following this paragraph will illustrate this. An agreeable point is thus an idealised point where all agents or sources agree with each other, despite practically such a point would often not be consistent with given knowledge. This collective framework of belief merging, that deals with individually consistent but jointly possibly inconsistent beliefs of a group of rational agents, was pioneered by Wilmers [30], and we are directly extending it here. An illustration is in Figure 1.

Example. *To illustrate, one scientific study could suggest that the proportion of people that develop a particular disease is somewhere between* 10% *and* 30% *while the other study could indicate that this value is between* 20% *and* 50%*. One way of constructing a point is to specify an ordered pair of individually admissible proportions such as* $(25\%, 40\%)$*, where the first number is admissible according to the first study and the second number is admissible according to the second study. Agreeable admissible points in this particular representation will be the points* (x, x)*,* $x \in [20\%, 30\%]$*, clearly representing the proportions on which the studies agree at the same time. There are other agreeable points that are not admissible, such as* $(35\%, 35\%)$*,* $(50\%, 50\%)$*,* $(0\%, 0\%)$*, and so on.*

The example above illustrates the kind of details we will need to go into before the intuitive concept that we develop here can be applied, but at this stage, working out

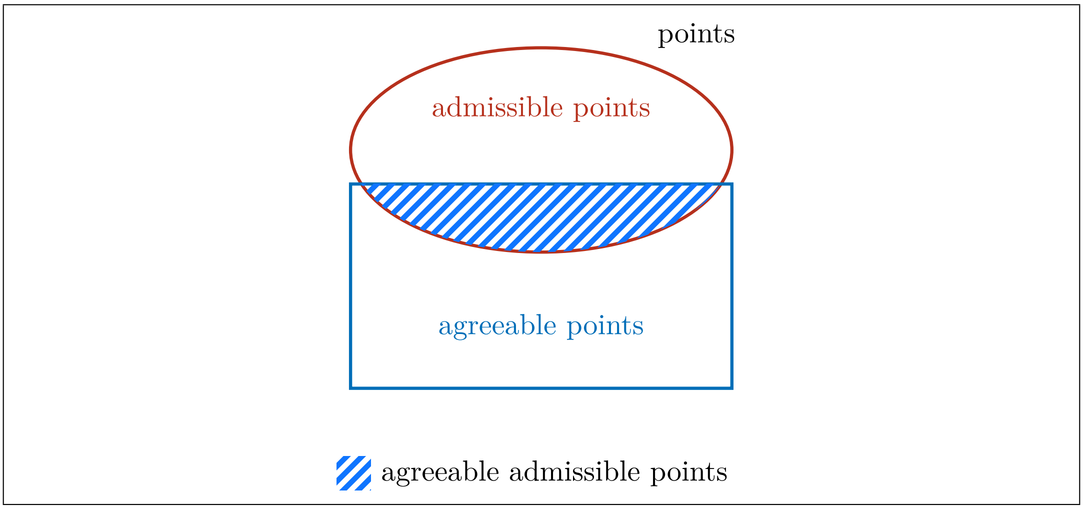


Figure 1: An illustration of the set of all points.

the details would only obstruct the general idea and the intuition behind it. We have therefore moved all technical examples and relevant references to Section 4. Here we only point out that our illustration fits Paris–Vencovská framework of uncertain reasoning as in [24] and that while the example above does not constitute the only line of application, representing agreeable points as those where all agents or sources agree is typical, and we will see this in Section 4.

The previous example was also straightforward enough in establishing agreeable points, but the following questions naturally arise:

1. What shall we do if admissible points contain no agreeable points?
2. How should we measure some kind of distance between an admissible point and an agreeable point in an effort to find the closest points of agreement?
3. Which intuitive principles such a notion of distance should satisfy?

These questions reflect the intuition that not all agreements, among those that a rational agent does not find admissible, are equally undesirable. With a notion of distance, we will be able to quantify this.

2.2 Information divergence

In the previous section, we saw the need for expressing some sort of information distance between two points, let us denote the set of all points as X, but we would not want to require much from this notion at this early stage. In particular, there is no apparent need for it to be a *metric*.

A metric is a symmetric distance between a pair of elements $\mathbf{x}$ and $\mathbf{y}$ of a set. It assigns to each pair $(\mathbf{x}, \mathbf{y})$ a non–negative real number $d(\mathbf{x}, \mathbf{y})$, this number is independent of the order of elements, it is zero if and only if the elements are identical and it satisfies the triangular inequality: $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z})$.

Instead, we will consider a much weaker notion of *information divergence*, a mapping D that assigns an ordered pair of points in X a non–negative real number:

$$D : X \times X \to \mathbb{R}, \text{ and } D(\mathbf{x}, \mathbf{y}) \geq 0.$$

We say that $D(\mathbf{x}, \mathbf{y})$ is the D *information divergence* from $\mathbf{x}$ to $\mathbf{y}$. Since symmetry is not required, the D information divergence from $\mathbf{y}$ to $\mathbf{x}$ could be different and therefore we do not call it a distance but a divergence.

Now, let $W \subseteq X$ be the set of all admissible points and $V \subseteq X$ be the set of agreeable points. Throughout the paper, we will assume that they are both non–empty, but we will not specifically require anything else from them before we reach Section 3. Formally, W and V are simply subsets of X, nothing more.

Let $\Delta(W) \subseteq V$ be the set of all those agreeable points $\mathbf{v}$ that are such that $D(\mathbf{v}, \mathbf{w})$ is minimal subject to $\mathbf{v} \in V$ and $\mathbf{w} \in W$:

$$\Delta(W) = \Big\{ \arg\min_{\mathbf{v} \in V} D(\mathbf{v}, \mathbf{w})\text{: subject to } \mathbf{w} \in W \Big\}.$$

In other words, we are looking here at all pairs $(\mathbf{v}, \mathbf{w})$, $\mathbf{v} \in V$ and $\mathbf{w} \in W$, establishing the minimal $D(\mathbf{v}, \mathbf{w})$ if it exists, and collecting all those $\mathbf{v}$ from V that give this minimal divergence into $\Delta(W)$. The purpose of the set $\Delta(W)$ is to determine those agreeable points that have the smallest D information divergence from them to admissible points and to use them as representatives of the set of all admissible points W. In other words, $\Delta(W) \subseteq V$ represents W; it is the agreement of a group of rational agents or a collection of information sources. We will call the points in $\Delta(W)$ *representative points*. See Figure 2 for an illustration.

Intuitively, if $W \cap V \neq \emptyset$; i.e., there are agreeable admissible points as in the example in the previous section, we expect the representation $\Delta(W) \subseteq V$ of W to be formed only by agreeable admissible points, although this intuition is not universally accepted. Williamson in [29] argues that such a principle is too strong as several rational agents may find an agreeable point to be admissible for inconsistent reasons. Nevertheless, as these reasons are in our setting unknown, it would not be rational to use them as an argument against our intuition. The following property of D guarantees that the above is the case:

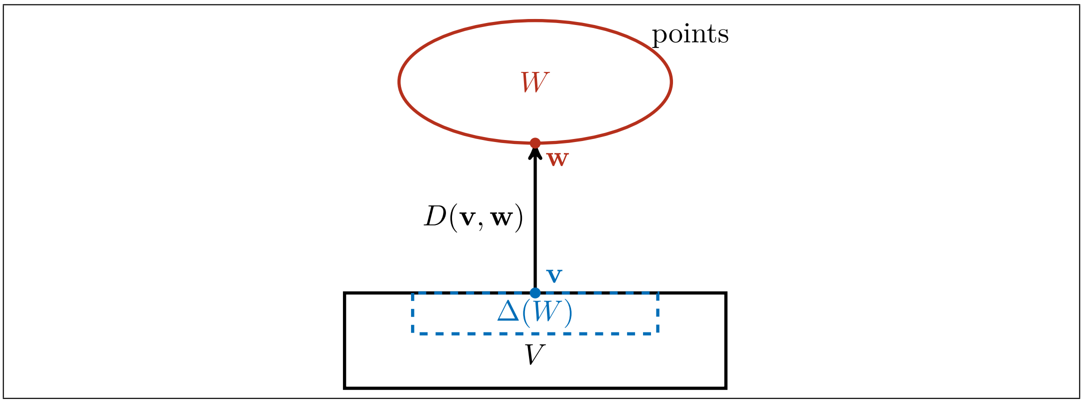


Figure 2: An illustration of the representative points. The direction of the arrow indicates that we consider here the divergence from $\mathbf{v}$ to $\mathbf{w}$. We should not, however, mistake this with the direction of Δ: It takes W and finds its (set) image in V.

Property 1 (Consistency)**.** *Let* $\mathbf{v}$ *and* $\mathbf{w}$ *be any two points in* X*. Then*

$$D(\mathbf{v},\mathbf{w}) = 0 \textit{ if and only if } \mathbf{v} = \mathbf{w}.$$

The meaning of $\mathbf{v} = \mathbf{w}$ for points is that $\mathbf{v}$ and $\mathbf{w}$ are identical: they denote the same point in X. In illustrations, if we draw two different points then they are not identical. For example, in Figure 2 we have that $\mathbf{w} \neq \mathbf{v}$.

Observation 1. *Let* D *be such that it satisfies the consistency property. If there are agreeable admissible points then agreeable admissible points form all representative points;*

$$\textit{if } W \cap V \neq \emptyset \textit{ then } \Delta(W) = W \cap V.$$

Proof. First, if $\mathbf{v} \in W \cap V$ then by the consistency property $D(\mathbf{v},\mathbf{v}) = 0$. We conclude that $\mathbf{v} \in \Delta(W)$ as $\mathbf{v}$ minimises $D(\mathbf{v},\mathbf{w})$ subject to $\mathbf{v} \in V$ and $\mathbf{w} \in W$. (Note that $D(\mathbf{v},\mathbf{v})$ cannot be smaller than zero by definition.) Hence $\Delta(W) \supseteq W \cap V$.

Second, assume that $W \cap V \neq \emptyset$ and $\mathbf{v} \in \Delta(W) \subseteq V$ is such that $\mathbf{v} \notin W$. Then $D(\mathbf{v},\mathbf{w}) = 0$ for some $\mathbf{w} \in W$, which by the consistency principle gives $\mathbf{v} = \mathbf{w}$. Hence $\Delta(W) \subseteq W \cap V$. □

The consistency property above is formulated more strongly than it is needed to prove Observation 1. Rather than considering any points $\mathbf{v}$ and $\mathbf{w}$, we could have required it only for $\mathbf{v} \in V$ and $\mathbf{w} \in W$. The reason for our choice is that we will need the stronger version later on.

In contrast, if $\mathbf{v} = \mathbf{w}$ implies $D(\mathbf{v}, \mathbf{w}) = 0$ but there are $\mathbf{v} \neq \mathbf{w}$ such that $D(\mathbf{v}, \mathbf{w}) = 0$, it could be possible to have $W \cap V \neq \emptyset$ and $\Delta(W) \supsetneqq W \cap V$, so further weakening of the consistency property would be undesirable.

2.3 Projections

Our notion of an information divergence is too general to have further useful properties on its own; in particular, if there are no agreeable admissible points, we cannot even say that the set of all representative points is always non–empty. We will keep adding assumptions concerning both D and sets of agreeable and admissible points V and W based on what appears rational to us in the context of information geometry. At some point, however, we will need to show that the list of our assumptions is consistent; we will need to find a particular information divergence, and sets W and V, that satisfy all those assumptions.

In this section, we will require D to have the following properties:

Property 2 (Projection)**.** *Assume that* $\mathbf{v}$ *is a given agreeable point. Then there is a unique admissible point* $\mathbf{w}$ *such that* $D(\mathbf{v}, \mathbf{w})$ *is minimal among all* $D(\mathbf{v}, \mathbf{y})$ *subject to* $\mathbf{y} \in W$*.*

The unique point $\mathbf{w}$ from the previous property will be denoted $\pi_W(\mathbf{v})$; it is the *D–projection* of $\mathbf{v}$ into W. An illustration is in Figure 3.

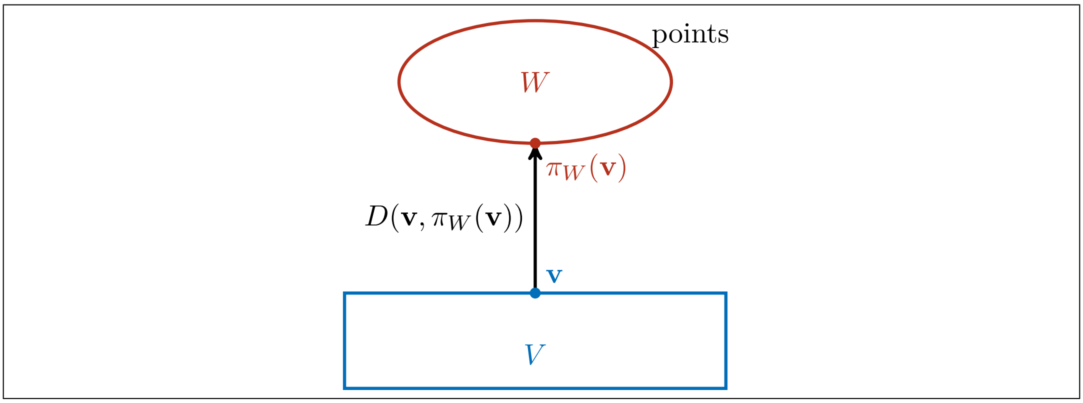


Figure 3: An illustration of the D–projection. The arrow again indicates the direction of the divergence. It happens that this direction coincides with the direction of the projection, but this is actually incidental as we will shortly see.

Property 3 (Conjugated Projection)**.** *Assume that* $\mathbf{w}$ *is a given admissible point. Then there is a unique agreeable point* $\mathbf{v}$ *such that* $D(\mathbf{v}, \mathbf{w})$ *is minimal among all* $D(\mathbf{x}, \mathbf{w})$ *subject to* $\mathbf{x} \in V$*.*

The unique point $\mathbf{v}$ from the previous property will be denoted $\widehat{\pi}_V(\mathbf{w})$; it is the *conjugated D–projection* of $\mathbf{w}$ into V. An illustration is in Figure 4.

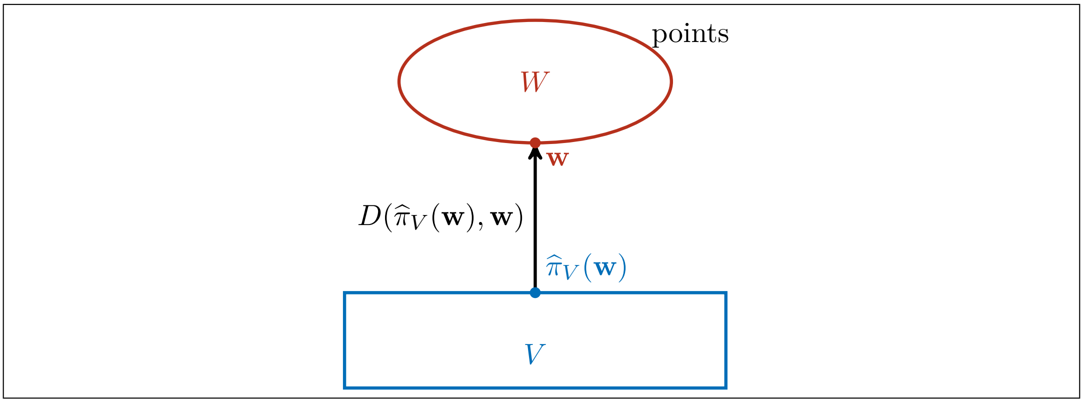

points
W
$\mathbf{w}$
$D(\widehat{\pi}_V(\mathbf{w}), \mathbf{w})$
$\widehat{\pi}_V(\mathbf{w})$
V

Figure 4: An illustration of the conjugated D–projection. Note that the direction of the divergence remains unchanged from Figure 3. We always consider a divergence from an agreeable point to an admissible point and not the other way around.

Intuitively, if we present a group of rational agents with a point of agreement, we expect them to find a single point among those they consider admissible as their personal opinion in view of the presented agreement. On the other hand, we should be able to establish agreement regardless of which specific admissible point the group presents to us. The required uniqueness of the projection and the conjugated projection contrasts with the possibility of finding multiple representative points, which are the solutions to a similar but more complex optimisation problem: We minimise $D(\mathbf{x}, \mathbf{y})$ subject to $\mathbf{x} \in V$ and $\mathbf{y} \in W$, being able to change two, not one variable freely.

Taking this further, we do not expect rational agents to solve complex optimisation problems at once. Instead, the following process taken from [1] and inspired by an earlier version of [30] could resemble a real–life agreement seeking:

Example. *Consider a group of rational agents with their set of admissible points W, which represents their individual knowledge or beliefs. The group elects a committee whose task is to find a single agreeable point from the set V: to merge their knowledge or beliefs. Naturally, the committee presents the group with their personal opinion or any other provisional starting point $\mathbf{v}_0$ that they see appropriate. The group then decides which point from those they consider admissible must have been the case to reach the conclusion suggested by the committee; they project the committee's point to their set W. At this stage, being present with a single admissible point, it is now possible for the committee to determine the conjugated projection of that* *admissible point to the set V: finding the corresponding agreeable point $\mathbf{v}_1$ of the group. Now, it is not at all necessary that $\mathbf{v}_1 = \mathbf{v}_0$. Nevertheless, the committee would be compelled to iterate the whole process until the above process stabilises on a single agreeable point. Otherwise, when they present the group with the conclusion (an agreeable point), the group would identify a different admissible point than the one that actually led the committee to make this conclusion.*

The points of interest from the previous example, although at this stage it is not clear if they even exist, will be called *fixed points*. More explicitly, an agreeable point $\mathbf{v} \in V$ is a fixed point if

$$\widehat{\pi}_V(\pi_W(\mathbf{v})) = \mathbf{v}.$$

The set of all fixed points will be denoted $\Theta(W)$⊆ V. See Figure 5 for an illustration.

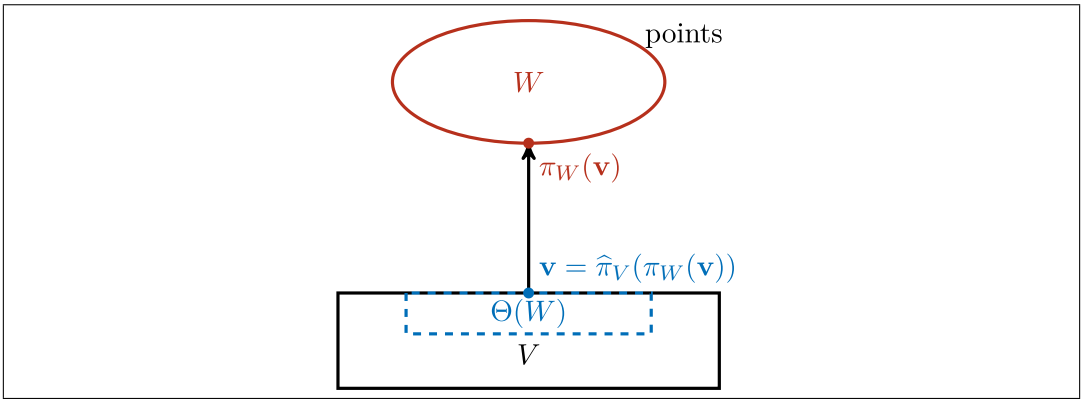


Figure 5: An illustration of the fixed points.

The example above is of course only one possible way of finding an agreement, although we argue that it is a rational one. An interesting question is how this way relates to previously suggested information divergence D minimisation, which yields the set $\Delta(W)$. There could be something:

Observation 2. *Let D be such that it satisfies the projection and conjugated projection properties. Then representative points are also fixed points:*

$$\Delta(W) \subseteq \Theta(W).$$

Proof. If $\Delta(W) = \emptyset$, which well could be the case, the statement holds trivially. So we investigate the case when $\Delta(W) \neq \emptyset$. Let $\mathbf{v} \in \Delta(W)$, and let d be the smallest D information divergence among all $D(\mathbf{x}, \mathbf{y})$ subject to $\mathbf{x} \in V$ and $\mathbf{y} \in W$. Such a real number exists by the definition of $\Delta(W)$, if this set is indeed non–empty, and note that in this paper we always assume that both V and W are non–empty.

Clearly, $D(\mathbf{v}, \pi_W(\mathbf{v})) \geq d$. Now assume that $D(\mathbf{v}, \pi_W(\mathbf{v})) > d$ so there must be $\mathbf{w} \in W$ such that $D(\mathbf{v}, \pi_W(\mathbf{v})) > D(\mathbf{v}, \mathbf{w})$. But this contradicts the definition of $\pi_W(\mathbf{v})$. So it must be that

$$D(\mathbf{v}, \pi_W(\mathbf{v})) = d.$$

Now, assume that $\widehat{\pi}_V(\pi_W(\mathbf{v})) \neq \mathbf{v}$. Nevertheless,

$$D(\widehat{\pi}_V(\pi_W(\mathbf{v})), \pi_W(\mathbf{v})) = D(\mathbf{v}, \pi_W(\mathbf{v})) = d,$$

otherwise we would contradict the definition of $\widehat{\pi}_V(\pi_W(\mathbf{v}))$. Finally, the equation above implies that both $\mathbf{x} = \mathbf{v}$ and $\mathbf{x} = \widehat{\pi}_V(\pi_W(\mathbf{v}))$ minimise $D(\mathbf{x}, \pi_W(\mathbf{v}))$ subject to $\mathbf{x} \in V$ for a given $\pi_W(\mathbf{v})$. Such a minimiser is, however, by the conjugated projection property required to be unique, thus

$$\widehat{\pi}_V(\pi_W(\mathbf{v})) = \mathbf{v}.$$

□

It seems that after concluding this section we have more questions than answers:

1. What properties should we require from an information divergence D, and sets W and V, so that $\Delta(W) = \Theta(W)$? If fixed points resemble real life agreement seeking, we would intuitively hope that they coincide with our representative points. But is it possible?

2. If we iterate the process from the example above; i.e., create a sequence $\{\mathbf{v}_i\}_{i=0}^{\infty}$, where $\mathbf{v}_{i+1} = \widehat{\pi}_V(\pi_W(\mathbf{v}_i))$, what properties should we require from the information divergence D, and sets W and V, so that we find an agreement in that way?

We shall find answers in the following sections.

2.4 Obdurate committee

> “The point is that we are not ignoring the dynamics, and we are not getting something from nothing, (. . .) for these all circumstances that are not under the experimenter’s control must, of necessity, be irrelevant. (. . .) Solution by the Maximum Entropy Principle is so unbelievably simple just because it eliminates those irrelevant details right at the beginning of the calculation by averaging over them.”

Edwin T. Jaynes, [18]

In this section, we further illustrate the setting by considering an obdurate committee that stubbornly refuses to iterate the process $\mathbf{v}_1 = \widehat{\pi}_V(\pi_W(\mathbf{v}_0))$, which we discussed in the previous section, after its first iteration. This will help us to connect our setting with some traditional concepts of information geometry.

First, we postulate the existence of the *prior agreeable point* $\mathbf{o}$ in the set of agreeable points V. This is an agreement that we would be compelled to select in the absence of any other information; of course, we have presented no justification as to why such a point should exist in our general context, we have only postulated it. However, the maximum entropy principle from the citation above gives such a justification in a more specific context [18]. Second, the group finds $\pi_W(\mathbf{o})$, a unique admissible point that has the smallest D–divergence from $\mathbf{o}$. If we wanted to represent W by a single admissible point, this is the most natural option as, with respect to D, it has the least 'distance' to $\mathbf{o}$ among the admissible points.

This generalises the concept of the famous *most entropic point* (also known as *MaxEnt*); we recover the usual concept if we choose a specific information divergence D and a specific concept of the point. We will elaborate on the details in Section 4. We only mention that the group is not ignoring the dynamics of the set of admissible points W by selecting that single point there as well as the experimenter is not doing so in the citation above. If the dynamics were laboriously worked out, we would have obtained this solution anyway.

Let us denote $\pi_W(\mathbf{o})$ by $\mathbf{ME}_D(W)$, and call it the most entropic point in W (with respect to D). Now, the committee wishes to find the corresponding agreeable point (if it is not already an agreeable admissible point). To that end, $\widehat{\pi}_V(\mathbf{ME}_D(W))$ is picked, and we denote $O(W) = \{\widehat{\pi}_V(\mathbf{ME}_D(W))\}$ the singleton containing it. An illustration is in Figure 6.

This *obdurate point* need not be a fixed point; and even less a representative point, considering Observation 2 on Page 14. It would indeed be a stubborn committee not to iterate the process further but be content with it. The committee would argue that the advantage of O is that it contains a single point. We would point out that $O(W) \neq W \cap V = \Delta(W)$, if $W \cap V \neq \emptyset$ and $W \cap V$ has at least two elements (given D satisfies the consistency property), as shown in Observation 1 on Page 11. Nevertheless, starting the whole iteration process from the prior agreeable point appears a well justified idea (what other point should the committee start with other than the 'prior', given such a point exists) that indeed might lead to a unique point as investigated in Section 3.2, and that has been practically applied as explained in Section 4.3.

Finally, let us point out the following obvious statements.

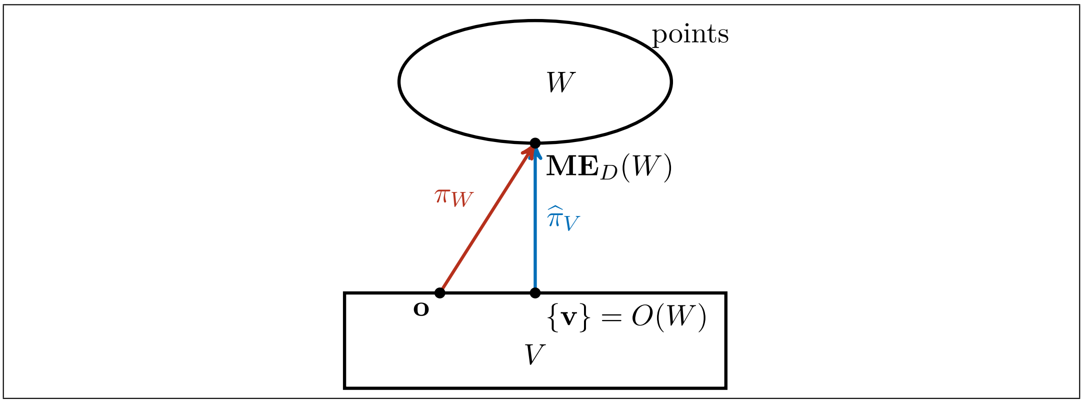


Figure 6: An illustration of an obdurate committee.

Observation 3. *If W is a singleton, then*

$$O(W) = \Delta(W).$$

Observation 4. *If $W \subseteq V$, then*

$$O(W) \subseteq \Delta(W).$$

The prior follows from Property 3, while the latter follows from Observation 1.

2.5 Pythagorean properties

The following property informally says that a group might establish the divergence of their agreement to an arbitrary admissible point by adding the divergence of their agreement to the conjugated projection of that admissible point and the divergence of the conjugated projection to the admissible point concerned.

Property 4 (Pythagorean for Agreeable Points)**.** *Let $\mathbf{v} \in V$ be an agreeable point and $\mathbf{w} \in W$ be an admissible point. Then*

$$D(\mathbf{v}, \widehat{\pi}_V(\mathbf{w})) + D(\widehat{\pi}_V(\mathbf{w}), \mathbf{w}) = D(\mathbf{v}, \mathbf{w}).$$

This property is counter–intuitive from the point of view of the classical Euclidean distance. Although it does not violate the triangular inequality, it is certainly not a property of the distance we are used to. On the other hand, it quite closely resembles how squares taken over the sides of a right–angled triangle behave in Euclidean geometry (hence the name), see Figures 7 and 8 for an illustration.

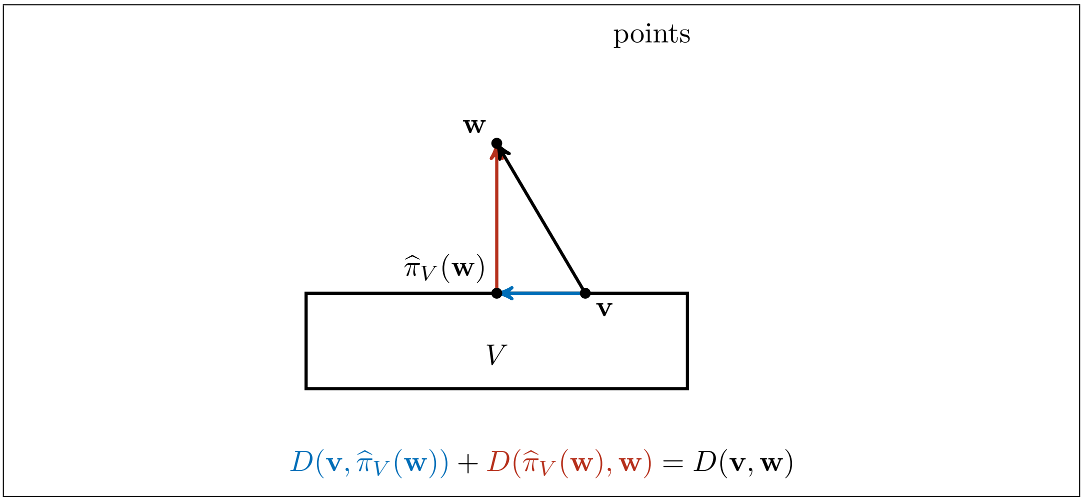

points
$\mathbf{w}$
$\widehat{\pi}_V(\mathbf{w})$
$\mathbf{v}$
V
$D(\mathbf{v}, \widehat{\pi}_V(\mathbf{w})) + D(\widehat{\pi}_V(\mathbf{w}), \mathbf{w}) = D(\mathbf{v}, \mathbf{w})$

Figure 7: An illustration of the Pythagorean property for agreeable points.

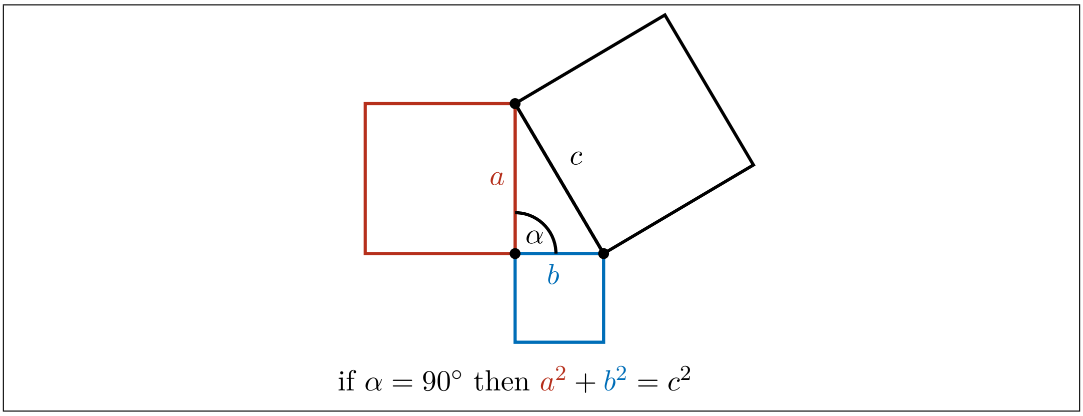


Figure 8: How squares behave in Euclidean geometry.

Intuitively, using an analogy from Euclidean geometry, we expect the set of agreeable points with respect to the conjugated D–projection to behave as a flat space into which we project admissible points. Something rather similar happens in least–squares linear regression, where a data vector is projected to a flat space defined by the linear model. Although this is clearly not an unusual idea, it is quite a strong requirement; we would not want to be so harsh on the set of admissible points. The following property will make admissible points to behave as a convex set.

Property 5 (Pythagorean for Admissible Points)**.** *Let* $\mathbf{v} \in V$ *be an agreeable point* *and* $\mathbf{w} \in W$ *be an admissible point. Then*

$$D(\mathbf{v}, \pi_W(\mathbf{v})) + D(\pi_W(\mathbf{v}), \mathbf{w}) \leq D(\mathbf{v}, \mathbf{w}).$$

While the Pythagorean naming convention here is now well established and used by Csiszár [12], who is the standard reference in the field, we may note that Shore and Johnson called the property above the triangular property in their work [27].

This property is similar to the Pythagorean property for agreeable points but it is weaker. And if the inequality from the statement actually holds in some cases for a particular D then this information divergence D is not a metric because it would not satisfy the triangular inequality. This is possibly the reason why the naming by Shore and Johnson did not catch up. See Figures 9 and 10 for an illustration.

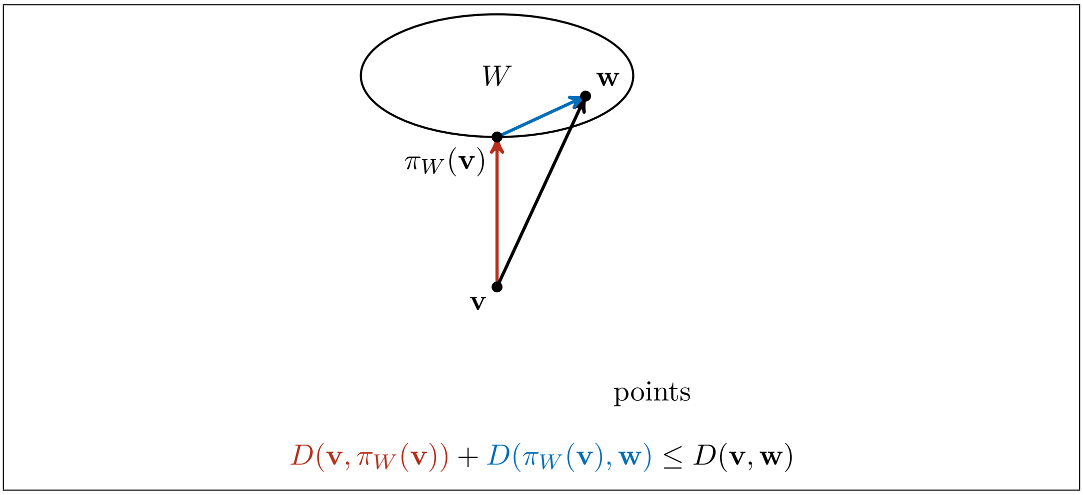

W
$\mathbf{w}$
$\pi_W(\mathbf{v})$
$\mathbf{v}$
points
$D(\mathbf{v}, \pi_W(\mathbf{v})) + D(\pi_W(\mathbf{v}), \mathbf{w}) \leq D(\mathbf{v}, \mathbf{w})$

Figure 9: An illustration of the Pythagorean property for admissible points.

The following observation gives us something that also follows from the consistency property on Page 11, but without assuming it.

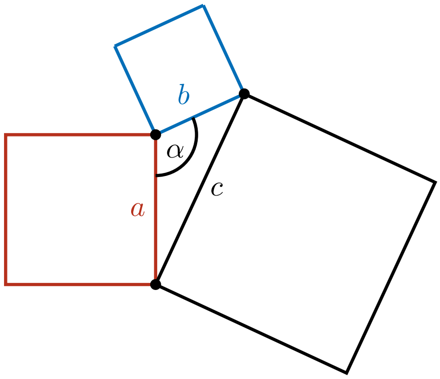
Observation 5. *Let D be such that it satisfies the projection and conjugated projection properties, and the Pythagorean properties for agreeable and admissible points. If $\mathbf{v} \in V$ is a fixed point then $D(\mathbf{v}, \mathbf{v}) = 0$ and $D(\pi_W(\mathbf{v}), \pi_W(\mathbf{v})) = 0$.*

Proof. By the Pythagorean property for agreeable points

$$D(\mathbf{v}, \widehat{\pi}_V(\pi_W(\mathbf{v}))) + D(\widehat{\pi}_V(\pi_W(\mathbf{v})), \pi_W(\mathbf{v})) = D(\mathbf{v}, \pi_W(\mathbf{v})).$$

But since $\mathbf{v}$ is fixed, the above is equivalent to

$$D(\mathbf{v}, \mathbf{v}) + D(\mathbf{v}, \pi_W(\mathbf{v})) = D(\mathbf{v}, \pi_W(\mathbf{v})),$$



if 90° $\leq \alpha \leq$ 180° then $a^2 + b^2 \leq c^2$

Figure 10: How squares behave in Euclidean geometry.

which is possible only if $D(\mathbf{v}, \mathbf{v}) = 0$.

Similarly, by the Pythagorean property for admissible points

$$D(\mathbf{v}, \pi_W(\mathbf{v})) + D(\pi_W(\mathbf{v}), \pi_W(\mathbf{v})) \leq D(\mathbf{v}, \pi_W(\mathbf{v})),$$

which is possible, due to non–negativity of information divergence, only if

$$D(\pi_W(\mathbf{v}), \pi_W(\mathbf{v})) = 0.$$

□

2.6 Fixed points are representative points

The following natural property says that the D information divergence from one admissible point to another admissible point should not be smaller than the D information divergence from and to the corresponding agreeable points. Intuitively, seeking an agreement should not take us further apart, see Figure 11. Please notice that we do not require that this takes us closer: equality is permitted.

Property 6 (Convexity)**.** *Let* $\mathbf{w}, \mathbf{u} \in W$*. Then*

$$D(\mathbf{w}, \mathbf{u}) \geq D(\widehat{\pi}_V(\mathbf{w}), \widehat{\pi}_V(\mathbf{u})).$$

We now have all the tools sufficient to prove that fixed points are also representative points, if there is actually a representative point.

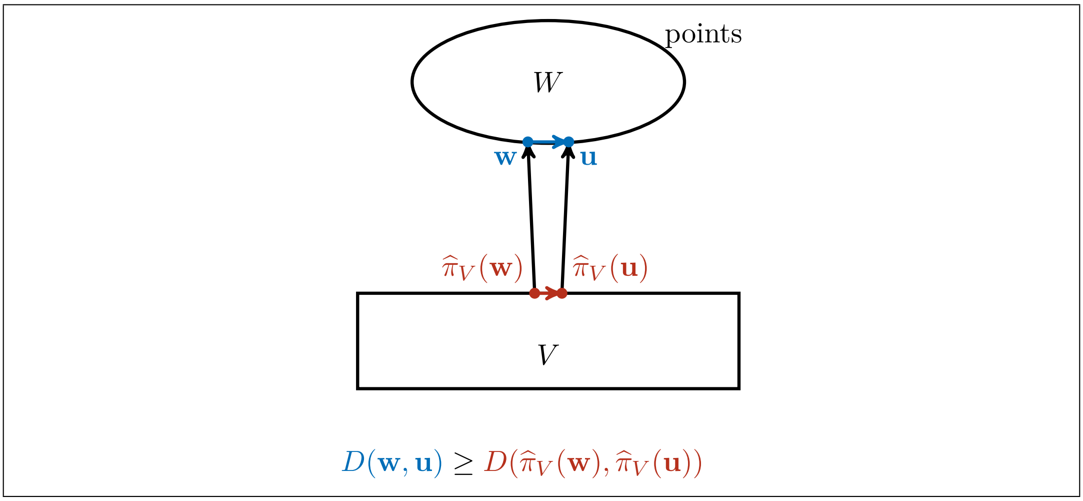

points
W
$\mathbf{w}$
$\mathbf{u}$
$\widehat{\pi}_V(\mathbf{w})$
$\widehat{\pi}_V(\mathbf{u})$
V
$D(\mathbf{w},\mathbf{u}) \geq D(\widehat{\pi}_V(\mathbf{w}),\widehat{\pi}_V(\mathbf{u}))$

Figure 11: An illustration of the convexity property.

Theorem 1 (Characterisation)**.** *Let D be such that it satisfies the projection and conjugated projection properties, the Pythagorean properties for both admissible and agreeable points, and the convexity property. If a representative point exists then the set of fixed points and the set of representative points are equal:*

$$\Delta(W) = \Theta(W).$$

Proof. By Observation 2 on Page 14 we already know that $\Delta(W) \subseteq \Theta(W)$, so it is sufficient to show that $\Delta(W) \supseteq \Theta(W)$.

Because we have assumed that a representative point exists and we already know that every representative point is also a fixed point, we may assume that $\widehat{\pi}_V(\mathbf{w}) \in \Delta(W)$, for some $\mathbf{w} \in W$. To make the argument, we now also assume that $\mathbf{v} \in \Theta(W)$ and show that $\mathbf{v} \in \Delta(W)$ in what follows.

The Pythagorean property for agreeable points

$$D(\mathbf{v}, \widehat{\pi}_V(\mathbf{w})) + D(\widehat{\pi}_V(\mathbf{w}), \mathbf{w}) = D(\mathbf{v}, \mathbf{w})$$

and the Pythagorean property for admissible points

$$D(\mathbf{v}, \mathbf{w}) \geq D(\mathbf{v}, \pi_W(\mathbf{v})) + D(\pi_W(\mathbf{v}), \mathbf{w})$$

give

$$D(\mathbf{v}, \widehat{\pi}_V(\mathbf{w})) + D(\widehat{\pi}_V(\mathbf{w}), \mathbf{w}) \geq D(\mathbf{v}, \pi_W(\mathbf{v})) + D(\pi_W(\mathbf{v}), \mathbf{w}), \tag{1}$$

see Figure 12 for an illustration.

Since $\mathbf{v}$ is a fixed point and hence $\mathbf{v} = \widehat{\pi}_V(\pi_W(\mathbf{v}))$, by the convexity property

$$D(\mathbf{v}, \widehat{\pi}_V(\mathbf{w})) \leq D(\pi_W(\mathbf{v}), \mathbf{w}). \tag{2}$$

Now, (1) and (2) give

$$D(\widehat{\pi}_V(\mathbf{w}), \mathbf{w}) \geq D(\mathbf{v}, \pi_W(\mathbf{v})).$$

Since $\widehat{\pi}_V(\mathbf{w}) \in \Delta(W)$, the above must hold with equality and therefore $\mathbf{v} \in \Delta(W)$. □

The proof above was based on ideas from [2].

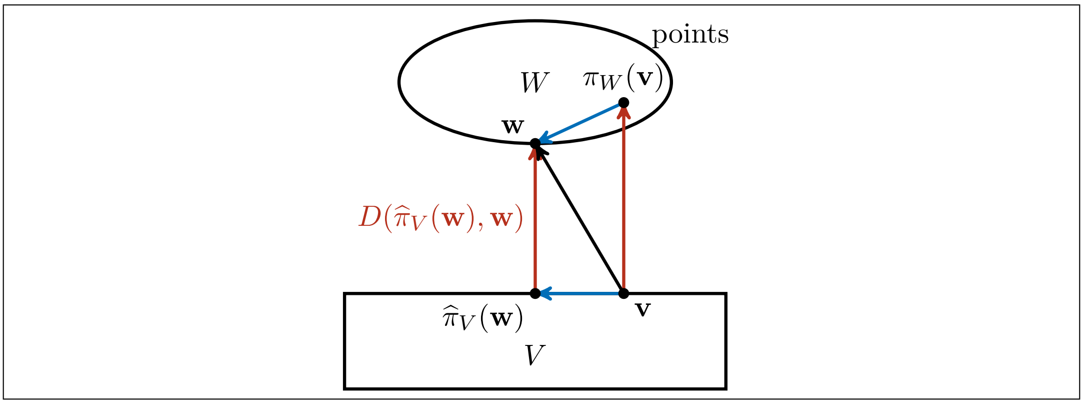

points
W
$\pi_W(\mathbf{v})$
$\mathbf{w}$
$D(\widehat{\pi}_V(\mathbf{w}), \mathbf{w})$
$\widehat{\pi}_V(\mathbf{w})$
$\mathbf{v}$
V

Figure 12: An illustration of the proof for Theorem 1.

Observation 6. *Let D be such that it satisfies the projection and conjugated projection properties, the Pythagorean properties for both admissible and agreeable points, and the convexity property. Let $\mathbf{v}, \mathbf{u} \in \Delta(W) = \Theta(W)$ be otherwise arbitrary. Then*

$$D(\mathbf{v}, \mathbf{u}) = D(\pi_W(\mathbf{v}), \pi_W(\mathbf{u})).$$

Proof. Looking at (1) in the previous proof, which employed identical assumptions, and taking $\mathbf{u} = \widehat{\pi}_V(\mathbf{w})$, we obtain

$$D(\mathbf{v}, \mathbf{u}) + D(\mathbf{u}, \mathbf{w}) \geq D(\mathbf{v}, \pi_W(\mathbf{v})) + D(\pi_W(\mathbf{v}), \mathbf{w}).$$

Since, following the previous theorem, we now know that $\mathbf{u}$ is also a fixed point, we can write $\mathbf{w} = \pi_W(\mathbf{u})$. Furthermore, because both $\mathbf{v}$ and $\mathbf{u}$ are representative points, we have $D(\mathbf{u}, \mathbf{w}) = D(\mathbf{v}, \pi_W(\mathbf{v}))$ and the above becomes

$$D(\mathbf{v}, \mathbf{u}) \geq D(\pi_W(\mathbf{v}), \mathbf{w}).$$

Finally, by the convexity property,

$$D(\mathbf{v},\mathbf{u}) = D(\widehat{\pi}_V(\pi_W(\mathbf{v})), \widehat{\pi}_V(\mathbf{w})) \leq D(\pi_W(\mathbf{v}), \mathbf{w})$$

so the above is possible only with the equality. □

3 Convergence

3.1 Enter metric topology

> "Every reasonable non–pathological space in topology will turn out to be a metric space. On the other hand, developments (. . .) showed there was a need to study a more general class of spaces than merely Euclidean spaces."

Donald W. Kahn, [21]

Thus far, we have avoided the need to introduce any topological structure on the set of all points, but this is going to change in this section. First, let us reintroduce a symbol for the set of points here considered as X. Then, let us equip the set of points X with a metric $d(\mathbf{x},\mathbf{y})$, where $\mathbf{x}$ and $\mathbf{y}$ are any points. Recall that the notion of metric was discussed on Page 10.

We say that a sequence $\{\mathbf{v}_i\}_{i=1}^{\infty}$ of points *converges* to a point $\mathbf{v}$ if for any real number $\epsilon > 0$ there is j such that $d(\mathbf{v}_i, \mathbf{v}) < \epsilon$ for all $i > j$. We call such a $\mathbf{v}$ a *limit point.*

What we need to establish now is a connection between the metric d and the divergence D, which is a mapping from a Cartesian product $X \times X$ to $\mathbb{R}$:

$$D : X \times X \to \mathbb{R}.$$

Therefore, we need to have a metric on the product, say a product metric

$$d_p((\mathbf{x}_1,\mathbf{y}_1),(\mathbf{x}_2,\mathbf{y}_2)) = \Big([d(\mathbf{x}_1,\mathbf{x}_2)]^p + [d(\mathbf{y}_1,\mathbf{y}_2)]^p\Big)^{\frac{1}{p}},$$

where $1 \leq p < \infty$, in place. Then we can define that a mapping $f : X \times X \to \mathbb{R}$ is *continuous*, if for any real number $\epsilon > 0$ there is $\delta > 0$ such that whenever $d_p((\mathbf{x}_1,\mathbf{y}_1),(\mathbf{x}_2,\mathbf{y}_2)) < \delta$ we have $|f(\mathbf{x}_1,\mathbf{y}_1) - f(\mathbf{x}_2,\mathbf{y}_2)| < \epsilon$. The last expression is just the standard metric on $\mathbb{R}$, and our definition follows the usual definition of continuity of a mapping between metric spaces. The connection we were looking for is then the following.

Property 7 (Continuity)**.** *D is continuous.*

Intuitively, the property above says that if two pairs of points are close to each other in the product metric, then D does not rip them apart in $\mathbb{R}$.

The following is a straightforward and intuitive consequence of D being continuous, and it is how we will employ continuity to obtain future results.

Observation 7. *Let D satisfy the continuity property. Assume that a sequence of agreeable points $\{\mathbf{v}_i\}_{i=1}^{\infty}$ converges to $\mathbf{v}$ and a sequence of admissible points $\{\mathbf{w}_i\}_{i=1}^{\infty}$ converges to $\mathbf{w}$. Then the sequence*

$$\{D(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^{\infty}$$

converges to $D(\mathbf{v}, \mathbf{w})$.

Proof. For any $\epsilon > 0$ we are tasked with finding j such that $|D(\mathbf{v}_i, \mathbf{w}_i) - D(\mathbf{v}, \mathbf{w})| < \epsilon$ for all $i > j$. Since D is continuous, for any $\epsilon > 0$ there is $\delta > 0$ such that whenever

$$\Big([d(\mathbf{v}_i, \mathbf{v})]^p + [d(\mathbf{w}_i, \mathbf{w})]^p\Big)^{\frac{1}{p}} < \delta$$

we have $|D(\mathbf{v}_i, \mathbf{w}_i) - D(\mathbf{v}, \mathbf{w})| < \epsilon$. Now we simply select j so that $[d(\mathbf{v}_i, \mathbf{v})]^p + [d(\mathbf{w}_i, \mathbf{w})]^p < \delta^p$ for all $i > j$. This is always possible since $\{\mathbf{v}_i\}_{i=1}^{\infty}$ converges to $\mathbf{v}$ and $\{\mathbf{w}_i\}_{i=1}^{\infty}$ converges to $\mathbf{w}$. □

Since we operate in a metric space, we can define that a subset of points X is *compact* if every sequence that can be constructed from its elements has a convergent subsequence and the limit point of this convergent subsequence lies in this subset. In other words, it has the Bolzano–Weierstrass property, which in metric spaces is equivalent to compactness.

Observation 8. *If V and W are compact, and D satisfies the continuity property, then a representative point exists.*

Proof. Consider the set of all real numbers $D(\mathbf{v}, \mathbf{w})$ such that $\mathbf{v} \in V$ and $\mathbf{w} \in W$. This set is bounded from below, so it also has the greatest lower bound (a basic property of real numbers). Let us denote it b.

Now, for every $\epsilon > 0$ there are $\mathbf{v} \in V$ and $\mathbf{w} \in W$ such that

$$\epsilon + b > D(\mathbf{v}, \mathbf{w}) \geq b,$$

otherwise b would not be the greatest lower bound. Therefore, we can construct a sequence $\{D(\mathbf{v}_i, \mathbf{w}_i)\}_{i=1}^{\infty}$ that converges to b. Now, due to the compactness of V

the sequence $\{\mathbf{v}_i\}_{i=1}^{\infty}$ has a convergent subsequence, say $\{\mathbf{v}_{i_j}\}_{j=1}^{\infty}$. Let $\{\mathbf{w}_{i_j}\}_{j=1}^{\infty}$ be the corresponding sequence in W, which is also compact, so it also has a convergent subsequence. Let $\mathbf{w} \in W$ be its limit point and let $\mathbf{v} \in V$ be the limit point of $\{\mathbf{v}_{i_j}\}_{j=1}^{\infty}$. Then, due to Observation 7

$$D(\mathbf{v}, \mathbf{w}) = b,$$

so $\mathbf{v}$ must be a representative point. □

Looking now at the statement of Theorem 1 on Page 21 we can replace the requirement for the existence of a representative point by requiring compactness of V and W, and asking D to satisfy the continuity property.

3.2 Alternating minimisation procedure

The following property will be needed to prove that a representative point can be reached by an iterative process.

Property 8 (Four Points)**.** *Let* $\mathbf{w}, \mathbf{u} \in W$ *and* $\mathbf{v} \in V$*. Then*

$$D(\widehat{\pi}_V(\mathbf{w}), \mathbf{u}) \leq D(\mathbf{w}, \mathbf{u}) + D(\mathbf{v}, \mathbf{u}).$$

The four point property is illustrated in Figure 13. The name comes from Csiszár and Tusnády, who gave a motivating example in [13, Page 213].

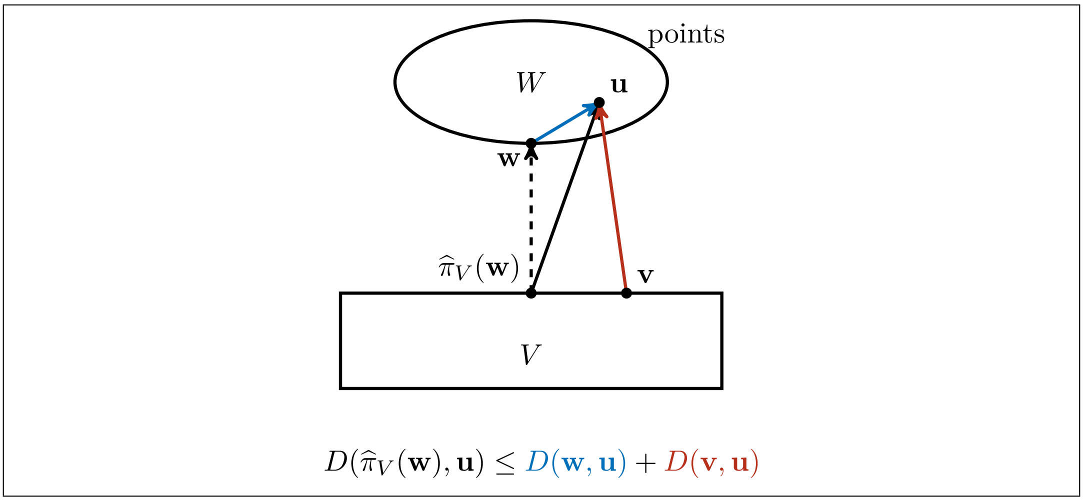

points
W
$\mathbf{u}$
$\mathbf{w}$
$\widehat{\pi}_V(\mathbf{w})$
$\mathbf{v}$
V
$D(\widehat{\pi}_V(\mathbf{w}), \mathbf{u}) \leq D(\mathbf{w}, \mathbf{u}) + D(\mathbf{v}, \mathbf{u})$

Figure 13: An illustration of the four point property.

Thus far, we had one property that linked the concept of divergence D and the metric topology given by d: it was the continuity property. Here we provide another one, which somewhat goes in the opposite direction. This connection is important as convergence results must be established using the metric; a divergence, in general, does not even generate a topology [12].

Property 9 (Connectivity). *If $\{D(\mathbf{v}_i,\mathbf{v})\}_{i=1}^{\infty}$ converges to zero then so does*

$$\{d(\mathbf{v}_i,\mathbf{v})\}_{i=1}^{\infty}.$$

If D is continuous, then $\{d(\mathbf{v}_i,\mathbf{v})\}_{i=1}^{\infty} \to 0$ implies $\{D(\mathbf{v}_i,\mathbf{v})\}_{i=1}^{\infty} \to 0$. The property above gives the opposite implication. It also implies that if $\{D(\mathbf{v}_i,\mathbf{v})\}_{i=1}^{\infty}$ converges to zero then $\mathbf{v}$ is the limit point of $\{\mathbf{v}_i\}_{i=1}^{\infty}$. We will use this in the following theorem.

Theorem 2 (Convergence). *Let D be such that it satisfies the projection and conjugated projection properties, the Pythagorean properties for both admissible and agreeable points, and the consistency, convexity, continuity, four point and connectivity properties. Let $\mathbf{v}_0 \in V$. Define a sequence $\{\mathbf{v}_i\}_{i=0}^{\infty}$ recursively by $\mathbf{v}_{i+1} = \widehat{\pi}_V(\pi_W(\mathbf{v}_i))$. If V and W are compact then the sequence $\{\mathbf{v}_i\}_{i=0}^{\infty}$ converges to a fixed point.*

Proof. First, notice that by the projection and conjugated projection properties

$$D(\mathbf{v}_i,\pi_W(\mathbf{v}_i)) \geq D(\widehat{\pi}_V(\pi_W(\mathbf{v}_i)),\pi_W(\mathbf{v}_i)) \geq$$

$$\geq D(\widehat{\pi}_V(\pi_W(\mathbf{v}_i)),\pi_W(\widehat{\pi}_V(\pi_W(\mathbf{v}_i)))),$$

so the sequence of non–negative real numbers $D(\mathbf{v}_i,\pi_W(\mathbf{v}_i))_{i=0}^{\infty}$ converges and its limit point exists (the closed interval $[0,D(\mathbf{v}_0,\pi_W(\mathbf{v}_0))]$ is compact in $\mathbb{R}$ equipped with the standard metric). We will denote this limit information divergence λ.

Furthermore, due to the compactness of V and W the sequences $\{\mathbf{v}_i\}_{i=0}^{\infty}$ and $\{\pi_W(\mathbf{v}_i)\}_{i=0}^{\infty}$ have both a convergent subsequence with a corresponding limit point, we denote these limit points $\mathbf{v} \in V$ and $\mathbf{w} \in W$ respectively. Therefore, by Observation 7 on Page 24,

$$D(\mathbf{v},\mathbf{w}) = \lambda.$$

What we need to prove at this stage is that the whole sequence $\{\mathbf{v}_i\}_{i=0}^{\infty}$, not just its subsequence, converges to $\mathbf{v}$. We will do this considering Figure 14.

By the four point property

$$D(\mathbf{v}_i,\mathbf{w}) \leq D(\pi_W(\mathbf{v}_{i-1}),\mathbf{w}) + D(\mathbf{v},\mathbf{w}).$$

and by the Pythagorean property for admissible points

$$D(\mathbf{v}_i, \pi_W(\mathbf{v}_i)) + D(\pi_W(\mathbf{v}_i), \mathbf{w}) \leq D(\mathbf{v}_i, \mathbf{w}).$$

Since

$$D(\mathbf{v}_i, \pi_W(\mathbf{v}_i)) \geq D(\mathbf{v}, \mathbf{w})$$

it follows that

$$D(\pi_W(\mathbf{v}_i), \mathbf{w}) \leq D(\pi_W(\mathbf{v}_{i-1}), \mathbf{w}).$$

However, we already know that a subsequence of $\{\pi_W(\mathbf{v}_i)\}_{i=0}^{\infty}$ converges to $\mathbf{w}$, so this means that $\{D(\pi_W(\mathbf{v}_i), \mathbf{w})\}_{i=0}^{\infty}$ converges, by Observation 7, to $D(\mathbf{w}, \mathbf{w})$, which is by the consistency property 0. Finally, using the connectivity property, the whole sequence $\{\pi_W(\mathbf{v}_i)\}_{i=0}^{\infty}$ must converge to $\mathbf{w}$.

By the convexity property $D(\pi_W(\mathbf{v}_i), \mathbf{w}) \geq D(\widehat{\pi}_V(\pi_W(\mathbf{v}_i)), \mathbf{v})$ for all i so also

$$\{D(\mathbf{v}_i, \mathbf{v})\}_{i=1}^{\infty}$$

converges to zero which in turn means, making the same argument as above, that $\{\mathbf{v}_i\}_{i=0}^{\infty}$ converges to $\mathbf{v}$ as desired.

However, in order to apply the convexity property above, we need first to establish that $\widehat{\pi}_V(\mathbf{w}) = \mathbf{v}$. For a contradiction, let us assume that $\mathbf{v} \neq \widehat{\pi}_V(\mathbf{w})$. By the Pythagorean property for agreeable points

$$D(\widehat{\pi}_V(\mathbf{w}), \widehat{\pi}_V(\mathbf{w}_i)) + D(\widehat{\pi}_V(\mathbf{w}_i), \mathbf{w}_i) = D(\widehat{\pi}_V(\mathbf{w}), \mathbf{w}_i),$$

for all i. Since $\{\mathbf{w}_i\}_{i=1}^{\infty}$ converges to $\mathbf{w}$, and $\{\mathbf{v}_i\}_{i=1}^{\infty}$ has a subsequence converging to $\mathbf{v}$ (so we focus only on it), and by Observation 7, we can also write

$$D(\widehat{\pi}_V(\mathbf{w}), \mathbf{v}) + D(\mathbf{v}, \mathbf{w}) = D(\widehat{\pi}_V(\mathbf{w}), \mathbf{w}).$$

By the assumption and the consistency property $D(\widehat{\pi}_V(\mathbf{w}), \mathbf{v}) > 0$, so we have that $D(\mathbf{v}, \mathbf{w}) < D(\widehat{\pi}_V(\mathbf{w}), \mathbf{w})$. But this is not possible, a contradiction.

Finally, we need also to establish that $\pi_W(\mathbf{v}) = \mathbf{w}$, which will give us together with the above that $\mathbf{v}$ is a fixed point. For a contradiction, let us assume that $\mathbf{w} \neq \pi_W(\mathbf{v})$. By the Pythagorean property for admissible points

$$D(\mathbf{v}_i, \pi_W(\mathbf{v}_i)) + D(\pi_W(\mathbf{v}_i), \pi_W(\mathbf{v})) \leq D(\mathbf{v}_i, \pi_W(\mathbf{v})),$$

for all i. Since $\{\mathbf{v}_i\}_{i=1}^{\infty}$ converges to $\mathbf{v}$ and $\{\pi_W(\mathbf{v}_i)\}_{i=1}^{\infty}$ converges to $\mathbf{w}$, and by Observation 7, we can also write

$$D(\mathbf{v}, \mathbf{w}) + D(\mathbf{w}, \pi_W(\mathbf{v})) \leq D(\mathbf{v}, \pi_W(\mathbf{v})).$$

By the assumption and the consistency property $D(\mathbf{w}, \pi_W(\mathbf{v})) > 0$, so we have that $D(\mathbf{v}, \mathbf{w}) < D(\mathbf{v}, \pi_W(\mathbf{v}))$. But this is not possible, a contradiction.

Therefore $\mathbf{v} = \widehat{\pi}_V(\pi_W(\mathbf{v}))$ and $\mathbf{v}$ is a fixed point. □

Finally, considering Theorem 1 on Page 21 and Observation 8 on Page 24, we may claim that the fixed point from the theorem above is also a representative point.

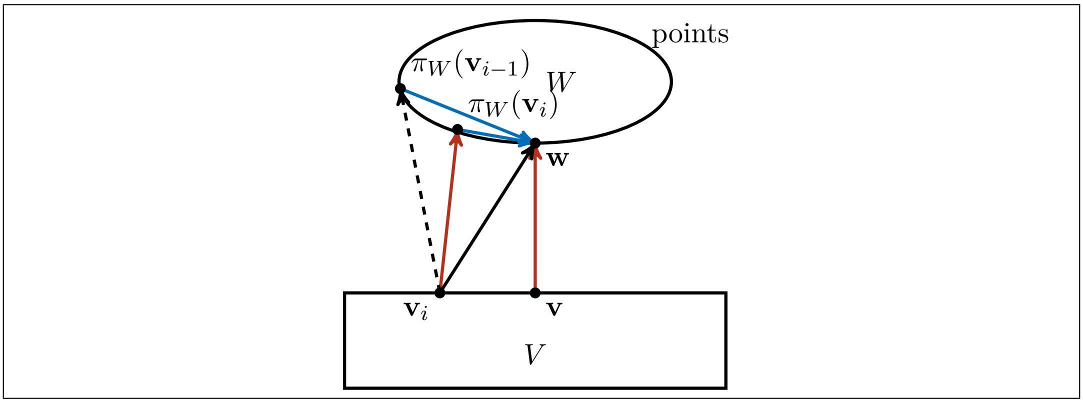


Figure 14: An illustration of the proof of Theorem 2. The arrows here indicate the direction of divergences as defined earlier, and they should not be interpreted as information propagation.

The algorithm (and in fact the idea of the proof presented above) is due to Csiszár and Tusnády [13], who developed it for a particular information divergence and it is known as an *alternating minimisation procedure*. The algorithm was then generalised many times in the literature, see e.g. [8], and the version above can be considered as another variant. Nevertheless, it is still the same idea developed before 1984 in many attempts, perhaps the first successful being [12].

3.3 Remarks

We have now achieved the goal as initially stated: we have introduced information geometry without actually specifying the exact nature of admissible and agreeable points we worked with. However, the paper is far from finished. First, we will formally establish the geometry in the Euclidean setting and mention a real practical application of such geometry.

Second, as this aspired to cover also non–Euclidean generalisation of information geometry, we will find a non–trivial and different formalisation of the intuitive concept. This is exciting as we recover information geometry on mathematical objects that are not as deeply connected with geometry as Euclidean space.

4 Euclidean space setting

4.1 Points

"One could not see the forest for the trees."

A Common Proverb

In this paper, we have accumulated a large number of properties that we require from an information divergence D and from the sets of agreeable and admissible points. Naturally, we should ask the following question: Is it actually possible to satisfy them all? In this section, we show particular examples that satisfy all the properties, but we will need some additional notions to define them.

We start with the J–dimensional *Euclidean space*, which is a set of all ordered J–tuples

$$\mathbf{v} = (v_1, \dots, v_J),$$

where every v_j is a real number. In other words, $\mathbf{v} \in \mathbb{R}^J$. A $(J-1)$–dimensional *probabilistic simplex* $\mathbb{D}^J$, $J \geq 2$, is a subspace of the J–dimensional Euclidean space defined as those $\mathbf{v} \in \mathbb{R}^J$ that satisfy

$$\sum_{j=1}^{J} v_j = 1.$$

We will confine ourselves to the case when $v_j > 0$, for all $1 \leq j \leq J$, to avoid any pathological cases, which makes $\mathbb{D}^J$ an open set. Such a defined *discrete probability distribution* $\mathbf{v}$ could perhaps represent a probabilistic opinion that an individual may have about the world, and thus it plays a central role in inductive logic, uncertain reasoning and belief merging [24, 30]. More recently, in [4] they have been used to represent the results of individual medical studies.

We say that a subset W of points in $\mathbb{R}^I$ is *convex* if for any two $\mathbf{v}, \mathbf{w} \in W$ we have that also

$$(\lambda \cdot v_1 + (1-\lambda) \cdot w_1, \dots, \lambda \cdot v_I + (1-\lambda) \cdot w_I) \in W,$$

for all $\lambda \in [0, 1]$. We say that a subset W of points in $\mathbb{R}^I$ is *closed* if the limit point of every convergent sequence constructed from the elements of W has its limit inside W, with respect to the standard Euclidean metric.

Now, let us consider a closed convex set of points

$$W \subseteq \underbrace{\mathbb{D}^J \times \dots \times \mathbb{D}^J}_{n}.$$

Note that $I = Jn$ (in the definition of convexity above) and $\mathbf{w}(i) \in W$ is of the form $\mathbf{w}(i) = (\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)})$, where each $\mathbf{w}^{(i)} \in \mathbb{D}^J$, $1 \leq i \leq n$, is a probability distribution admissible by the member i of a group of n individuals. It might be helpful to think of $\mathbf{w}(i)$ as a discrete function $\{1, \dots, n\} \to \mathbb{D}^J$. The set W will be an example of a set of admissible points discussed earlier in the paper.

Finally, let

$$V \subseteq \underbrace{\mathbb{D}^J \times \ldots \times \mathbb{D}^J}_{n}$$

be such that in each $\mathbf{v} \in V$ all members are in agreement: $\mathbf{v} = (\mathbf{v}^{(1)}, \dots, \mathbf{v}^{(n)})$, where $\mathbf{v}^{(1)} = \ldots = \mathbf{v}^{(n)}$. With some abuse of notation, we will often write $\mathbf{v}$ in place of every $\mathbf{v}^{(i)}$. This set V is not closed yet (because $\mathbb{D}^J$ is not), but we will fix a sufficiently small $\epsilon > 0$ and ask every $v_j > \epsilon$, $1 \leq j \leq J$. A suitable ϵ exists (in the sense that $W \subseteq V$ must be possible), since W is assumed closed. Such a closed set V will be an example of a set of agreeable points discussed earlier in the paper.

Clearly, it could be that there are some agreeable points in W (when $\mathbf{w}^{(1)} = \ldots = \mathbf{w}^{(n)}$ for some $\mathbf{w}(i)$), but V and W could be disjoint as well. In any case, W is assumed non–empty, while V is non–empty by definition. Both W and V are defined as closed and bounded, and hence in this Euclidean setting they are both compact. Note that compactness was required in Observation 8 on Page 24, and Theorem 2 on Page 26.

4.2 Divergences

After we have introduced the points, let us now define a divergence from one point to another. In [5], the following divergence from $\mathbf{v} \in V$ to $\mathbf{w}(i) \in W$ based on the Rényi entropy [26] was defined:

$$D_r(\mathbf{v}, \mathbf{w}(i)) = \sum_{i=1}^{n} \lambda_i \sum_{j=1}^{J} [(w_j^{(i)})^r - (v_j)^r - r(w_j^{(i)} - v_j)(v_j)^{r-1}],$$

where $2 \geq r > 1$; and $\sum_{i=1}^{n} \lambda_i = 1$ represents fixed positive weights that we wish to assign to different agents or sources of information.

For $r = 2$ this divergence becomes a weighted sum of the well known *squared Euclidean distances*

$$D_2(\mathbf{v}, \mathbf{w}(i)) = \sum_{i=1}^{n} \lambda_i \sum_{j=1}^{J} (v_j - w_j^{(i)})^2,$$

exceptionally a symmetric divergence. If $\mathbf{w}(i)$ is an agreeable point, in this case we simply write $\mathbf{w} = \mathbf{w}^{(1)} = \ldots = \mathbf{w}^{(n)}$, the above actually becomes the squared

Euclidean distance:

$$D_2(\mathbf{v}, \mathbf{w}(i)) = D_2(\mathbf{v}, \mathbf{w}) = \sum_{j=1}^{J} (v_j - w_j)^2.$$

Additionally, we should note that in this case

$$d(\mathbf{v}, \mathbf{w}) = \sqrt{D_2(\mathbf{v}, \mathbf{w})}$$

is the standard Euclidean metric. The proof that the set of representative points $\Delta^{D_r}(W)$ based on the Rényi entropy is well defined is in [1].

Another way to define the divergence D from $\mathbf{v} \in V$ to $\mathbf{w}(i) \in W$ is to take the *Kullback–Leibler divergence* (also known as cross–entropy)

$$\mathrm{KL}(\mathbf{v}, \mathbf{w}(i)) = \sum_{i=1}^{n} \lambda_i \sum_{j=1}^{J} w_j^{(i)} \log \frac{w_j^{(i)}}{v_j}.$$

Again, this becomes the usual Kullback–Leibler divergence if all components of $\mathbf{w}(i)$ agree, in which case we write $\mathbf{w}$ in place of $\mathbf{w}(i)$ and

$$\mathrm{KL}(\mathbf{v}, \mathbf{w}) = \sum_{j=1}^{J} w_j \log \frac{w_j}{v_j}.$$

In literature, it is common to write the arguments the other way around as $\mathrm{KL}(\mathbf{w} \| \mathbf{v})$, but we wanted to stick here with the more intuitive notation adopted earlier in this paper. Also, our choice of V avoids the usual headache of defining the above when $\mathbf{v}$ is permitted to be zero in some coordinates.

A limit theorem relating the set of representative points $\Delta^{D_r}(W)$ based on the Rényi entropy to the set of representative points $\Delta^{\mathrm{KL}}(W)$ based on the Kullback–Leibler divergence has been proven in [5]:

$$\emptyset \neq \lim_{r \searrow 1} \Delta^{D_r}(W) \subseteq \Delta^{\mathrm{KL}}(W).$$

Whether or not the above holds with equality is an open problem.

The proofs that the divergences defined above satisfy all Properties 1 to 9 discussed in this paper are scattered in [1] and [2], and they are all special cases of general convex Bregman divergences [9]. As those represent another increase in difficulty, we will mention them in more detail later in Section 5.2.

What we discussed in this paper now gives us

$$\Delta^{D_r}(W) = \Theta^{D_r}(W), \text{ and } \Delta^{\mathrm{KL}}(W) = \Theta^{\mathrm{KL}}(W),$$

the representative and fixed points are the same points, and we can get a representative point by iterating projections and conjugated projections; in other words, using the alternating minimisation procedure.

4.3 Applications

Regardless of which of the above mentioned divergences is taken for D, the conjugated D–projection of an admissible point $\mathbf{w}(i) = (\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)}) \in W$ to the set of agreeable points $\underbrace{(\mathbf{w}, \dots, \mathbf{w})}_{n} \in V$ in fact gives

$$\mathbf{w} = \sum_{i=1}^{n} \lambda_i \mathbf{w}^{(i)},$$

where $\sum_{i=1}^{n} \lambda_i = 1$ is a fixed positive weighting, and

$$\mathbf{w} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{w}^{(i)}$$

if the weights are equal. Recall that, with some abuse of notation, we write $\mathbf{w}$ to denote both the component computed by the formulas above and the whole n–tuple in the set V. This will allow us to greatly simplify the notation of the next section, making everything more intuitive.

For example, for a given $\mathbf{w}(i)$,

$$\arg\min_{\mathbf{x} \in V} D_2(\mathbf{x}, \mathbf{w}(i)) = \arg\min_{\mathbf{x} \in V} \sum_{i=1}^{n} \lambda_i \sum_{j=1}^{J} (x_j - w_j^{(i)})^2,$$

gives after differentiation $2 \sum_{i=1}^{n} \lambda_i (x_j - w_j^{(i)}) = 0$, which is equivalent to $x_j = \sum_{i=1}^{n} \lambda_i w_j^{(i)}$, for all $1 \leq j \leq J$. The claim can now be established using the fact that the function is strictly convex. A proof for other divergences can be derived by proving a Euclidean analogy to Observation 9 on Page 41, and this Euclidean analogy can also be found in [2, Page 6343].

In short, we have obtained above the ordinary (weighted) arithmetic mean applied to J coordinates respectively. This result, as the expected value, has also been established in terms of random variables by Banerjee, Guo and Wang [7] in 2005. In the literature, this arithmetic mean operator is known as the *linear pooling operator*, see [16]. It is a common choice of representing different opinions $\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)} \in \mathbb{D}^J$ of n individuals as a single point in $\mathbb{D}^J$, a natural agreeable point. We should note that there are also other pooling operators, and even arguments against using the linear pooling operator in uncertain reasoning and belief merging, see [30] for one based on the so–called *locality principle*.

Now, define the prior agreeable point

$$\mathbf{o} = (\underbrace{\mathbf{v}, \dots, \mathbf{v}}_{n}) \in V$$

using the *uniform probability distribution*

$$\mathbf{v} = \left(\frac{1}{J}, \dots, \frac{1}{J}\right) \in \mathbb{D}^J.$$

Then $\mathbf{ME}_{\mathrm{KL}}(W)$, discussed in Section 2.4 on Page 15 as the KL–projection of $\mathbf{o}$, is the usual most entropic point in W. For every i, it is defined as that $\mathbf{w}^{(i)}$ that maximises the Shannon entropy

$$-\sum_{j=1}^{J} w_j^{(i)} \log w_j^{(i)}.$$

An obdurate committee would then take this most entropic point and find the conjugated KL–projection in the set of agreeable points V, which we now know to be equivalent to applying the linear pooling operator, and be content with it. We suggest that a rational committee would iterate the whole process endlessly until a representative point in $\Delta^{\mathrm{KL}}(W) = \Theta^{\mathrm{KL}}(W) \subseteq V$ is reached, which will happen by Theorem 2 on Page 26.

Exactly this procedure, starting with the prior agreeable point and using the Kullback–Leibler divergence, was applied in [4] to combine several medical studies investigating the one–year incidence in the diagnosis of cancer in patients with unprovoked venous thromboembolism. This condition is a formation of a blood clot in a vein or lungs without an apparent reason, and we are interested in the proportion of patients that are subsequently diagnosed with cancer and how we should detect it. The particular studies investigated in [4] were presenting some linear constraints on how patients are distributed over several mutually exhaustive and exclusive categories made of different screening outcomes and cancer diagnosis results. While these constraints across all studies were jointly conflicting, each individual study gave rise to a non–empty closed convex set of discrete probability distributions. The Cartesian product of these sets defines an example of a set of admissible points W.

The weights λ_i, $1 \leq i \leq n$, associated with each study (as a source of information) were given by the corresponding study sample sizes:

$$\lambda_i = \frac{(\text{study } i \text{ sample size})}{(\text{pooled sample size})}.$$

Should the constraints given by each study provided the full information and resulted in a single probability distribution $\mathbf{w}^{(i)}$, then the result of the above procedure would simply be the weighted arithmetic mean of the individual probability distributions

$$\mathbf{w} = \sum_{i=1}^{n} \lambda_i \mathbf{w}^{(i)},$$

as there would be only one point $\mathbf{w}(i) = (\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)})$ in the set W. However, that was not the case and the alternating minimisation procedure was iterated a thousand times to approach a point in $\Delta^{\mathrm{KL}}(W)$. A combinatorial argument in favour of using $\Delta^{\mathrm{KL}}(W)$ in a context of merging heterogeneous studies (i.e., they are reporting results that are more different than statistically expected), where this heterogeneity is unexplained, and with large sample sizes was presented in [3]. Reference [4] argues that these assumptions on heterogeneity and sample sizes were in this case satisfied, and it thus presents a real and seemingly justified application of information geometry and of the alternating minimisation procedure introduced to the reader in this paper.

It is important to note that if the assumption of unexplained heterogeneity is not satisfied, the method is not justified. If we know what is causing the studies to be more different than statistically expected then we need to use this information (this is the same requirement as when the maximum entropy principle is applied). In contrast, in the described application, we do not judge the quality of information coming from different sources other than by the provided sample size: the higher the relative sample size, the higher the weight.

On the other hand, should there be only one agent (or one source of information) $n = 1$, then $W = \Delta^{\mathrm{KL}}(W) = \Theta^{\mathrm{KL}}(W) \subseteq V$, so there would be no need to iterate the process as $\mathbf{ME}_{\mathrm{KL}}(W)$ would be trivially, see Observation 4 on Page 17, a fixed point. This would correspond to the classical most entropic solution when there are no conflicting sources of information.

5 Hilbertian space setting

5.1 $\mathbf{L^2}$ space

In the previous section, we worked with the Euclidean space equipped with the *inner product* $\langle \mathbf{v}, \mathbf{w} \rangle = \sum_{j=1}^{J} v_j w_j$. A natural generalisation of the Euclidean space is a *Hilbert space* equipped with a real *inner product* $\langle \cdot, \cdot \rangle : H \times H \to \mathbb{R}$. Csiszár and Tusnády suggested this setting for the alternating minimisation procedure already in their 1984 paper [13].

In the following, we will be working in the L^2 space, which is a non–Euclidean Hilbert space. A reader unfamiliar with function spaces can have a look at [15], or perhaps [11], for an introduction. Elements of this space are Lebesgue measurable functions

$$\mathbf{w} : [0,1] \to \mathbb{R},$$

where

$$\langle \mathbf{v}, \mathbf{w} \rangle = \int_0^1 \mathbf{v}\mathbf{w} d\mu$$

is a Lebesgue integral and μ is a Lebesgue measure, and we will require

$$\int_0^1 \mathbf{w} d\mu = 1 \tag{3}$$

on top of the usual condition that

$$\sqrt{\int_0^1 (\mathbf{w})^2 d\mu} < \infty.$$

Again, a reader unfamiliar with the Lebesgue measure could consider reading [23] first.

The idea here is to move from discrete probability distributions of the Euclidean setting to *continuous probability distributions*, that is the reasoning behind (3). As in the previous section, we will try to avoid pathological cases by requiring $\mathbf{w}$ to be non–zero in the domain, with the exception of a set with Lebesgue measure zero.

Now, considering a discrete number of n individuals seeking agreement as in the Euclidean setting, let us denote their respective functions $\mathbf{w}^{(i)}$, $1 \leq i \leq n$. An exciting line of research could be instead changing the function $\mathbf{w}(t)$ continuously with time $t \in [0,1]$ and using the mean value function as the agreement, which would invite us to think of a single agent changing its mind about the continuous probability distribution $\mathbf{w}$. Although much of what follows would be similar in this set up, there are some complications that we are not prepared to deal with and thus we will not discuss this idea further.

We say that a subset W_i of functions in L^2 is *convex* if for any two $\mathbf{v}, \mathbf{w} \in W_i$ we have that also

$$\lambda \cdot \mathbf{v} + (1-\lambda) \cdot \mathbf{w} \in W_i,$$

for all $\lambda \in [0,1]$. We say that a subset W_i of points in L^2 is *closed* if the limit point of every convergent sequence constructed from the elements of W_i has its limit inside W_i, with respect to the standard L^2 metric.

We take a Cartesian product W of closed convex sets W_i of bounded L^2 functions $\mathbf{w}^{(i)}$, satisfying (3) and non–zero in the domain (with the exception of a set of

Lebesgue measure zero) for every $1 \leq i \leq n$, as our set of admissible points. This is consistent with what we did in the Euclidean setting (Section 4). Each member of the set W will be denoted by $\mathbf{w}(i) = (\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)})$, but if all the components agree we will simply write $\mathbf{w} = (\mathbf{w}, \dots, \mathbf{w})$. For a small price of accepting this confusion, we will be shortly rewarded with more intuitive formulas.

This set W is not necessarily compact, but compactness was needed in our key result: Theorem 2 on Page 26. For compactness, and since a Hilbert space is a complete metric space, we need the union of all W_i to be *totally bounded*, see [11]. A necessary and sufficient condition for an L^p space to be totally bounded is given by the Kolmogorov—Riesz theorem, see [17]. Informally, we need every W_i to be bounded and a small change in the argument of its functions should make a uniformly small change in the function values (across all functions in W_i, $1 \leq i \leq n$): For all $\epsilon > 0$ there is $\delta > 0$ such that for all $\mathbf{w} \in W_i$, all $1 \leq i \leq n$ and all $|h| < \delta$ we have

$$\int_0^1 [\mathbf{w}(x+h) - \mathbf{w}(x)]^2 d\mu < \epsilon. \tag{4}$$

We could get an agreement from $\mathbf{w}(i) = (\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)}) \in W$ say as its weighted arithmetic mean

$$\mathbf{w} = \sum_{i=1}^{n} \lambda_i \mathbf{w}^{(i)},$$

where $\sum_{i=1}^{n} \lambda_i = 1$ is a fixed positive weighting (a motivation was outlined in Section 4.3: we typically assign weights to different medical studies proportionally to their sample sizes), and then with some abuse of notation write $\mathbf{w} = (\mathbf{w}, \dots, \mathbf{w})$ to make the agreement the same kind of object as the admissible points are.

Clearly, the resulting function satisfies (3), because $\mathbf{w}^{(i)}$ satisfies (3) for every $1 \leq i \leq n$. It is also non–zero in the domain with the exception of a set with Lebesgue measure zero. This is because $\mathbf{w}^{(i)} \geq 0$, $1 \leq i \leq n$, so whenever $\sum_{i=1}^{n} \lambda_i \mathbf{w}^{(i)}$ is zero for some x it must be that all $\mathbf{w}^{(i)}$ are zero at that x. If this happens on a set with non–zero Lebesgue measure, then all $\mathbf{w}^{(i)}$ would have the same property: a contradiction. Additionally, since all $\mathbf{w}^{(i)}$, $1 \leq i \leq n$, are bounded, the weighted arithmetic mean is also bounded. For a given W, we indeed define the set of agreeable points V as the corresponding set of weighted arithmetic means.

Finally, if W is compact, then V is compact as well: We need to prove that for all $\epsilon > 0$ there is $\delta > 0$ such that for all $\sum_{i=1}^{n} \lambda_i \mathbf{w}^{(i)} \in V$ and $|h| < \delta$ we have

$$\int_0^1 \Big[\sum_{i=1}^{n} \lambda_i \mathbf{w}^{(i)}(x+h) - \sum_{i=1}^{n} \lambda_i \mathbf{w}^{(i)}(x) \Big]^2 d\mu < \epsilon.$$

We can find this δ as the one assumed to exist in (4), since that one exists uniformly for all $1 \leq i \leq n$: The quadratic function is a convex function so using Jensen's inequality for integrals [19] (which is valid only for probability measures: $\int_0^1 d\mu = 1$) together with (4) we obtain

$$\epsilon = \sum_{i=1}^{n} \lambda_i \epsilon \geq \sum_{i=1}^{n} \lambda_i \Big(\int_0^1 [\mathbf{w}^{(i)}(x+h) - \mathbf{w}^{(i)}(x)]^2 d\mu \Big) \geq$$

$$\geq \int_0^1 \Big[\sum_{i=1}^{n} \lambda_i \mathbf{w}^{(i)}(x+h) - \sum_{i=1}^{n} \lambda_i \mathbf{w}^{(i)}(x) \Big]^2 d\mu$$

as required.

5.2 Bregman divergences

> "A functional Bregman divergence acts on functions or distributions, and generalizes the standard Bregman divergence for vectors and a previous pointwise Bregman divergence that was defined for functions. A recent result showed that the mean minimizes the expected Bregman divergence. The new functional definition enables the extension of this result to the continuous case to show that the mean minimizes the expected functional Bregman divergence over a set of functions or distributions."
>
> Béla A. Frigyik, Santosh Srivastava and Maya R. Gupta, [14]

Recall that a functional $f : L^2 \to \mathbb{R}$ is *Fréchet differentiable* [15] if there exists a bounded linear operator $A : L^2 \to \mathbb{R}$ such that

$$\lim_{\|\mathbf{h}\| \to 0} \frac{|f(\mathbf{v}+\mathbf{h}) - f(\mathbf{v}) - A(\mathbf{h})|}{\|\mathbf{h}\|} = 0,$$

where $\| \cdot \|$ is the norm in the L^2 space.

Every Fréchet differentiable functional has its *Gâteaux differential* (but not the other way around):

$$\delta f(\mathbf{v}; \mathbf{h}) = \lim_{h \to 0} \frac{f(\mathbf{v} + h\mathbf{h}) - f(\mathbf{v})}{h}. \tag{5}$$

The differential is at $\mathbf{v}$ in the direction of $\mathbf{h}$. Since f is Fréchet differentiable, the functional $\delta f(\mathbf{v}; \cdot) : L^2 \to \mathbb{R}$ is equal to the Fréchet derivative and thus it is a bounded linear functional. By the Riesz representation theorem [6] we can then represent it by an inner product for a given $\nabla f(\mathbf{v}) \in L^2$:

$$\delta f(\mathbf{v}; \cdot) = \langle \cdot, \nabla f(\mathbf{v}) \rangle.$$

With some abuse of notation $\nabla f(\mathbf{v})$ is often called the Fréchet derivative of f at $\mathbf{v}$.

Now we define a rather general divergence based on the *Bregman divergence* [9]:

$$D_f(\mathbf{v}, \mathbf{w}(i)) = \sum_{i=1}^{n} \lambda_i \int_0^1 f(\mathbf{w}^{(i)}) - f(\mathbf{v}) - \langle \mathbf{w}^{(i)} - \mathbf{v}, \nabla f(\mathbf{v}) \rangle d\mu,$$

where $f : L^2 \to \mathbb{R}$ is a strictly convex, continuous and continuously Fréchet differentiable functional and $\sum_{i=1}^{n} \lambda_i = 1$ is a fixed positive weighting. In literature, it is common to write the arguments the other way around as $D_f(\mathbf{w}(i) \| \mathbf{v})$, but we wanted to stick here with the more intuitive notation adopted earlier in this paper. We should also note that $\mathbf{v}$ as well as $\mathbf{w}(i)$ is considered a Cartesian product, but whenever all its components agree we will drop the index. Similarly, if in addition all components of $\mathbf{w}(i)$ agree, we may even drop the sum entirely.

Looking at existing literature, the geometry of Bregman divergences and convergence conditions for alternating minimisation procedures in Hilbert spaces have already been developed by Burachik and Iusem [10] in 1998. Bregman divergences over L^p function spaces, among them L^2 is the only Hilbert space, were introduced by Frigyik, Srivastava and Gupta [14] in 2008. We are confined here to the L^2 space setting, which lies in the intersection of the two more general approaches.

If we take a very specific choice of $f(\mathbf{x}) = \int_0^1 (\mathbf{x})^2 d\mu$ in the definition above, then

$$\delta f(\mathbf{v}; \mathbf{h}) = \lim_{h \to 0} \int_0^1 \frac{(\mathbf{v} + h\mathbf{h})^2 - (\mathbf{v})^2}{h} d\mu =$$

$$= \lim_{h \to 0} \int_0^1 \frac{2h\mathbf{h}\mathbf{v} + h^2(\mathbf{h})^2}{h} d\mu = \int_0^1 \mathbf{h}(2\mathbf{v}) d\mu = \langle \mathbf{h}, 2\mathbf{v} \rangle. \tag{6}$$

Therefore,

$$D_f(\mathbf{v}, \mathbf{w}(i)) = \sum_{i=1}^{n} \lambda_i \int_0^1 [\mathbf{w}^{(i)}]^2 - (\mathbf{v})^2 - \langle \mathbf{w}^{(i)} - \mathbf{v}, 2\mathbf{v} \rangle d\mu =$$

$$\sum_{i=1}^{n} \lambda_i \int_0^1 [\mathbf{w}^{(i)}]^2 - (\mathbf{v})^2 - (\mathbf{w}^{(i)} - \mathbf{v}) 2\mathbf{v} d\mu = \sum_{i=1}^{n} \lambda_i \int_0^1 [\mathbf{w}^{(i)} - \mathbf{v}]^2 d\mu.$$

This divergence corresponds to what the squared Euclidean distance was in the Euclidean setting. This is because $\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}$ defines in a Hilbert space the norm, and the above corresponds thus to $\sum_{i=1}^{n} \lambda_i \|\mathbf{w}^{(i)} - \mathbf{v}\|^2$, which in the Euclidean setting was a weighted sum of squared Euclidean distances, and $d(\mathbf{v}, \mathbf{w}^{(i)}) = \|\mathbf{w}^{(i)} - \mathbf{v}\|$ is the standard L^2 metric. Should the reader find working with the general divergence above difficult to grasp, confining the considerations to $D(\mathbf{v}, \mathbf{w}(i)) =$

$\sum_{i=1}^{n} \lambda_i [d(\mathbf{v}, \mathbf{w}^{(i)})]^2$ would be sufficient to illustrate this non–Euclidean setting. Nevertheless, the motivation to use the more general Bregman divergence is that it encompasses many other popular divergences. The Kullback–Leibler divergence is among the most prominent ones, and it is significant to us as it has been utilised in the motivating example elaborated in Section 4.3.

We can put the above simplified setting already in use when we investigate whether Property 1 (consistency) is satisfied. This property asks for any two $\mathbf{v}$ and $\mathbf{w}$ that $D_f(\mathbf{v}, \mathbf{w}) = 0$ be equivalent to $\mathbf{v} = \mathbf{w}$. Nevertheless, we get $D_f(\mathbf{v}, \mathbf{w}) = 0$ even for those $\mathbf{v}$ and $\mathbf{w}$ that differ on a set with Lebesgue measure zero. What we then want to work with are actually equivalence classes of Lebesgue measurable functions, and we shall never consider them outside integrals.

Another requirement of a divergence is that $D_f(\mathbf{v}, \mathbf{w}(i)) > 0$ for different $\mathbf{v}$ and $\mathbf{w}(i)$. Of course, this holds for the special example of D_f above, but in general one establishes this by observing that the argument of the integral in the definition of D_f is positive for all μ and i where $\mathbf{v}$ and $\mathbf{w}^{(i)}$ are different, as depicted in Figure 15. As long as they are different at a set that has a non–zero Lebesgue measure, the resulting integrals and the sum are positive as well.

We use similar reasoning to establish that D_f satisfies Property 9 (connectivity): If $\{D_f(\mathbf{v}_i, \mathbf{v})\}_{i=1}^{\infty} \to 0$, then

$$\liminf_{i\to\infty} \int_0^1 f(\mathbf{v}) - f(\mathbf{v}_i) - \langle \mathbf{v} - \mathbf{v}_i, \nabla f(\mathbf{v}_i)\rangle d\mu = 0,$$

where, without loss of generality, we assumed that all components of the Cartesian products agree and dropped the sum. Since we have already established that the argument of the integral is positive, we can use Fatou's lemma to establish that also

$$\int_0^1 \liminf_{i\to\infty} [f(\mathbf{v}) - f(\mathbf{v}_i) - \langle \mathbf{v} - \mathbf{v}_i, \nabla f(\mathbf{v}_i)\rangle] d\mu = 0.$$

Looking at Figure 15, we again know that if $\lim_{i\to\infty} d(\mathbf{v}_i, \mathbf{v}) > 0$ then

$$\liminf_{i\to\infty} [f(\mathbf{v}) - f(\mathbf{v}_i) - \langle \mathbf{v} - \mathbf{v}_i, \nabla f(\mathbf{v}_i)\rangle]$$

would be a function (the limit is defined for each function argument separately) non–zero on a set with a non–zero Lebesgue measure. That would be a contradiction.

The argument is more straightforward for the special case of

$$f(\mathbf{x}) = \int_0^1 (\mathbf{x})^2 d\mu,$$

when $D_f(\mathbf{v}_i, \mathbf{v}) = [d(\mathbf{v}_i, \mathbf{v})]^2$ and we get the connectivity property straight away.

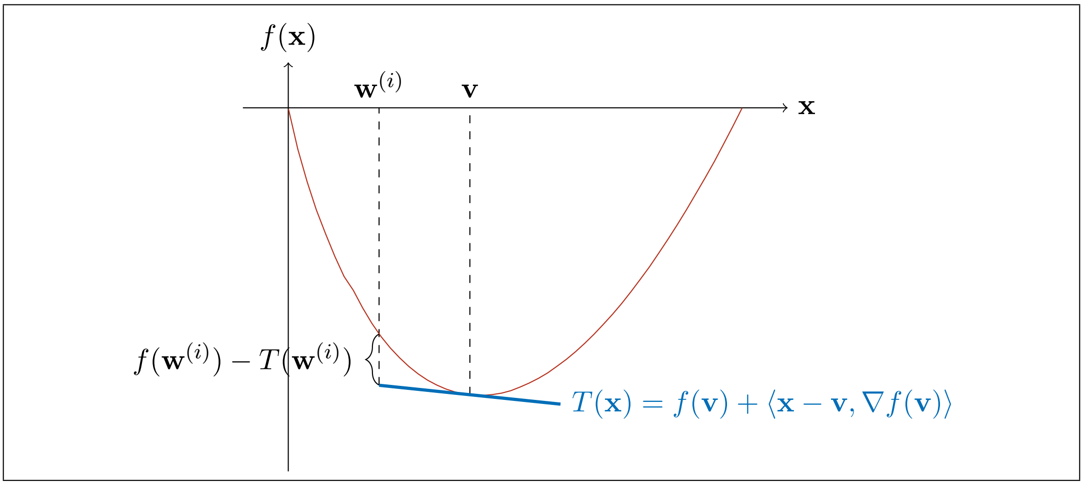

$f(\mathbf{x})$
$\mathbf{w}^{(i)}$
$\mathbf{v}$
$\mathbf{x}$
$f(\mathbf{w}^{(i)}) - T(\mathbf{w}^{(i)})$
$T(\mathbf{x}) = f(\mathbf{v}) + \langle \mathbf{x} - \mathbf{v}, \nabla f(\mathbf{v}) \rangle$

Figure 15: A single–dimensional impression of the integral argument in a Bregman divergence D_f is depicted below. $T(\mathbf{x}) = f(\mathbf{v}) + \langle \mathbf{x} - \mathbf{v}, \nabla f(\mathbf{v}) \rangle$, where $\nabla f(\mathbf{v})$ is the Fréchet derivative, is in this single–dimensional impression the tangent line to f at $\mathbf{v}$ in direction to $\mathbf{w}^{(i)}$. The divergence from $\mathbf{v}$ to $\mathbf{w}(i)$ is then given by integrating the difference $f(\mathbf{w}^{(i)}) - T(\mathbf{w}^{(i)})$, which is explicitly shown above, over μ and summing the results across i with corresponding weights λ_i.

Lemma 1. *For any* $\mathbf{v}$*,* $\mathbf{u}$ *and* $\mathbf{w}$*, and a Bregman divergence* D_f *we have*

$$D_f(\mathbf{v}, \mathbf{w}) = D_f(\mathbf{v}, \mathbf{u}) + D_f(\mathbf{u}, \mathbf{w}) + \langle \mathbf{w} - \mathbf{u}, \nabla f(\mathbf{u}) - \nabla f(\mathbf{v}) \rangle.$$

Proof. First, we rewrite the right–hand side in detail:

$$D_f(\mathbf{v}, \mathbf{u}) + D_f(\mathbf{u}, \mathbf{w}) + \langle \mathbf{u} - \mathbf{w}, \nabla f(\mathbf{w}) - \nabla f(\mathbf{v}) \rangle =$$

$$= \int_0^1 f(\mathbf{w}) - f(\mathbf{u}) - \langle \mathbf{w} - \mathbf{u}, \nabla f(\mathbf{u}) \rangle d\mu +$$
$$+ \int_0^1 f(\mathbf{u}) - f(\mathbf{v}) - \langle \mathbf{u} - \mathbf{v}, \nabla f(\mathbf{v}) \rangle d\mu +$$
$$+ \langle \mathbf{w} - \mathbf{u}, \nabla f(\mathbf{u}) - \nabla f(\mathbf{v}) \rangle. \tag{7}$$

Now, using the linearity of Lebesgue integration, (7) becomes

$$= \int_0^1 f(\mathbf{w}) - f(\mathbf{u}) - \langle \mathbf{w} - \mathbf{u}, \nabla f(\mathbf{u}) \rangle + f(\mathbf{u}) - f(\mathbf{v}) -$$
$$- \langle \mathbf{u} - \mathbf{v}, \nabla f(\mathbf{v}) \rangle + \langle \mathbf{w} - \mathbf{u}, \nabla f(\mathbf{u}) - \nabla f(\mathbf{v}) \rangle d\mu,$$

which, using the linearity of the real inner product, can be rewritten as

$$\int_0^1 f(\mathbf{w}) - f(\mathbf{u}) - \langle \mathbf{w}, \nabla f(\mathbf{u})\rangle + \langle \mathbf{u}, \nabla f(\mathbf{u})\rangle + f(\mathbf{u}) - f(\mathbf{v}) -$$
$$-\langle \mathbf{u}, \nabla f(\mathbf{v})\rangle + \langle \mathbf{v}, \nabla f(\mathbf{v})\rangle + \langle \mathbf{w}, \nabla f(\mathbf{u})\rangle - \langle \mathbf{u}, \nabla f(\mathbf{u})\rangle -$$
$$-\langle \mathbf{w}, \nabla f(\mathbf{v})\rangle + \langle \mathbf{u}, \nabla f(\mathbf{v})\rangle d\mu.$$

After simplifying, we obtain what is on the left–hand side:

$$\int_0^1 f(\mathbf{w}) - f(\mathbf{v}) + \langle \mathbf{v}, \nabla f(\mathbf{v})\rangle - \langle \mathbf{w}, \nabla f(\mathbf{v})\rangle d\mu =$$
$$= \int_0^1 f(\mathbf{w}) - f(\mathbf{v}) - \langle \mathbf{w} - \mathbf{v}, \nabla f(\mathbf{v})\rangle d\mu = D_f(\mathbf{v}, \mathbf{w}).$$

□

The lemma above was noted in [14] for Bregman divergences in L^p spaces, and it was called a generalized Pythagorean inequality. We will use this lemma above in the proof of the following key observation, and it will indeed lead to our Pythagorean properties. Recall that the sets W and V for the considered L^2 space setting were defined in Section 5.1.

Observation 9. *Let* $\mathbf{w}(i) = (\mathbf{w}^{(1)}, \dots, \mathbf{w}^{(n)}) \in W$*,* $\mathbf{w} = \sum_{i=1}^n \lambda_i \mathbf{w}^{(i)}$ *and* $\mathbf{v} \in V$*, and* D_f *be a Bregman divergence. Then*

$$D_f(\mathbf{v}, \mathbf{w}) + D_f(\mathbf{w}, \mathbf{w}(i)) = D_f(\mathbf{v}, \mathbf{w}(i)).$$

Proof. By Lemma 1, applying it to each summand in the weighted sum,

$$D_f(\mathbf{v}, \mathbf{w}(i)) = D_f(\mathbf{v}, \mathbf{w}) + D_f(\mathbf{w}, \mathbf{w}(i)) + \sum_{i=1}^n \lambda_i \langle \mathbf{w}^{(i)} - \mathbf{w}, \nabla f(\mathbf{w}) - \nabla f(\mathbf{v})\rangle,$$

so we wish to prove that

$$\sum_{i=1}^n \lambda_i \langle \mathbf{w}^{(i)} - \mathbf{w}, \nabla f(\mathbf{w}) - \nabla f(\mathbf{v})\rangle = \sum_{i=1}^n \lambda_i \int_0^1 (\mathbf{w}^{(i)} - \mathbf{w})(\nabla f(\mathbf{w}) - \nabla f(\mathbf{v})) d\mu$$

is zero. Since $\nabla f(\mathbf{w}) - \nabla f(\mathbf{v})$ is independent of i, the above is

$$\int_0^1 (\nabla f(\mathbf{w}) - \nabla f(\mathbf{v}))\Big[\sum_{i=1}^n \lambda_i (\mathbf{w}^{(i)} - \mathbf{w})\Big] d\mu =$$
$$= \int_0^1 (\nabla f(\mathbf{w}) - \nabla f(\mathbf{v}))[0] d\mu = 0$$

as required, since $\mathbf{w} = \sum_{i=1}^n \lambda_i \mathbf{w}^{(i)}$ by its definition. □

This observation proves that $\mathbf{w} = \sum_{i=1}^n \lambda_i \mathbf{w}^{(i)}$ is the conjugated D_f–projection of $\mathbf{w}(i)$ into V. If there was another point giving a smaller divergence, we could substitute it in place of $\mathbf{v}$ in the statement of the observation and since the divergence is always non–negative, we would have obtained a contradiction.

If the conjugated projection was not unique, repeating the same procedure we would obtain $D_f(\mathbf{v}, \mathbf{w}) = 0$, which is possible only if they differ at most at a set that has Lebesgue measure zero. So the projection would not need to be unique if we were not actually working with equivalence classes, but we do.

Writing $\mathbf{w} = \widehat{\pi}_V(\mathbf{w}(i))$, which is our notation for the unique conjugated projection, Observation 9 gives us the statement of the Pythagorean property for agreeable points. Indeed, for the special D_f where $f(\mathbf{x}) = \int_0^1 (\mathbf{x})^2 d\mu$,

$$\sum_{i=1}^n \lambda_i \langle \mathbf{w}^{(i)} - \mathbf{w}, \nabla f(\mathbf{w}) - \nabla f(\mathbf{v}) \rangle = \sum_{i=1}^n \lambda_i \langle \mathbf{w}^{(i)} - \mathbf{w}, 2\mathbf{w} - 2\mathbf{v} \rangle =$$

$$= 2 \sum_{i=1}^n \lambda_i \langle \mathbf{w}^{(i)} - \mathbf{w}, \mathbf{w} - \mathbf{v} \rangle = 0,$$

where we have used (6) on Page 38, gives for every i orthogonality of vectors $\mathbf{w}^{(i)} - \mathbf{w}$ and $\mathbf{w} - \mathbf{v}$ in a Hilbert space, and the observation becomes the usual Pythagorean identity of a Hilbert space.

5.3 Projections in Hilbert spaces

Since W_i defined in Section 5.1 are closed convex sets in L^2, a Hilbert space, by the Hilbert projection theorem [6] there is a unique $\mathbf{w}^{(i)} \in W_i$ such that

$$\mathbf{w}^{(i)} = \arg \min_{\mathbf{y}^{(i)} \in W_i} d(\mathbf{v}, \mathbf{y}^{(i)}) = \arg \min_{\mathbf{y}^{(i)} \in W_i} \|\mathbf{y}^{(i)} - \mathbf{v}\|$$

for a given $\mathbf{v}$. Since $\|\mathbf{w}^{(i)} - \mathbf{v}\|^2 < \|\mathbf{y}^{(i)} - \mathbf{v}\|^2$ is equivalent to $\|\mathbf{w}^{(i)} - \mathbf{v}\| < \|\mathbf{y}^{(i)} - \mathbf{v}\|$, there exists also a unique D_f–projection $\pi_W(\mathbf{v})$ if $f(\mathbf{x}) = \int_0^1 (\mathbf{x})^2 d\mu$ and

$$D_f(\mathbf{v}, \mathbf{y}(i)) = \sum_{i=1}^n \lambda_i [d(\mathbf{v}, \mathbf{y}^{(i)})]^2.$$

In the following, we will assume that D_f is such that the D_f–projection of every $\mathbf{v} \in V$ into W exists (above we have established that this is the case for our special divergence corresponding to the squared Euclidean distance), and having this assumption in place we can establish the Pythagorean property for admissible points.

Observation 10. *Let $\mathbf{u}(i) \in W$ and $\mathbf{v} \in V$, $\mathbf{w}(i) = \pi_W(\mathbf{v})$ exist and be unique, and D_f be a Bregman divergence. Then*

$$D_f(\mathbf{v}, \mathbf{w}(i)) + D_f(\mathbf{w}(i), \mathbf{u}(i)) \leq D_f(\mathbf{v}, \mathbf{u}(i)).$$

Proof. By Lemma 1 on Page 40, applying it to each summand in the weighted sum,

$$D_f(\mathbf{v}, \mathbf{u}(i)) = D_f(\mathbf{v}, \mathbf{w}(i)) + D_f(\mathbf{w}(i), \mathbf{u}(i)) + \sum_{i=1}^{n} \lambda_i \langle \mathbf{u}^{(i)} - \mathbf{w}^{(i)}, \nabla f(\mathbf{w}^{(i)}) - \nabla f(\mathbf{v}) \rangle,$$

so we wish to prove that

$$\sum_{i=1}^{n} \lambda_i \langle \mathbf{u}^{(i)} - \mathbf{w}^{(i)}, \nabla f(\mathbf{w}^{(i)}) - \nabla f(\mathbf{v}) \rangle =$$

$$= \sum_{i=1}^{n} \lambda_i \int_0^1 (\mathbf{u}^{(i)} - \mathbf{w}^{(i)})(\nabla f(\mathbf{w}^{(i)}) - \nabla f(\mathbf{v})) d\mu \geq 0.$$

Let us consider the convex combination of $\mathbf{w}(i)$ and $\mathbf{u}(i)$: $\lambda \mathbf{u}(i) + (1-\lambda)\mathbf{w}(i) \in W$, $\lambda \in [0,1]$, since W was assumed to be a convex set. If $\mathbf{w}(i) = \pi_W(\mathbf{v})$ is the D_f–projection of $\mathbf{v}$ into W, then

$$\frac{d}{d\lambda} D_f(\mathbf{v}, \lambda \mathbf{u}(i) + (1-\lambda)\mathbf{w}(i)) \Big|_{\lambda=0} \geq 0.$$

Spelling out the divergence we get

$$\frac{d}{d\lambda} \sum_{i=1}^{n} \lambda_i \int_0^1 f(\lambda \mathbf{u}^{(i)} + (1-\lambda)\mathbf{w}^{(i)}) - f(\mathbf{v}) - \langle \lambda \mathbf{u}^{(i)} + (1-\lambda)\mathbf{w}^{(i)} - \mathbf{v}, \nabla f(\mathbf{v}) \rangle d\mu,$$

which is by the Leibniz integral rule [28] (f is continuous and continuously differentiable by the assumption)

$$\sum_{i=1}^{n} \lambda_i \int_0^1 \frac{d}{d\lambda} f(\lambda \mathbf{u}^{(i)} + (1-\lambda)\mathbf{w}^{(i)}) \Big|_{\lambda=0} - \Big\langle \frac{d}{d\lambda} [\lambda \mathbf{u}^{(i)} + (1-\lambda)\mathbf{w}^{(i)}]_{\lambda=0}, \nabla f(\mathbf{v}) \Big\rangle d\mu =$$

$$= \sum_{i=1}^{n} \lambda_i \int_0^1 \langle \mathbf{u}^{(i)} - \mathbf{w}^{(i)}, \nabla f(\mathbf{w}^{(i)}) \rangle - \langle \mathbf{u}^{(i)} - \mathbf{w}^{(i)}, \nabla f(\mathbf{v}) \rangle d\mu =$$

$$= \sum_{i=1}^{n} \lambda_i \int_0^1 \langle \mathbf{u}^{(i)} - \mathbf{w}^{(i)}, \nabla f(\mathbf{w}^{(i)}) - \nabla f(\mathbf{v}) \rangle d\mu \geq 0,$$

as required. Note that the last equality is due to the linearity of an inner product. □

Having the Pythagorean and projection properties established, we turn our attention to Property 6 (Convexity). Let us first make the following simple observation.

Lemma 2. *For a given* $\mathbf{v} \in V$, $D_f(\mathbf{v}, \cdot)$ *is a convex function.*

Proof. Due to the monotonicity of the Lebesgue integral (i.e., $f \leq g$ implies $\int_0^1 f d\mu \leq \int_0^1 g d\mu$) and since f is a convex function, for a convex combination of $\mathbf{w}_1$ and $\mathbf{w}_2$: $\lambda \mathbf{w}_1 + (1-\lambda)\mathbf{w}_2$, we have

$$D(\mathbf{v}, \lambda \mathbf{w}_1 + (1-\lambda)\mathbf{w}_2) =$$

$$= \int_0^1 f(\lambda \mathbf{w}_1 + (1-\lambda)\mathbf{w}_2) - f(\mathbf{v}) - \langle \lambda \mathbf{w}_1 + (1-\lambda)\mathbf{w}_2 - \mathbf{v}, \nabla f(\mathbf{v})\rangle d\mu \leq$$

$$\leq \int_0^1 \lambda f(\mathbf{w}_1) + (1-\lambda) f(\mathbf{w}_2) - [\lambda + (1-\lambda)] f(\mathbf{v}) -$$

$$-\lambda \langle \mathbf{w}_1 - \mathbf{v}, \nabla f(\mathbf{v})\rangle + (1-\lambda)\langle \mathbf{w}_2 - \mathbf{v}, \nabla f(\mathbf{v})\rangle d\mu =$$

$$= \lambda \int_0^1 f(\mathbf{w}_1) - f(\mathbf{v}) - \langle \mathbf{w}_1 - \mathbf{v}, \nabla f(\mathbf{v})\rangle d\mu +$$

$$+(1-\lambda) \int_0^1 f(\mathbf{w}_2) - f(\mathbf{v}) - \langle \mathbf{w}_2 - \mathbf{v}, \nabla f(\mathbf{v})\rangle d\mu =$$

$$= \lambda D(\mathbf{v}, \mathbf{w}_1) + (1-\lambda) D(\mathbf{v}, \mathbf{w}_2).$$

Note that in the above we have used the linearity of an inner product. □

Even in Euclidean space, Bregman divergences are not necessarily convex in the other argument. Please note again that our order of arguments is opposite to how it is usually written in the literature, so by this other argument they would mean the second and not the first argument as we do.

In what follows, we will need to consider only those D_f that are convex functions in both arguments. Again, the choice $f(\mathbf{x}) = \int_0^1 (\mathbf{x})^2 d\mu$ gives us such a function. The following lemma then establishes the convexity property for such Bregman divergences.

Lemma 3. *Let* $\mathbf{w}(i), \mathbf{u}(i) \in W$, *and* $\mathbf{w} = \sum_{i=1}^n \lambda_i \mathbf{w}^{(i)}$ *and* $\mathbf{u} = \sum_{i=1}^n \lambda_i \mathbf{u}^{(i)}$. *We assume that* $D_f(\cdot, \cdot)$ *is a convex function (jointly in both arguments). Then*

$$D_f(\mathbf{w}(i), \mathbf{u}(i)) \geq D_f(\mathbf{w}, \mathbf{u}).$$

Proof. Since $D_f(\cdot,\cdot)$ is a convex function, using Jensen's inequality for integrals we obtain

$$\sum_{i=1}^{n} \lambda_i D_f(\mathbf{w}^{(i)}, \mathbf{u}^{(i)}) \geq D_f\Big(\sum_{i=1}^{n} \lambda_i \mathbf{w}^{(i)}, \sum_{i=1}^{n} \lambda_i \mathbf{u}^{(i)}\Big).$$

However, the left–hand side is just $D_f(\mathbf{w}(i), \mathbf{u}(i))$, which is

$$D_f(\mathbf{w}(i), \mathbf{u}(i)) = \sum_{i=1}^{n} \lambda_i \int_0^1 f(\mathbf{u}^{(i)}) - f(\mathbf{w}^{(i)}) - \langle \mathbf{u}^{(i)} - \mathbf{w}^{(i)}, \nabla f(\mathbf{w}^{(i)})\rangle d\mu,$$

because for each fixed i

$$D_f(\mathbf{w}^{(i)}, \mathbf{u}^{(i)}) = \int_0^1 f(\mathbf{u}^{(i)}) - f(\mathbf{w}^{(i)}) - \langle \mathbf{u}^{(i)} - \mathbf{w}^{(i)}, \nabla f(\mathbf{w}^{(i)})\rangle d\mu$$

for the same reason as

$$D_f(\mathbf{v}, \mathbf{w}) = \int_0^1 f(\mathbf{w}) - f(\mathbf{v}) - \langle \mathbf{w} - \mathbf{v}, \nabla f(\mathbf{v})\rangle d\mu.$$

Finally, as the right–hand side is by the definition $D_f(\mathbf{w}, \mathbf{u})$, the statement of the lemma follows. □

At this point, we have proven all properties needed to establish Theorem 1 on Page 21 for a convex Bregman divergence D_f in the L^2 space setting.

Corollary 1. *Let D_f be a convex Bregman divergence such that D_f–projections to W exist and are unique. If a representative point exists then the set of fixed points and the set of representative points are equal:*

$$\Delta(W) = \Theta(W).$$

5.4 Four point property

$D_f(\mathbf{v}, \cdot) : L^2 \to \mathbb{R}$ is Gâteaux differentiable in its second argument if there exists a functional $\delta_{\mathbf{w}(i)} D_f(\mathbf{v}, \mathbf{w}(i); \cdot) : L^2 \to \mathbb{R}$ such that

$$\delta_{\mathbf{w}(i)} D_f(\mathbf{v}, \mathbf{w}(i); \mathbf{h}(i)) = \lim_{h \to 0} \frac{D_f(\mathbf{v}, \mathbf{w}(i) + h\mathbf{h}(i)) - D_f(\mathbf{v}, \mathbf{w}(i))}{h}.$$

Since, by the linearity of the real inner product,

$$\sum_{i=1}^{n} \lambda_i \int_0^1 f(\mathbf{w}^{(i)} + h\mathbf{h}^{(i)}) - f(\mathbf{v}) - \langle \mathbf{w}^{(i)} + h\mathbf{h}^{(i)} - \mathbf{v}, \nabla f(\mathbf{v})\rangle d\mu -$$

$$-\sum_{i=1}^{n}\lambda_i\int_0^1 f(\mathbf{w}^{(i)})-f(\mathbf{v})-\langle\mathbf{w}^{(i)}-\mathbf{v},\nabla f(\mathbf{v})\rangle d\mu=$$

$$=\sum_{i=1}^{n}\lambda_i\int_0^1 f(\mathbf{w}^{(i)}+h\mathbf{h}^{(i)})-f(\mathbf{w}^{(i)})-h\langle\mathbf{h}^{(i)},\nabla f(\mathbf{v})\rangle d\mu,$$

we have that $D_f(\mathbf{v},\cdot)$ is indeed Gâteaux differentiable (and continuous), and

$$\delta_{\mathbf{w}(i)}D_f(\mathbf{v},\mathbf{w}(i);\mathbf{h}(i))=\lim_{h\to 0}\sum_{i=1}^{n}\lambda_i\int_0^1\frac{f(\mathbf{w}^{(i)}+h\mathbf{h}^{(i)})-f(\mathbf{w}^{(i)})}{h}-\langle\mathbf{h}^{(i)},\nabla f(\mathbf{v})\rangle d\mu=$$

$$=\sum_{i=1}^{n}\lambda_i\Big[\delta f(\mathbf{w}^{(i)};\mathbf{h}^{(i)})-\langle\mathbf{h}^{(i)},\nabla f(\mathbf{v})\rangle\Big]=$$

$$=\sum_{i=1}^{n}\lambda_i\Big[\langle\mathbf{h}^{(i)},\nabla f(\mathbf{w}^{(i)})\rangle-\langle\mathbf{h}^{(i)},\nabla f(\mathbf{v})\rangle\Big],\tag{8}$$

where in the second equality we have used Gâteaux differentiability of f from (5) on Page 37. Recall once again that due to the assumed Fréchet differentiability of f and the Riesz representation theorem [6] we can express the above differential as an inner product. As this real inner product is linear, the same is true for the differential so it is also Fréchet differentiable.

The special choice $f(\mathbf{x})=\int_0^1(\mathbf{x})^2d\mu$ gives a symmetric $D_f(\cdot,\cdot)$, so this divergence is continuous and Fréchet differentiable in both arguments. In order to prove the four point property for a general D_f, we will need to impose such a strong assumption on its differentiability:

Observation 11. *Let $\mathbf{w}(i),\mathbf{u}(i)\in W$, and $\mathbf{w}=\sum_{i=1}^{n}\lambda_i\mathbf{w}^{(i)}$ and $\mathbf{v}\in V$. Assume that $D_f(\cdot,\cdot)$ is Fréchet differentiable (and hence continuous) and convex (jointly in both arguments). Then*

$$D_f(\mathbf{w},\mathbf{u}(i))\le D_f(\mathbf{w}(i),\mathbf{u}(i))+D_f(\mathbf{v},\mathbf{u}(i)).$$

Proof. Since D_f is a Fréchet differentiable convex function, we can write

$$D_f(\mathbf{v},\mathbf{u}(i))\ge D_f(\mathbf{w},\mathbf{w}(i))+$$

$$+\delta_{\mathbf{w}}D_f(\mathbf{w},\mathbf{w}(i);\mathbf{v}-\mathbf{w})+\delta_{\mathbf{w}(i)}D_f(\mathbf{w},\mathbf{w}(i);\mathbf{u}(i)-\mathbf{w}(i)).$$

Using (5.4), the above becomes

$$D_f(\mathbf{v},\mathbf{u}(i))\ge D_f(\mathbf{w},\mathbf{w}(i))+$$

$$+\delta_{\mathbf{w}} D_f(\mathbf{w}, \mathbf{w}(i); \mathbf{v}-\mathbf{w}) + \sum_{i=1}^{n} \lambda_i \langle \mathbf{u}^{(i)} - \mathbf{w}^{(i)}, \nabla f(\mathbf{w}^{(i)}) - \nabla f(\mathbf{w}) \rangle. \tag{9}$$

By Lemma 1 on Page 40,

$$D_f(\mathbf{w}, \mathbf{u}(i)) = D_f(\mathbf{w}, \mathbf{w}(i)) + D_f(\mathbf{w}(i), \mathbf{u}(i)) + \sum_{i=1}^{n} \lambda_i \langle \mathbf{u}^{(i)} - \mathbf{w}^{(i)}, \nabla f(\mathbf{w}^{(i)}) - \nabla f(\mathbf{w}) \rangle,$$

so replacing $D_f(\mathbf{w}, \mathbf{w}(i))$ in (9) with this we obtain

$$D_f(\mathbf{v}, \mathbf{u}(i)) \geq D_f(\mathbf{w}, \mathbf{u}(i)) - D_f(\mathbf{w}(i), \mathbf{u}(i)) +$$
$$+\delta_{\mathbf{w}} D_f(\mathbf{w}, \mathbf{w}(i); \mathbf{v}-\mathbf{w}). \tag{10}$$

Finally, since $\mathbf{x} = \mathbf{w} = \sum_{i=1}^{n} \lambda_i \mathbf{w}^{(i)} = \widehat{\pi}_V(\mathbf{w}(i))$ minimises $D_f(\mathbf{x}, \mathbf{w}(i))$ subject to $\mathbf{x} \in V$ for a given $\mathbf{w}(i)$, the Gâteaux differential in (10) must be non–negative and the observation follows. □

In the above observation we have established the four point property in a Hilbertian space setting, see Figure 16 for an illustration. We have done that however only under the further assumption that D_f is a Fréchet differentiable convex function (jointly in both arguments), which also implies that D_f satisfies the last property that we have not touched so far: Property 7, continuity.

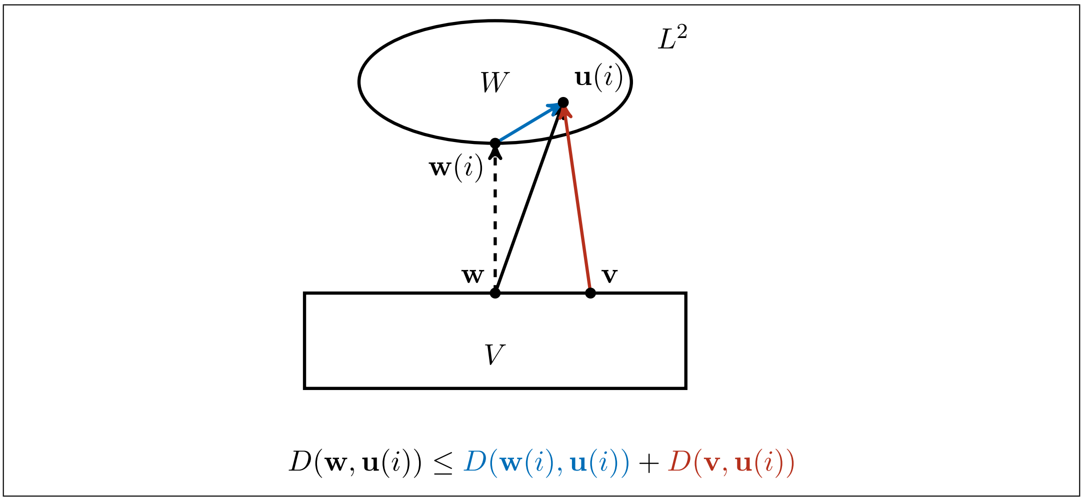

L^2
W
$\mathbf{u}(i)$
$\mathbf{w}(i)$
$\mathbf{w}$
$\mathbf{v}$
V
$D(\mathbf{w}, \mathbf{u}(i)) \leq D(\mathbf{w}(i), \mathbf{u}(i)) + D(\mathbf{v}, \mathbf{u}(i))$

Figure 16: An illustration of the four points property in a Hilbertian space setting.

At this point, we have proven all properties needed to establish Theorem 2 on Page 26 for a convex Fréchet differentiable Bregman divergence D_f in the L^2 space setting. Recall that conditions needed for compactness of sets W and V for the considered L^2 space setting were spelt out in Section 5.1 starting on Page 34.

Corollary 2. *Let D_f be a convex Fréchet differentiable Bregman divergence such that D_f–projections to W exist and are unique. Let $\mathbf{v}_0 \in V$. Define a sequence $\{\mathbf{v}_i\}_{i=0}^{\infty}$ recursively by $\mathbf{v}_{i+1} = \widehat{\pi}_V(\pi_W(\mathbf{v}_i))$. If V and W are compact then the sequence $\{\mathbf{v}_i\}_{i=0}^{\infty}$ converges to a fixed point.*

Let us repeat that D_f where $f(\mathbf{x}) = \int_0^1 (\mathbf{x})^2 d\mu$ satisfies all the requirements above.

6 Conclusion

In this paper we have provided an introduction to information geometry inspired by uncertain reasoning and belief merging (Section 2), explained the alternating minimisation procedure (Theorem 2 on Page 26) and an application to merging study findings (Section 4.3), and proven related results in the L^2 space setting (Section 5). The main aim was to make this paper accessible to logicians and philosophers, and to that end, a concept of admissible and agreeable points was invented. Consider how many readers would have been lost should the L^2 space setting of Section 5 were to be presented first.

We hope that this paper contributes to education, inspires logicians and philosophers to combine information geometry and logical principles, and covers what appears to be a lack of related proofs in the intersection of two more popular general approaches: Hilbertian and function space settings. After all, mathematics was always about working in the right set up. And the L^2 space setting is just fairly intuitive.

Finally, we should also admit that the results presented in Sections 2 and 3 were somewhat easy; we imposed enough properties so that the proofs of Theorems 1 and 2 went through. The hard job is the opposite: What properties are necessary? For the notions of cross–entropy and entropy this has been achieved by Shore and Johnson [27], and Paris and Vencovská [25], respectively. In our context, the question of axiomatising the alternating minimisation procedure remains open for now and we would like to encourage other researchers to investigate it.

References

[1] Martin Adamčík. *Collective Reasoning under Uncertainty and Inconsistency.* Phd thesis, University of Manchester, 2014.

[2] Martin Adamčík. The information geometry of Bregman divergences and some applications in multi–expert reasoning. *Entropy*, 16:6338–6381, 2014.

[3] Martin Adamčík. On the applicability of the 'number of possible states' argument in multi–expert reasoning. *Journal of Applied Logic*, 19:20–49, 2016.

[4] Martin Adamčík. A logician's approach to meta–analysis with unexplained heterogeneity. *Journal of Biomedical Informatics*, 71:110–129, 2017.

[5] Martin Adamčík. A note on how Rényi entropy can create a spectrum of probabilistic merging operators. *Kybernetika*, 55:605–617, 2019.

[6] Ben Adler. Hilbert spaces and the Riesz representation theorem. `https://math.uchicago.edu/~may/REU2021/REUPapers/Adler.pdf`, 2021. [Online; accessed 15–May–2023].

[7] Arindam Banerjee, Xin Guo, and Hui Wang. On the optimality of conditional expectation as a Bregman predictor. *IEEE Transactions on Information Theory*, 51:2664–2669, 2005.

[8] Heinz H. Bauschke, Patrick L. Combettes, and Dominikus Noll. Joint minimization with alternating Bregman proximity operators. *Pacific Journal of Optimization*, 2:401–524, 2006.

[9] Lev M. Bregman. The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 1:200–217, 1967.

[10] Regina S. Burachik and Alfredo N. Iusem. A generalized proximal point algorithm for the variational inequality problem in a Hilbert space. *SIAM Journal on Optimization*, 8:197–216, 1998.

[11] Vaughn Climenhaga. Function spaces and compactness. `https://www.math.uh.edu/~climenha/blog-posts/function-spaces.pdf`, 2013. [Online; accessed 15–May–2023].

[12] Imre Csiszár. I–divergence geometry of probability distributions and minimization problems. *The Anals of Probability*, 3:146–158, 1975.

[13] Imre Csiszár and Gábor Tusnády. Informational geometry and alternating minimization procedures. *Statistic and Decisions*, 1:205–237, 1984.

[14] Béla A. Frigyik, Santosh Srivastava, and Maya R. Gupta. Functional Bregman divergence and Bayesian estimation of distributions. *IEEE Transactions on Information Theory*, 54:5130–5139, 2008.

[15] I. M. Gelfand and S. V. Fomin. *Calculus of Variations.* Prentice–Hall, New Jersey, 1963. English translation by Richard A. Silverman.

[16] Christian Genest and Carl G. Wagner. Further evidence against independence preservation in expert judgement synthesis. *Aequationes Mathematicae*, 32:74–86, 1987.

[17] Harald Hanche-Olsen and Helge Holden. The Kolmogorov—Riesz compactness theorem. *Expositiones Mathematicae*, 28:385–394, 2010.

[18] Edwin T. Jaynes. Where do we stand on maximum entropy? In R. D. Levine and M. Tribus, editors, *The Maximum Entropy Formalism*, pages 15–118. M.I.T. Press,

1979.

[19] Rob Johnson. Jensen's inequality for integrals. `https://math.stackexchange.com/questions/171599/jensens-inequality-for-integrals/171625#171625`, 2012. [Online; accessed 15–May–2023].

[20] David E. Joyce. Online eddition of Euclid's Elements. `http://aleph0.clarku.edu/~djoyce/java/elements`, 1998. [Online; accessed 12–December–2016].

[21] Donald W. Kahn. *Topology.* Dover Publications, New York, 1995.

[22] Sébastien Konieczny and Ramón Pino-Pérez. On the logic of merging. In A. G. Cohn, L. Schubert, and S. C. Shapiro, editors, *Proceedings of the 6th International Conference on Principles of Knowledge Representation and Reasoning*, pages 488–498. Morgan Kaufmann Publishers, 1998.

[23] Gary H. Meisters. Lebesgue measure on the real line. `https://math.unl.edu/~gmeisters1/papers/Measure/measure.pdf`, 1997. [Online; accessed 15–May–2023].

[24] Jeff B. Paris. *The Uncertain Reasoner Companion.* Cambridge University Press, Cambridge, 1994.

[25] Jeff B. Paris and Alena Vencovská. A note on the inevitability of maximum entropy. *International Journal of Approximate Reasoning*, 4:183–223, 1990.

[26] Alfréd Rényi. On measures of entropy and information. In J. Neyman, editor, *Proc. Fourth Berkeley Symposium on Mathematics, Statistics and Probability*, pages 547–561. University of California Press, 1961.

[27] John E. Shore and Rodney W. Johnson. Properties of cross–entropy minimization. *IEEE Transactions on Information Theory*, IT–27:472–482, 1981.

[28] Erik Talvila. Necessary and sufficient conditions for differentiating under the integral sign. *The American Mathematical Monthly*, 108:544–548, 2001.

[29] Jon Williamson. Deliberation, judgement and the nature of evidence. *Economics and Philosophy*, 31:27–65, 2015.

[30] George M. Wilmers. A foundational approach to generalising the maximum entropy inference process to the multi–agent context. *Entropy*, 17:594–645, 2015.

Received January 2023

Analogical Proportion-Based Induction: From Classification to Creativity

Henri Prade
IRIT – CNRS, 118, route de Narbonne, 31062 Toulouse Cedex 9, France [*]
henri.prade@irit.fr

Gilles Richard
IRIT – CNRS, 118, route de Narbonne, 31062 Toulouse Cedex 9, France
gilles.richard@irit.fr

Abstract

The first aim of this article is to position analogical inference (or at least a particular form of it) in relation to induction. After a brief reminder on the induction of plausible conclusions in a probabilistic, logical, possibilistic settings, and with J. S. Mill's methods of induction, we turn our attention to analogical inference, based on analogical proportions. Analogical proportions that hold between Boolean vectors are emphasized as a matter of pairs belonging to the same equivalence class. Then the mechanism of analogical proportions-based classification is explained and the main algorithms and results obtained so far are surveyed. After which, steps towards a logic of creativity are presented. The approach starts from the observation that analogical proportions belong to a larger set of quaternary relations called logical proportions. The six logical proportions giving birth to an equivalence relation between pairs are identified. This includes two important cases: i) a logic of conditional events known as being a basis for non monotonic reasoning (which is a form of plausible deduction); ii) a logic of ordered pairs preserving positive changes, closely related to analogical proportions. Within this framework, we revisit the creative nature of analogical proportions and introduce a creative inference mechanism that works on the basis of a specific situation and a collection of ordered pairs representing possible changes.

Special thanks are due to two anonymous referees for their careful reading and their suggestions to improve clarity and avoid confusion.

[*] This research has been supported by the ANR project "Analogies: from theory to tools and applications" (AT2TA), ANR-22-CE23-002.

1 Introduction

As it is well-known, Charles Sanders Peirce [44] distinguished between three main forms of reasoning: deduction (for deriving necessary consequences), induction (for extrapolating and generalizing from facts), and abduction (for finding out an hypothesis explaining a particular situation).

Peirce [45] was viewing analogical reasoning as a composite form of reasoning that combines induction and abduction (or retroduction, using Peirce's word). Roughly speaking, the idea is to see the fact of relating a given situation to another known situation as a kind of abduction, and the fact of projecting / extrapolating what happens in the latter situation onto the given situation (in order to predict something about it) as a form of induction. In fact, Peirce did not detail the way the two modes of reasoning are combined in analogical reasoning; see [41] and [42] for different understandings of Peirce's suggestion that may also involve deduction.

Indeed analogical reasoning involves drawing a plausible inference that a property found (or holding) in one situation is likely to also apply/hold to/in a second situation when there are significant similarities between the two situations. Such a parallel between two situations is also the starting point of Gentner [27]'s structure-mapping theory, where analogy is viewed as a mapping from one source domain to a target domain which conveys that a system of relations which holds among the source entities also holds among the target entities. Such a view was also proposed in [57]. A restricted rendering of this idea, stated in logical terms, making a parallel between two entities x and y, amounts to infer $Q(y)$ from $P(x), Q(x)$, and $P(y)$ (where P and Q are predicates applicable to situations x and y). Such an inference is clearly an "analogical jump", which offers no guarantee on the conclusion. Some authors [14, 54] looked for what could be added to such premises in order to ensure the truth of the conclusion. But this view is too demanding since it amounts to reduce analogical inference to a form of deduction.

This view of the "analogical jump" pattern of inference is not so far from case-based reasoning [25]. Indeed in this pattern, a "case" y under consideration where P is true is related to a known "case" x where P is true as well[1] (i.e., we have a perfect similarity between x and y) as well as another property Q also true for x. So as in case-based reasoning, this suggests that Q may be true as well for y.

In this paper, we use a slightly different view of analogical inference that is based on analogical proportions. Analogical proportions are statements of the form "a is to b as c is to d", as for instance, "a calf is to a cow as a foal is to a mare". Although it involves four items, we can observe that a calf and a cow are bovids, which corresponds to a first situation, while a foal and a mare are equids, which is the second situation. In each sit-

[1] Note that P may be a compound property involving a number of elementary properties.

uation, properties or relation such as “young”, “adult”, “mother” apply to the entities put in correspondence by the analogical proportion. More generally the pair (a, b) with all the properties and relations attached to items a and b is paralleled with the pair (c, d) with all the properties and relations attached to items c and d. A philosopher, Mary Hesse [30], pointed out and discussed the link between analogy viewed as a parallel between two situations and analogical proportions.

In the following, we shall use an inference based on analogical proportions. Although its link with the “analogical jump” pattern (as a particular instance) has been established [8], this inference is different from the one in case-base reasoning where the known cases in the repertory are considered one by one in isolation [50][2]. In contrast, in analogical proportion-based inference, the case under consideration, say d, is related to a triplet a, b, c of known cases, each case being described by a set of attribute values [36]. .

Analogical inference is known to be useful in a wide range of applications [28], from aiding explanation [5] to aiding creativity, e.g., in mathematics [47]. In this paper, we shall review how analogical proportion-based inference can be applied to classification tasks, and we shall advocate a new view of analogical proportions that will lead us to propose some elements of a new logic that manipulates ordered pairs of vectors, and that may be used in a creativity task.

The paper is organized as follows. Section 2 discusses analogical inference as a form of induction, together with a brief review of some non conventional form of induction. Section 3 provides a formal survey of analogical proportions. It emphasizes a new view of (Boolean) analogical proportions that express an *equivalence* relation between two ordered pairs of vectors describing items in terms of Boolean features. Section 4 recalls analogical inference (based on analogical proportions) and the main results obtained with this inference in classification. Section 5 first briefly restates the general setting of so-called logical proportions, to which analogical proportions belong. We focus on those logical proportions that define equivalence relations. We thus identify two families of such proportions, one which defines equivalences between “conditional objects” and which is at the basis of nonmonotonic reasoning, and the other which is related to analogical proportions and from which we outline a logic of ordered pairs, applicable to creativity issues.

2 Analogical inference and transduction

Induction is classically opposed to deduction. While deduction applies generic knowledge (represented by general rules (“all men are mortal”) to factual information (“Socrate is a

[2] Usually in case-based reasoning the case under consideration y is put in similarity relation with several known cases $x_1, \cdots, x_k$ where each x_i leads to a partial conclusion for y (provided that the similarity between x_i and y is judged sufficient), and the partial conclusions have then to be combined in some way.

man")[3], induction works in the opposite direction, attempting to infer general laws from a limited set of observed facts. Mathematical models of induction predominantly stem from probability and statistics theories. Sometimes what is covered by the word 'induction' also includes the deductive step which corresponds to the application of the general laws induced to a factual situation.

But there is another form of induction named *transduction* [26] that directly infers a particular factual conclusion from a set of data. A well-known transduction-based mechanism is the k-nearest neighbor method where the prediction for a new item is only based on the observation of closed neighbors. It is also known as lazy learning. In both cases, induction is a form of inference that provides plausible conclusions, unlike deduction, which provides safe conclusions.

Let us consider a classification problem. Given a set of items or entities (objects, situations, profiles, ...) that are all described in terms of the values of a collection of observable attributes applicable to all of them, and that belong to a known class, the classification problem amounts to assign a class to a new item whose class is unknown. More formally, let S be a set of m items, each one $\vec{a^i}$ is represented by a vector of n attribute values $\vec{a^i} = (a^i_1, a^i_2, \cdots, a^i_n)$ $i = 1, ..., m$, together with its class/label $cl(\vec{a^i}) \in \mathcal{C}$, where $\mathcal{C}$ is a finite set of labels: for instance $\mathcal{C} = \{good, bad\}$, or $\mathcal{C} = \{bovid, equid, canid\}$. The set of attributes used to describe an observable piece of data is fixed: for instance $\{color, age, weight, position, \ldots\}$. Thus, a class gathers items of the same kind in a categorization process. Each class $C \in \mathcal{C}$ divides the set of items into the $\vec{a^i}$'s that are examples of C if $\vec{a^i} \in C$, and those that are counter-examples for C (and examples of other classes). In general, attributes can take into account values of various types, such as integers, real numbers, words, and more. In a context of binary attributes (i.e., attributes with values belonging to $\{0, 1\}$), attributes can be regarded as properties, for instance, representing whether an observable individual is wealthy (1 for 'rich,' 0 for 'not rich').

Given a new item $\vec{a^\star} = (a^\star_1, a^\star_2, \cdots, a^\star_n) \notin S$, the problem is then to predict its class.

When applying Bayesian classification with the assumption that the attributes are statistically independent of each other given the class, we calculate, for each $C \in \mathcal{C}$:

$$Prob(C|\vec{a^\star}) = \frac{Prob(\vec{a^\star}|C) \cdot Prob(C)}{Prob(\vec{a^\star})} = \frac{1}{Prob(\vec{a^\star})} \cdot Prob(C) \cdot \prod_{k=1}^{n} Prob(a^\star_k|C)$$

and $\vec{a^\star}$ is assigned to the class $C \in \mathcal{C}$ providing the highest value for $Prob(C|\vec{a^\star})$.

The probabilistic approach is the prevailing viewpoint for induction and transduction. Still there exist other options. Let us briefly review them.

[3] Even if deduction can also infer new generic rules from given rules.

Possibility theory [19] also has a Bayesian-like rule, which writes [20],

$$\Pi(C|\vec{a^{\star}}) \circledast \Pi(\vec{a^{\star}}) = \Pi(\vec{a^{\star}}|C) \circledast \Pi(C)$$

where $\circledast = \min$ or $\circledast =$ product. These two options separate the theories of qualitative and quantitative possibilities where conditioning is based on min and product respectively [19]. In case of product-based conditioning on obtains the expression

$$\Pi(C|\vec{a^{\star}}) = \frac{\Pi(\vec{a^{\star}}|C)}{\max_{j:C^j \in \mathcal{C}} \Pi(a^{\star}|C^j)}$$

under the hypothesis of no prior information, i.e., $\Pi(C) = 1$. More generally $\forall j, \Pi(C^j) = 1$, since $\Pi(\vec{a^{\star}}) = \max_j \Pi(a^{\star}|C^j) \cdot \Pi(C^j)$. The result is very close to Edwards' notion of likelihood [22], who advocates a non-Bayesian view, i.e., without priors. See, for instance, [7] for experiments with possibilistic classification.

If we use the min-based qualitative conditioning, i.e., $\circledast = \min$, assuming that class C is fully possible (no prior, i.e., $\Pi(C) = 1$), and taking $\Pi(\vec{a^{\star}}) = 1$, the equality $\Pi(C|\vec{a^{\star}}) \circledast \Pi(\vec{a^{\star}}) = \Pi(\vec{a^{\star}}|C) \circledast \Pi(C)$ reduces to $\Pi(C|\vec{a^{\star}}) = \Pi(\vec{a^{\star}}|C)$ and finally with an hypothesis of non-interactivity (logical independence) of attributes we obtain

$$\Pi(C|\vec{a^{\star}}) = \min_{k=1}^{n} \Pi(a^{\star}_k|C)$$

It expresses that C is all the more a possible class for $\vec{a^{\star}}$ as all its attribute values are possible in class C, or better that the less possible is one of the attribute values for the class C, the less possible class C is for $\vec{a^{\star}}$. We then come closer to a logical analysis of data that we examine now.

Considering a particular class $C \in \mathcal{C}$, a simple logical reading for Boolean data sets can be done from the sets of examples $\mathcal{E}_C$ and counter-examples $\mathcal{E}'_C$. Let $\varphi(C)$ be a logical formula that describes class C, i.e., $\varphi(C)$ is true for the description of any item belonging to $\varphi(C)$. Then we have [21]

$$\bigvee_{i:\vec{a^i}\in\mathcal{E}} (a^i_1 \wedge a^i_2 \wedge \cdots \wedge a^i_n) \models \varphi(C) \models \bigwedge_{j:\vec{a'^j}\in\mathcal{E}'} (\neg a'^j_1 \vee \neg a'^j_2 \vee \cdots \vee \neg a'^j_n)$$

This means that the description of any example in terms of the n Boolean attributes makes $\varphi(C)$ true, and that any model of $\varphi(C)$ falsifies at least one attribute value of any counter-example. Provided that the data are not noisy, this provides a consistent bracketing of $\varphi(C)$. This looks like the version space approach [43] where the hypotheses space is bracketed between a lower and an upper bound computed from the examples and counter-examples, except that here no representation bias is introduced.

In such a view, examples and counter-examples are exploited one by one. This is also the case for the possibilistic approach, although differently. In other words, the informational contributions made by each example (or counter-example) are combined, but no comparison of examples belonging to the same class, or to different classes, is made. This contrasts, as already mentioned in the introduction, with the analogical proportion-based inference where items are handled three by three ; see Sections 3 and 4 for details. For their part, k-nearest neighbors methods [24, 55], which are based on ideas similar to the ones at work in case-based reasoning, also handle examples one by one (even if there is some cumulative counting across the k examples considered).

The XIXth English philosopher, logician and economist, John Stuart Mill is known, among many other things, for his "methods of induction". He indeed proposed five methods of induction in his 1843 treatise of logic [40]. Strictly speaking, it is a matter of abduction rather than induction, but it was long before Charles Sanders Peirce distinguished between the two notions! Indeed these "induction" methods look for the simplest and most likely hypothesis that explains some observations. But what contrasts Mill's methods with the previously reviewed approaches to induction is that they heavily rely on the ideas of *agreement* and *difference*, which makes them somewhat similar in that respect to analogical proportion-based inference, where examples are compared within pairs, as we shall see in Section 3.

We cite here only the two main methods. The first one, called the *(Direct) Method of Agreement* is stated like that [40]:

$ABCD$ occur together with $wxyz$
$AEFG$ occur together with $wtuv$

Therefore A is the cause, or the effect, of w

Mill is not at all precise about what the letters in his induction patterns refer to. However the causality flavor suggests to regard letters as properties or attributes that characterize circumstances and their consequences (hence the use of two kinds of letters). The number of attributes involved in the two first statements that differ from one to the other has no particular meaning here. What is important here is that two states of affairs $ABCDwxyz$ on the one hand and $AEFGwtuv$ on the other hand are *compared*.

In other words, what $\{A, B, C, D\}$ and $\{A, E, F, G\}$ have in common, i.e., $\{A\}$), corresponds to what $\{w, x, y, z\}$ and $\{w, t, u, v\}$ have in common, i.e., $\{w\}$.

A similar analysis applies to the *Method of Difference*, which reads [40]:

$ABCD$ occur together with $wxyz$
BCD occur together with xyz

Therefore A is the cause, or the effect, or a part of the cause of w

Observe in the second pattern of inference that the set differences between $\{A, B, C, D\}$ and $\{B, C, D\}$ correspond to the set differences between $\{w, x, y, z\}$ and $\{x, y, z\}$.

Still Mill was making no link between his Methods of Difference and Agreement and his view of analogy, although he considered that analogical reasoning (viewed according to the pattern discussed in the introduction) was also a form of induction. The reader is referred to [4] for some further discussions on links between Mill's methods of induction and the analogical proportions-based inference that we are going to present in the next section.

3 Analogical proportions

This section provides a refresher on Boolean-valued analogical proportions together with a new view in terms of an equivalence relation between pairs.

3.1 Postulates

Generally speaking, analogical proportions are statements of the form "a is to b as c is to d" linking four entities, which are supposed to satisfy the following postulates, according to a long tradition that dates back to Aristotle. These postulates were inspired by a parallel with numerical proportions, namely, arithmetic proportions ($a - b = c - d$) and geometric proportions ($\frac{a}{b} = \frac{c}{d}$) which equalize differences and ratios respectively. Thus, proportions operates a double comparison (inside, and between, pairs) as also suggested by the expression of an analogical proportion "a is to b as c is to d". Such a proportion, considered here as a relation between 4 items, is denoted by $a : b :: c : d$.

These postulates, which are the only ones classically associated to an analogical proportion, are:

P1 *reflexivity*: $a : b :: a : b$;

P2 *symmetry*: $a : b :: c : d \Rightarrow c : d :: a : b$;

P3 stability under *central permutation*: $a : b :: c : d \Rightarrow a : c :: b : d$.

As immediate consequences, an analogical proportion also satisfies

- $a : a :: b : b$ (sameness) ;
- $a : b :: c : d \Rightarrow d : b :: c : a$ (external permutation);
- $a : b :: c : d \Rightarrow b : a :: d : c$ (internal reversal);
- $a : b :: c : d \Rightarrow d : c :: b : a$ (complete reversal).

It is also worth noticing that the 3 postulates *do not allow* for other permutations such that $a : b :: c : d \Rightarrow b : a :: c : d$, or $a : b :: c : d \Rightarrow c : b :: a : d$. In particular, $a : b :: b : a$ is not supposed to hold. Thus one cannot say that "black is to white as white is to black" (even if this might be advocated on the basis that the relation of opposition between 'black' and 'white' is the same as the one between 'white' and 'black'; see [37] for a discussion of a relation-based view of analogical proportions, maybe closer to natural language practice.

The entities involved in an analogical proportion can be of different natures: it may be numbers, words, drawings, images, sentences, ... [52]. This may lead to question the validity of the central permutation postulate, and to replace it by a weaker postulate such as, e.g., the internal reversal property ; see [37, 2] for discussions.

It is assumed in this article that the entities considered can be represented by vectors of Boolean feature values. Moreover, we require that the features used in the representation are applicable to the four entities involved in the analogical proportions we deal with.[4] Note that in this paper we are primarily interested in the inference mechanism associated with the logical modeling of analogical proportions between Boolean-valued vectors. Even if we use some analogical proportion stated in natural language for illustration, we do not intent to discuss analogical proportions between words in ordinary language here ; see [37] on this point.

Thus, any entity is represented here by a vector $\vec{a} = (a_1, ..., a_n)$ where a_i is the value of feature (or attribute) i. We define the analogical proportion relation among n-tuples by applying it componentwise. Namely,

$$\vec{a} : \vec{b} :: \vec{c} : \vec{d} \text{ if and only if } \forall i \in [1, n], a_i : b_i :: c_i : d_i$$

We need now to recall the definition of an analogical proportion for four Boolean variables a_i, b_i, c_i, d_i representing the value of some feature i for four distinct items.

3.2 Boolean proportions

The reflexivity postulate $a : b :: a : b$ forces a Boolean analogical proportion to be true for any values of a and b in $\{0, 1\}$, and therefore an analogical proportion $a : b :: c : d$ is true for the valuations $(0, 0, 0, 0)$, $(0, 1, 0, 1)$, $(1, 0, 1, 0)$, and $(1, 1, 1, 1)$. The unique *minimal* Boolean model that satisfies the three postulates P1, P2, P3 is made up of the 6 valuations shown in Table 1, where the valuations $(0, 0, 1, 1)$, $(1, 1, 0, 0)$ are added due to P3. As can be seen the 6 patterns are symmetrical (i.e., satisfy P2). The $2^4 - 6 = 10$ other valuations

[4] This assumption excludes analogical proportions such as "beer is to the English what wine is to the French" where two different conceptual universes are present (beverages and people, in the example). Note that the central permutation does not hold for such analogical proportions, which require a more sophisticated modeling [3].

a	b	c	d
0	0	0	0
1	1	1	1
0	0	1	1
1	1	0	0
0	1	0	1
1	0	1	0

Table 1: Minimal Boolean model of $a : b :: c : d$

are excluded, namely $a : b :: c : d$ is false for $(0, 1, 1, 0)$, for $(1, 0, 0, 1)$, for the 4 valuations with only one 0, and for the 4 valuations with only one 1 [51]:

There are several remarkable quaternary logical formulas for an analogical proportion $a : b :: c : d$, all of which are logically equivalent. Thus, they are all true only for the 6 valuations of Table 1 (and false for the 10 remaining valuations). The first formula uses *dissimilarity* indicators only, inside pairs (a, b) and (c, d). Indeed we have [39]:

$$a : b :: c : d = ((a \wedge \neg b) \equiv (c \wedge \neg d)) \wedge ((\neg a \wedge b) \equiv (\neg c \wedge d)) \tag{1}$$

It precisely expresses that "a differs from b as c differs from d and b differs from a as d differs from c" (and "when a and b do not differ, c and d do not differ").

A second formula, logically equivalent to expression (1), is [39]:

$$a : b :: c : d = ((a \wedge d) \equiv (b \wedge c)) \wedge ((\neg a \wedge \neg d) \equiv (\neg b \wedge \neg c)) \tag{2}$$

It uses *similarity* indicators only and can be read as "what a and d have in common (positively or negatively) b and c have it also"[5].

Another expression, equivalent to the two above logical formulas [51], which is closer to the description of the contents of Table 1, is given by

$$a : b :: c : d = ((a \equiv b) \wedge (c \equiv d)) \vee ((a \equiv c) \wedge (b \equiv d))$$

Beyond the three postulates P1, P2, P3 and their consequences, the minimal Boolean model of an analogical proportion (described in Table 1) also satisfies two noticeable properties [48]:

- *code independence*: $a : b :: c : d \Rightarrow \neg a : \neg b :: \neg c : \neg d$;

[5] It also says that "when a and d differ (one is true, the other is false) then b and c also differ"! Rewriting it as $a : b :: c : d = ((a \wedge d) \equiv (b \wedge c)) \wedge ((a \vee d) \equiv (b \vee c))$, emphasizes that the conjunctions of the extremes and of the means are equivalent, as well as their disjunctions.

- *transitivity*: $(a : b :: c : d) \wedge (c : d :: e : f) \Rightarrow a : b :: e : f$.

The first property expresses that any feature can be encoded positively or negatively without harming the analogical proportion.

The second property that expresses transitivity does not follow from the postulates either, but is true for Boolean variables.[6]. It has an important consequence (together with reflexivity and symmetry properties): The analogical proportion $a : b :: c : d$ defines *an equivalence relation* between the ordered pairs (a, b) and (c, d) in the Boolean setting.

The description of items may involve *nominal* attributes, i.e., attributes with a finite domain $\mathcal{A}$ whose cardinality is larger than 2. Then $a : b :: c : d$ holds for nominal variables if and only if (as first suggested in [46])

$$(a, b, c, d) \in \{(s, s, s, s), (s, t, s, t), (s, s, t, t) \mid s, t \in \mathcal{A}\} \tag{3}$$

When the cardinality of $\mathcal{A}$ is equal to 2, we retrieve the Boolean model. As can be checked, (3) is the unique nominal model that satisfies P1, P2, and P3. All the properties discussed above still hold for nominal attributes.

Let us take the example mentioned in the Introduction "a calf is to a cow as a foal is to a mare". The animals there can be described in terms of attributes such as `mammal`, `carnivore`, `young`, `adult`, `ruminant`, `single-toed`, or `family` (the last attribute being nominal). The vector describing each animal is given horizontally in Table 2. We can see vertically that a perfect analogical proportion holds component by component.

	mammal	*carnivore*	*young*	*adult*	*ruminant*	*single-toed*	*family*
`calf`	1	0	1	0	1	0	bovidae
`cow`	1	0	0	1	1	0	bovidae
`foal`	1	0	1	0	0	1	equidae
`mare`	1	0	0	1	0	1	equidae

Table 2: A calf is to a cow as a foal is to a mare

[6] Some readers might object that analogical proportions may not be transitive. In a general context, their observation is valid, and this issue becomes more apparent when dealing with multiple attributes. Specifically $\vec{a} : \vec{b} :: \vec{c} : \vec{d}$ may hold with respect to some attributes and $\vec{c} : \vec{d} :: \vec{e} : \vec{f}$ may hold with respect to a different set of attributes leading to a failure of transitivity, as in the following abstract example. Assume $\vec{a}, \vec{b}, \vec{c}, \vec{d}, \vec{e}, \vec{f}$ can be described in terms of 4 Boolean attributes i_1, i_2, i_3, i_4, and $\vec{a} = (1, 1, 0, 0)$, $\vec{b} = (1, 1, 1, 0)$, $\vec{c} = (1, 0, 0, 0)$, $\vec{d} = (1, 0, 1, 1)$, $\vec{e} = (1, 1, 1, 0)$, and $\vec{f} = (1, 1, 1, 1)$. Let us denote by $(\vec{a} : \vec{b} :: \vec{c} : \vec{d})_S$ the fact that the analogical proportion holds componentwise for all attributes $i \in S$. Then it can be easily checked that $(\vec{a} : \vec{b} :: \vec{c} : \vec{d})_{\{i_1,i_2,i_3\}}$ holds as well as $(\vec{c} : \vec{d} :: \vec{e} : \vec{f})_{\{i_1,i_2,i_4\}}$, while $(\vec{a} : \vec{b} :: \vec{e} : \vec{f})_{\{i_1,i_2,i_3,i_4\}}$ does not hold. Still, it can be observed that here transitivity is preserved if we restrict ourselves to the set $S = \{i_1, i_2\}$.

Note that as soon as $\vec{a}, \vec{b}, \vec{c}, \vec{d}$ have at least two components, $\vec{a} : \vec{b} :: \vec{c} : \vec{d}$ can hold with 4 distinct vectors: for instance, in dimension 2, $\vec{a} = (0,0), \vec{b} = (0,1), \vec{c} = (1,0), \vec{d} = (1,1)$ build a proper analogical proportion.

3.3 Analogical proportions in terms of sets

As already said, the entities or items involved in analogical proportions that we consider are represented by vectors of Boolean values, such as $\vec{a} = (a_1, ..., a_n)$, and analogical proportions between vectors are defined componentwise:

$$\vec{a} : \vec{b} :: \vec{c} : \vec{d} \Leftrightarrow \forall i \in \{1, \cdots, n\},\ a_i : b_i :: c_i : d_i$$

To a vector $\vec{a} = (a_1, ..., a_n)$ where $\forall i, a_i \in \{0,1\}$, one can associate the set of features $A = \{i \in \{1, \cdots, n\} \mid a_i = 1\}$ possessed by $\vec{a}$. Clearly, A and $\vec{a}$ are equivalent representations. In the same way, we associate $\vec{b}$, $\vec{c}$ and $\vec{d}$ with B, C and D. The set representation provides a maybe more intuitive view of an analogical proportion.

Indeed, it has been noticed [56] (see also [3]) that $A : B :: C : D$ holds if and only if it exists *non overlapping* subsets U, V, X, Y, and Z, such that

- $A = U \cup X \cup Z$;
- $B = U \cup Y \cup Z$;
- $C = V \cup X \cup Z$;
- $D = V \cup Y \cup Z$.

This makes clear that $A \setminus B = C \setminus D = X$ and $B \setminus A = D \setminus C = Y$, which is in agreement with equation (1). Note also that $A \setminus C = B \setminus D = U$ and $C \setminus A = D \setminus B = V$ in agreement with the stability under central permutation.[7]

Moreover, with this set representation, it can be easily checked that an analogical proportion $A : B :: C : D$ holds as soon as

A: made of what is common to A and C together with what is common to A and B
B: made of what is common to B and D together with what is common to A and B
C: made of what is common to A and C together with what is common to C and D
D: made of what is common to B and D together with what is common to C and D.

[7] This is also in agreement with equation (2) since $A \cap D = B \cap C = Z$ and $A \cup D = B \cup C = U \cup V \cup X \cup Y \cup Z$.

where ‘made of’ refers to union operation and “what is common” to intersection operation. Interestingly enough, the four descriptions of A, B, C, D above make a perfect analogical proportion, term by term (applying nominal definition (3))!

Remark. The set-based view of analogical proportion allows us to better suggest a relationship between Mill’s rules of induction and analogy. If we reconsider Mill’s pattern of (Direct) Method of Agreement as the analogical proportion (keeping the notations of Section 2) “$\mathcal{A} = ABCD$ is to $\mathcal{B} = wxyz$ as $\mathcal{C} = AEFG$ is to $\mathcal{D} = wtuv$” (where ‘is to’ is understood as ‘co-occurs with’), we can observe that indeed $\mathcal{A} \setminus \mathcal{C} = \mathcal{C} \setminus \mathcal{A} = A$ co-occurs with $\mathcal{B} \setminus \mathcal{D} = \mathcal{D} \setminus \mathcal{B} = w$. Mill’s Method of Difference could be read similarly in an analogical proportion manner in spite of the presence of items of two different natures in those patterns.

3.4 Analogical proportions as equivalence relations between pairs

Analogical proportions are a matter of i) comparing items inside an ordered pair, and then. ii) pairing pairs $(\vec{a}, \vec{b})$ and $(\vec{c}, \vec{d})$. Let us examine these two steps.

Let $\vec{a} = (a_1, ..., a_n)$, $\vec{b} = (b_1, ..., b_n)$, etc. be items described by means of n Boolean attributes or features. Given two vectors $\vec{a}$, $\vec{b}$, their comparison leads to consider the subsets of attributes where they are equal (to 1 or to 0), and the subsets of attributes where they differ (by going from 0 to 1, or from 1 to 0), when we go from $\vec{a}$ to $\vec{b}$. This leads to define

$$Equ^0(\vec{a}, \vec{b}) = \{i \mid a_i = b_i = 0\},$$
$$Equ^1(\vec{a}, \vec{b}) = \{i \mid a_i = b_i = 1\},$$
$$Equ(\vec{a}, \vec{b}) = \{i \mid a_i = b_i\} = Equ^0(\vec{a}, \vec{b}) \cup Equ^1(\vec{a}, \vec{b}),$$

and

$$Dif^{10}(\vec{a}, \vec{b}) = \{i \mid a_i = 1, b_i = 0\},$$
$$Dif^{01}(\vec{a}, \vec{b}) = \{i \mid a_i = 0, b_i = 1\},$$
$$Dif(\vec{a}, \vec{b}) = \{i \mid a_i \neq b_i\} = Dif^{01}(\vec{a}, \vec{b}) \cup Dif^{10}(\vec{a}, \vec{b}).$$

This allows us to state the following result:

$$\vec{a} : \vec{b} :: \vec{c} : \vec{d} \text{ if and only if } \begin{cases} Equ(\vec{a}, \vec{b}) = Equ(\vec{c}, \vec{d}) \\ Dif^{10}(\vec{a}, \vec{b}) = Dif^{10}(\vec{c}, \vec{d}) \\ Dif^{01}(\vec{a}, \vec{b}) = Dif^{01}(\vec{c}, \vec{d}) \end{cases}$$

We see that what matters in an analogical proportion is the orientation of the differences, whereas it does not matter with which value the equality is realized. Table 3 highlights the structure of an analogical proportion, in three subsets of attribute(s), one where the 4 items are equal, one where they are equal within the pairs, but not in the same way, and finally the subset of attribute(s) whose value(s) change(s), in the same direction, from $\vec{a}$ to $\vec{b}$ and from $\vec{c}$ to $\vec{d}$.

$items$	$All\ equal$	$Equality\ by\ pairs$	$Change\ (Dif)$
$\vec{a}$	1 0	1 0	1 0
$\vec{b}$	1 0	1 0	0 1
$\vec{c}$	1 0	0 1	1 0
$\vec{d}$	1 0	0 1	0 1

Table 3: The 3 parts of analogical proportion and the associated valuations

As shown in Table 3, the set of attributes with which the four items involved in an analogical proportion $a : b :: c : d$ are supposed to be represented can be partitioned in three subsets corresponding to the way the attribute values are possibly modified from an item to another item. As we can see, the central permutation of $\vec{b}$ and $\vec{c}$ exchanges the content of the columns “Equality by pairs" and “Change" (but does not affect the “All equal" column). Neither of these two subsets must be empty if we want the analogical proportion to be non-trivial, i.e., $\vec{a}$, $\vec{b}$, $\vec{c}$, $\vec{d}$ are distinct vectors (for $n = 2$, $\vec{a} = (1, 1)$, $\vec{b} = (1, 0)$, $\vec{c} = (0, 1)$, $\vec{d} = (0, 0)$ realize an analogical proportion with distinct vectors, as already said). Besides, the subset of attribute(s) “All equal” can be empty. If the subset “Equality by pairs” or the subset “Change” is empty, then $\vec{a} = \vec{c}$ and $\vec{b} = \vec{d}$ or $\vec{a} = \vec{b}$ and $\vec{c} = \vec{d}$ respectively.

Besides, analogical proportions are implicitly present when comparing two items represented with the same set of n attributes ($n \geq 2$): From a formal viewpoint, for any two distinct vectors $\vec{a}$ and $\vec{d}$ *differing on at least two attributes*, there exist two other distinct vectors $\vec{b}$ and $\vec{c}$ such as $\vec{a} : \vec{b} :: \vec{c} : \vec{d}$ [12]. This does not mean that these two vectors $\vec{b}$ and $\vec{c}$ represent items existing in the real world.

As already noticed when dealing with one component, analogical proportions express an equivalence relation between two ordered pairs. Given four distinct vectors $\vec{a}, \vec{b}, \vec{c}, \vec{d}$, we have the following result:

Two ordered pairs $(\vec{a}, \vec{b})$ and $(\vec{c}, \vec{d})$ are in the same equivalence class if and only if[8]

1. $Dif(\vec{a}, \vec{b}) = Dif(\vec{c}, \vec{d})$;
2. $\forall j \in Dif(\vec{a}, \vec{b})\ a_j = c_j$ and $b_j = d_j$.

Condition 1 ensures that the change concerns the same attributes in both pairs, condition 2 that the change applies in the same direction in both pairs. It is clear that any two pairs $(\vec{a}, \vec{b})$ and $(\vec{c}, \vec{d})$ taken in the same equivalence class together form an analogical proportion

[8] A further condition should be added, namely $Dif(\vec{a}, \vec{b}) \neq \emptyset$ and $\exists i\ a_i \neq c_i$ in case the vectors might not be distinct.

$\vec{a} : \vec{b} :: \vec{c} : \vec{d}$. This notion of equivalence class is similar to the idea of "analogical cluster" introduced in [34] in a context of computational linguistics.

4 Analogical proportions-based inference and classification

In the Boolean and nominal cases, analogical inference relies on the solving of analogical equations, i.e. finding $\vec{x}$ such that $\vec{a} : \vec{b} :: \vec{c} : \vec{x}$ holds, working component by component.[9]. This kind of extrapolation is a counterpart of the "rule of three" based on geometric proportions $x = \frac{b \cdot c}{a}$ (its arithmetic counterpart is $x = b + c - a$).

Since a triplet $a, b, c \in \{0, 1\}^3$ may take $2^3 = 8$ values, while $a : b :: c : d$ is true only for six distinct 4-tuples, there are cases where the equation $a : b :: c : x$ in the Boolean case has no solution. Indeed, the equations $1 : 0 :: 0 : x$ and $0 : 1 :: 1 : x$ have no solution. It is easy to prove that the Boolean analogical equation $a : b :: c : x$ is solvable if and only if $(a \equiv b) \vee (a \equiv c)$ holds true. In that case, the *unique* solution is given by $x = a \equiv (b \equiv c)$; thus $x = b$ if $a = c$ and $x = c$ if $a = b$.

The situation in the nominal case is quite similar: $s : t :: t : x$ has no solution (for $s \neq t$). Only the equations $s : t :: s : x$ and $s : s :: t : x$ are solvable, with unique solution $x = t$. In the nominal case, where s, t, u can take more than 2 values, the equation $s : t :: u : x$ is also not solvable as soon as s, t, u are distinct.

Analogical proportion-based inference, as described by the inference rule (4), applies to classification and relies on a simple principle: in the Boolean or nominal cases, if four vectors $\vec{a}$, $\vec{b}$, $\vec{c}$ and $\vec{d}$ make a valid analogical proportion component-wise for each 4-tuple of values pertaining to the same attribute, then it is expected that their class labels also make a valid proportion ([58], see also [9]).

$$\frac{\vec{a} : \vec{b} :: \vec{c} : \vec{d}}{cl(\vec{a}) : cl(\vec{b}) :: cl(\vec{c}) : cl(\vec{d})} \tag{4}$$

Assuming that the class labels for vectors $\vec{a}$, $\vec{b}$ and $\vec{c}$ are known (i.e., they belong to the sample set), the classification of a new Boolean or nominal vector $\vec{d}$ is only possible i) when the equation $cl(\vec{a}) : cl(\vec{b}) :: cl(\vec{c}) : x$ is solvable (since a Boolean or a nominal equation may have no solution if the equation is of the form $s : t :: t : x$ or $s : t :: u : x$), and ii) the analogical proportion $\vec{a} : \vec{b} :: \vec{c} : \vec{d}$ holds true on all components. If these two conditions are met, we take $cl(\vec{d})$ as the unique solution for x.

Clearly, the inference rule (4) offers no guarantee on the truth of the prediction for $cl(\vec{d})$; this prediction may be only regarded as a plausible conclusion, just as in the case of

[9] The three other equations $\vec{x} : \vec{b} :: \vec{c} : \vec{d}$, $\vec{a} : \vec{x} :: \vec{c} : \vec{d}$, $\vec{a} : \vec{b} :: \vec{x} : \vec{d}$ can be equivalently stated in the above form (with x in d position), applying internal and complete reversal properties.

the "analogical jump" inference (recalled in the introduction), of which (4) can be shown to a particular instance [9]. Note that this inference rule is a transduction rule that infers a factual conclusion about the class of a new item from a set of items and their class. It is an extrapolation process [35].

However as already said, there is no guarantee that the conclusion of the inference rule is not erroneous, and it may happen that different predictions coexist for $cl(\vec{d})$ as shown in Table 4, where three pairs $(\vec{a}, \vec{b})$, $(\vec{c}, \vec{d})$, $(\vec{a'}, \vec{b'})$, belonging to the same equivalence class (since each pair has the same Dif-pattern) are put in parallel leading to divide the equality part of the pairs into four subparts $Equ_{sss}, Equ_{sst}, Equ_{sts}, Equ_{stt}$ corresponding to the different possible arrangements of these equality parts while preserving analogical proportions between the first two pairs and between the last two.[10] Indeed, on the one hand, one can check that $\vec{a} : \vec{b} :: \vec{c} : \vec{d}$ holds, and the analogical inference yields $cl(\vec{d}) = cl(\vec{b}) = t$; on the other hand, it can be seen that $\vec{a'} : \vec{b'} :: \vec{c} : \vec{d}$ holds also, and the analogical inference then gives $cl(\vec{d}) = cl(\vec{c}) = s$. This takes place in spite of the fact the pairs $(\vec{a}, \vec{b})$ and $(\vec{a'}, \vec{b'})$ are in the same class of equivalence.

$case$	Equ_{sss}	Equ_{sst}	Equ_{sts}	Equ_{stt}	Dif	$class$
$\vec{a}$	s_1	s_2	s_3	s_4	s_5	s
$\vec{b}$	s_1	s_2	s_3	s_4	t_5	t
$\vec{c}$	s_1	s_2	t_3	t_4	s_5	s
$\vec{d}$	s_1	s_2	t_3	t_4	t_5	t / s
$\vec{a'}$	s_1	t_2	s_3	t_4	s_5	s
$\vec{b'}$	s_1	t_2	s_3	t_4	t_5	s

Table 4: Inconsistent prediction

This situation of inconsistent predictions is very general. The unique exception (where the situation is impossible) is when the classification function is a linear function in case of Boolean attributes [11], or is quasi-linear in case of nominal attributes [13]. Still this does not mean that those special cases are the only cases where the analogical proportion-based inference can be used. Indeed, even if there is some inconsistent predictions, one may just retain the prediction made by the majority of the triplets.

Thus, the brute-force algorithm consists in looking for all triplets $(\vec{a}, \vec{b}, \vec{c})$ for which the corresponding analogical equation on class is solvable, and which makes an analogical proportion with the item $\vec{d}$ for which one wants to predict the class. This has a clear cubic complexity, and is costly. However, the accuracy results are good enough on real data

[10] In Table 4, each of the 5 columns from Equ_{sss} to Dif stands for a subset of attributes and the corresponding s_k is here a sub-vector of Boolean or nominal attribute values.

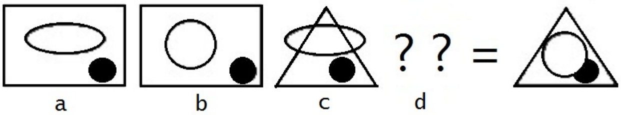


Figure 1: Example of a visual analogical proportion to be completed

benchmarks to be compared with other more classical approaches [38, 9]. It is worth noticing that it is possible to make an offline pre-compilation of pairs, to focus on pairs which are differing on a few attributes only, and to choose $\vec{c}$ as a close neighbor of $\vec{d}$, without harm for the accuracy results [9].

Moreover, a natural idea for restricting the number of triplets is to look only for those triplets involving "competent" pairs. Competent pairs are those in an equivalence class whose pattern for the class is in the majority. For instance, in Table 4, the pairs $(\vec{a}, \vec{b})$ and $(\vec{a'}, \vec{b'})$ are in the the same equivalence class ; their respective pattern for the class are (s, t) and (s, s); the pairs whose pattern for the class is is in the majority in the equivalence class will be considered as "competent" for extrapolating $cl(\vec{d})$ on the basis of such a pair and a $\vec{c}$ forming an analogical proportion $\vec{a} : \vec{b} :: \vec{c} : \vec{d}$ with $\vec{d}$. The use of competent pairs has proved to be experimentally interesting in terms of accuracy and computational cost [35, 21].

5 Towards a logic of creativity

Analogical reasoning has long had a reputation for creativity [47, 17, 29]. First, let us explain in what limited sense we can talk about creativity in the case of analogical proportion-based inference.

For doing that, we use a simple IQ test-like example[11]. The problem is to find, among a given set of candidate solutions, the figure X that gives the best fit to the analogical proportion "A is to B as C is to X". This kind of problem was successfully addressed very early in artificial intelligence by Thomas Evans [23].

In the following, we show that it can be solved *without the knowledge of candidate solutions*. Let us consider the example of Figure 1. The figures in this example of Figure 1 can be encoded with 5 Boolean predicates, namely $hasRectangle(hR)$, $hasBlackDot(hBD)$,

[11] More difficult tests, like Raven IQ tests where a series of instances having the format of a 3×3 matrix whose cells contain diverse geometric figures, except the last cell which is empty and has to be completed by selecting a solution among 8 candidates, can be also solved by analogical proportion-based reasoning without the help of candidate solutions [10].

	hR	hBD	hT	hC	hE
$\boldsymbol{a}$	1	1	0	0	1
$\boldsymbol{b}$	1	1	0	1	0
$\boldsymbol{c}$	0	1	1	0	1
$\boldsymbol{x}$	?	?	?	?	?

Table 5: Solving the example: $\boldsymbol{x} = (0, 1, 1, 1, 0)$

$hasTriangle(hT)$, $hasCircle(hC)$, and $hasEllipse(hE)$. They appear in that order in Table 5, where the example is encoded.

It can be observed that the description of figure x, namely, $hR = 0$, $hBD = 1$, $hT = 1$, $hC = 1$, $hE = 0$ can be obtained by solving the analogical equation in each column, which in this example, has a solution for each feature. Thus, from three items $\vec{a}$, $\vec{b}$, $\vec{c}$, we are able to build (*create*!) a *novel* item $\vec{d}$ different from the three others. Since in the above example we have not encoded the position of the figures in a picture, each figure could be drawn inside, outside or intersecting the other ones. We could also take care of the positions with respect to, e.g., the basis of the rectangles and triangles, by using more attributes. Thus, for instance, the Black Dot will remain at the same place in the different figures. But what is important in the handling of such riddles is to use independent attributes in the representation; but simple relationships, such as "the black dot is outside the ellipse", "the black dot is inside the rectangle" could be also coded directly; see [10] for more discussions.

This form of creativity can be summarized by the following inference pattern (remember $\equiv$ is associative), where $\vec{a}, \vec{b}, \vec{c}$ are vectors of Boolean attribute values defined on the same set of attributes:

$$\frac{\vec{a}, \vec{b}, \vec{c}}{\vec{d} = (\vec{a} \equiv \vec{b} \equiv \vec{c})} \tag{5}$$

This pattern of inference has been considered a fundamental element of human creativity [32]. However, note that we consider this pattern as *valid only if* on each feature i the Boolean analogical equation $a_i : b_i :: c_i : x_i$ is solvable, which requires that

$$(\vec{a} \equiv \vec{b}) \vee (\vec{a} \equiv \vec{c})$$

holds true for each vector component.

Remark 1. *Strictly speaking, we might accept the inference pattern (5) without any restriction, since $\vec{a} \equiv \vec{b} \equiv \vec{c}$ is always defined. Such a view was defended by S. Klein [31] who was a forerunner of the Boolean modeling of analogical proportions used here. But* *this leads to debatable consequences. Indeed Klein's view of analogy was deeply influenced by his anthropological interest in cultural devices such as Navaho sand paintings or mandalas [31]. Such paintings upon the ground have a square structure and can be contemplated from any side (there is no top or bottom); it makes natural a property such as* $\vec{a} : \vec{b} :: \vec{c} : \vec{d} \Rightarrow \vec{b} : \vec{c} :: \vec{d} : \vec{a}$*, where* $\vec{a}, \vec{b}, \vec{c}, \vec{d}$ *refer to the descriptions of the four corners of a sand painting. If we iterate the property, namely,* $\vec{b} : \vec{c} :: \vec{d} : \vec{a} \Rightarrow \vec{c} : \vec{d} :: \vec{a} : \vec{b}$*, we see it entails symmetry. But while this property preserves* $0 : 0 :: 0 : 0$ *and* $1 : 1 :: 1 : 1$*, exchanges* $0 : 1 :: 0 : 1$ *and* $1 : 0 :: 1 : 0$*, it changes* $0 : 0 :: 1 : 1$ *into* $0 : 1 :: 1 : 0$ *and* $1 : 1 :: 0 : 0$ *into* $1 : 0 :: 0 : 1$*, thus introducing two patterns excluded by the condition* $(\vec{a} \equiv \vec{b}) \vee (\vec{a} \equiv \vec{c})$*. This latter condition preserves strict analogical proportions (otherwise the undesirable property* $\vec{a} : \vec{b} :: \vec{c} : \vec{d} \Rightarrow \vec{b} : \vec{a} :: \vec{c} : \vec{d}$ *would hold, and break the oriented nature of the comparisons inside the analogical proportion).*

Our aim in the following is to investigate what consequence relation could be defined between ordered pairs. This relation, once symmetrized, must give rise to an equivalence relation between ordered pairs, which must therefore be reflexive, symmetrical and transitive. In a Boolean framework, such a relation corresponds to a logical connector between four variables (two per pair). Analogical proportions are a particular case of logical proportions. This is why we first start by a short journey among logical proportions [48] in the next subsection, looking for those proportions that are reflexive, symmetrical and transitive when considered as operators on pairs. The contents of the rest of this section expands the presentation of a logic outlined in [53].

5.1 Logical proportions

The logical proportions [48] offer a framework, in propositional logic, of quaternary connectors expressing relations between pairs. It is from this Boolean framework, the essence of which we now recall, that we start our investigations.

In the Boolean framework, we have four comparison indicators to relate two variables a to b.

- Two indicators express *similarity*, either *positively* as $a \wedge b$ (which is true if a and b are true), or *negatively* as $\neg a \wedge \neg b$ (which is true if a and b are false).
- The other two are indicators of *dissimilarity* $\neg a \wedge b$ (which is true if a is false and b is true) and $a \wedge \neg b$ (which is true if a is true and b is false).

Definition 1. *A logical proportion* $T(a, b, c, d)$ *is the conjunction of two equivalences between an indicator for* (a, b) *and an indicator for* (c, d)*.*

The expression $\qquad ((a \wedge \neg b) \equiv (c \wedge \neg d)) \wedge ((a \wedge b) \equiv (c \wedge d))$

provides an example of a logical proportion, where the same dissimilarity operator and the same similarity operator are applied to both pairs. As can be seen, it expresses that "a differs from b as c differs from d" and that "a is similar to b as c is similar to d". It seems to refer to the comparison of the elements within each ordered pair, but we shall see that this is not in the sense of an analogical proportion.

It has been established [48] that there are 120 syntactically and semantically distinct logical proportions. Because of the way they are built, all these proportions share a remarkable property: They are true for exactly 6 patterns of $abcd$ values among $2^4 = 16$ candidate patterns. For instance, the above proportion is true for 0000, 1111, 1010, 0101, 0001, and 0100. The interested reader is invited to consult [48, 49] for in-depth studies of the different types of logical proportions.

Among all 120 logical proportions T, only 6 are reflexive (i.e., $T(a, b, a, b)$ holds true) [48].

Proposition 1. *Only 6 logical proportions are reflexive. They are*
- the analogical proportion

$$A(a, b, c, d) = a : b :: c : d = ((a \wedge \neg b) \equiv (c \wedge \neg d)) \wedge ((\neg a \wedge b) \equiv (\neg c \wedge d))$$

- the so-called paralogy
$P(a, b, c, d) = ((a \wedge b) \equiv (c \wedge d)) \wedge ((\neg a \wedge \neg b) \equiv (\neg c \wedge \neg d))$
- and the 4 following conditional logical proportions
$((a \wedge b) \equiv (c \wedge d)) \wedge ((a \wedge \neg b) \equiv (c \wedge \neg d))$ *;*
$((a \wedge b) \equiv (c \wedge d)) \wedge ((\neg a \wedge b) \equiv (\neg c \wedge d))$ *;*
$((a \wedge \neg b) \equiv (c \wedge \neg d)) \wedge ((\neg a \wedge \neg b) \equiv (\neg c \wedge \neg d))$ *;*
$((\neg a \wedge b) \equiv (\neg c \wedge d)) \wedge ((\neg a \wedge \neg b) \equiv (\neg c \wedge \neg d))$.

The paralogy expresses that "what a and b have in common (positively or negatively), c and d also have, and vice versa". This proportion is (only) true for 0000, 1111, 1010, 0101, 1001, and 0110. It can be checked that P$(a, b, c, d)) \Leftrightarrow$ A(c, b, a, d). It has been already mentioned in subsection 3.2 that A(a, b, c, d) is equivalent to P(a, d, b, c). The reason of the name "conditional logical proportions" will appear in next subsection.

These 6 logical proportions are also among the 12 logical proportions that are symmetrical ($T(a, b, c, d) \Leftrightarrow T(c, d, a, b)$) and among the 54 logical proportions that are transitive ($T(a, b, c, d), T(c, d, e, f) \Rightarrow T(a, b, e, f)$) [48].

Proposition 2. *A, P and the 4 conditional logical proportions of Proposition 1 are the only logical proportions that define equivalence relations between ordered pairs.*

It should be also clear that while the 6 above logical proportions satisfy the first two postulates P1 and P2 (reflexivity and symmetry), only the analogical proportion is stable under central permutation (i.e., satisfies P3). The paralogy $\mathrm{P}(a, b, c, d))$ is stable under the permutation of the first two items (or the last two, due to symmetry), i.e., we have $\mathrm{P}(a, b, c, d)) \Leftrightarrow \mathrm{P}(b, a, c, d))$. The last 4 logical proportions are not stable for any permutation of two items. This shows that these logical proportions are quite different and serve different purposes.

Indeed it turns out that this result covers two important cases:

- the logic of conditional events which is a basis of non monotonic reasoning, recalled in the next subsection, and which will also be a source of inspiration for the rest of the paper,

- a logic of ordered pairs preserving positive changes, which may contribute to a "controlled" creativity process, outlined in the rest of the section.

We first recall the logic associated with the conditional logical proportions, since the way a logic is associated with these proportions will guide us for building a logic associated with the analogical proportion, and another one associated with the paralogy proportion (which will be only briefly mentioned, since out of the analogical scope of the paper).

5.2 Conditional events as a basis of non monotonic reasoning

Let us consider the 4 conditional proportions which are related to our subject, since they are reflexive, symmetrical and transitive. Let us first explain the word "conditional". It comes from the fact that these proportions express equivalences between conditional statements. Indeed, it was pointed out in [18] that a rule "if a then b" can be considered as a three-valued entity referred as a "conditional object" or a "conditional event", and denoted $b|a$. This entity is *tri-valued* [15] as follows:

- $b|a$ is true if $a \wedge b$ is true. The elements which make true $a \wedge b$ are the *examples* of the rule "if a then b";

- $b|a$ is false if $a \wedge \neg b$ is true. The elements which make true $a \wedge \neg b$ are the *counter-examples* of the rule "if a then b";

- $b|a$ is undefined if a is false. The rule "if a then b" is then not applicable.

Consider the conditional proportion appearing in Proposition 1 and which was our first example of a logical proportion:

$$((a \wedge b) \equiv (c \wedge d)) \wedge ((a \wedge \neg b) \equiv (c \wedge \neg d)) \qquad (6)$$

The above logical proportion can then be denoted $b|a :: d|c$ by combining the notation of conditional objects with that of the analogical proportion. Indeed, the proportion $b|a :: d|c$

expresses a semantic equivalence between the two rules “if a then b” and “if c then d” by stating that:

- they have the same examples, i.e., $(a \wedge b) \equiv (c \wedge d)$;
- they have the same counter-examples, i.e., $(a \wedge \neg b) \equiv (c \wedge \neg d)$;
- if $b|a$ is not applicable, i.e., a is false, then necessarily c is false (otherwise (6) would be false), which means that $d|c$ is not applicable.

The logical consequence relation between conditional objects $b|a \vDash d|c$ is defined as [15]:

$$a \wedge b \vDash c \wedge d \text{ and } c \wedge \neg d \vDash a \wedge \neg b \tag{7}$$

which expresses that the examples of the first conditional object are examples of the second one, and the counter-examples of the second conditional object are counter-examples of the first one. This entailment is naturally associated with the conditional proportion $b|a :: d|c$, since

$$b|a :: d|c \Leftrightarrow b|a \vDash d|c \text{ and } d|c \vDash b|a.$$

The transitivity of the 4 conditional proportions of the Proposition 1 reflects the fact that they express equivalences between conditional objects (and thus between rules), namely respectively $b|a :: d|c$, $a|b :: c|d$, $a|\neg b :: c|\neg d$, and $b|\neg a :: d|\neg c$.

The conditional object $b|a$ must therefore be thought of as a rule “if a then b". A rule may have exceptions. That is, we can have at the same time the rule “if a then b” and a rule “if $(a \wedge c)$ then $\neg b$”. The two conditional objects $b|a$ and $\neg b|(a \wedge c)$ do not lead to a contradiction in the presence of the facts a and c (unlike a modeling of rules by material implication), in the setting of a tri-valued logic where the conjunction & is defined by [18]:

$$b|a \,\&\, d|c \triangleq ((a \rightarrow b) \wedge (c \rightarrow d))|(a \vee c)$$

where $\rightarrow$ is the material implication ($a \rightarrow b \triangleq \neg a \vee b$) and with the following semantics: $val(o_1 \& o_2) = \min(val(o_1), val(o_2))$ where the three truth values are ordered as follows: undefined $>$ true $>$ false.[12]

It can be shown that this quasi-conjunction ‘&’ (that is its name) is associative. It expresses that the set constituted by the two rules “if a then b” and “if c then d” is triggerable if a or c is true, and in this case the triggered rule behaves like the material implication. This logic constitutes the simplest semantics [6] of the system P of non-monotonic inference of Kraus, Lehmann, and Magidor [33]. The reader can consult [18, 6] for more details.

[12] The negation is defined by $\neg(b|a) = (\neg b|a)$; $\neg(b|a)$ is undefined if and only if $b|a$ is.

Although nonmonotonic reasoning only yields plausible conclusions, it is not clear that it might be considered as a special form of induction. Indeed nonmonotonic reasoning is here a two steps-process. First from a set of conditional events representing a set of default rules, we *deduce* a new conditional event whose condition part corresponds exactly to our knowledge of the current situation under consideration, and then - second step - we apply the new default rule thus inferred to the current situation.

We now consider the two other logical proportions that define an equivalence relation between ordered pairs, namely the analogical proportion and the paralogy, and we try to identify what consequence relations can be associated with them.

5.3 A new consequence relation between ordered pairs

In the following subsections, we try to identify some elements of a comparative logic of ordered pairs. The items to be compared are described by vectors of attribute values (here Boolean). When $a_i = 1$ (resp. $a_i = 0$) we understand it as item $\vec{a}$ has (resp. has not) feature / property i.

As usual, logical connectives extend to vectors componentwise:
1. $\neg\vec{a} = (\neg a_1, ..., \neg a_n)$;
2. $\vec{a} \wedge \vec{b} = (a_1 \wedge b_1, ..., a_n \wedge b_n)$;
3. $\vec{a} \vee \vec{b} = (a_1 \vee b_1, ..., a_n \vee b_n)$.

Taking inspiration from the case of conditional logical proportions (namely definition Definition 7), we are led to define the following, new logical consequence relation between pairs (still denoted $\vDash$) from the definition of an analogical proportion:

$$(\vec{a}, \vec{b}) \vDash (\vec{c}, \vec{d}) \Leftrightarrow \neg\vec{a} \wedge \vec{b} \vDash \neg\vec{c} \wedge \vec{d} \text{ and } \vec{c} \wedge \neg\vec{d} \vDash \vec{a} \wedge \neg\vec{b} \tag{8}$$

When we deal with pairs, the valuation $(a_i, b_i) = (0, 1)$ can be understood as when we go from $\vec{a}$ to $\vec{b}$, we acquire feature i. Thus the meaning of entailment (8) is the following: features that are acquired when going from $\vec{a}$ to $\vec{b}$ remain acquired when going from $\vec{c}$ to $\vec{d}$. Moreover if when going from $\vec{c}$ to $\vec{d}$ a feature is lost, it was already the case when going from $\vec{a}$ to $\vec{b}$.[13]

Proposition 3. *We have the following equivalence:*

$$(\vec{a}, \vec{b}) \vDash (\vec{c}, \vec{d}) \textit{ and } (\vec{c}, \vec{d}) \vDash (\vec{a}, \vec{b}) \textit{ iff } A(\vec{a}, \vec{b}, \vec{c}, \vec{d})$$

[13] The choice of definition (8), rather than $(\vec{a}, \vec{b}) \vDash (\vec{c}, \vec{d}) \Leftrightarrow \vec{a} \wedge \neg\vec{b} \vDash \vec{c} \wedge \neg\vec{d}$ and $\neg\vec{c} \wedge \vec{d} \vDash \neg\vec{a} \wedge \vec{b}$, is governed by the need here to privilege the acquisition of features rather than their loss. Indeed the alternative definition given in this footnote says that features that are lost when going from $\vec{a}$ to $\vec{b}$ remain lost when going from $\vec{c}$ to $\vec{d}$, and that if when going from $\vec{c}$ to $\vec{d}$ a feature is acquired, it was already the case when going from $\vec{a}$ to $\vec{b}$.

Proof: Let us see the precise meaning of this definition for pairs. Because we are working componentwise, it is enough to consider the consequence of this definition on one component. Two cases have to be considered:

- Case $a = b$ (representing 8 valuations among the 16 candidates for a, b, c, d). Because $\neg a \wedge b$ and $a \wedge \neg b$ are 0, the only constraint is that $c \wedge \neg d = 0$ which is valid only if $(c, d) \neq (1, 0)$, eliminating (0010) and (1110) as valid valuations, leaving 6 valuations still valid.

- Case $a \neq b$ (representing the 8 remaining valuations): if $(a, b) = (1, 0)$, there is no constraint on (c, d). If $(a, b) = (0, 1)$, only $(c, d) = (0, 1)$ is valid eliminating 3 valuations among the 8: $(0100), (0110), (0111)$

Having both $(\vec{a}, \vec{b}) \models (\vec{c}, \vec{d})$ and $(\vec{c}, \vec{d}) \models (\vec{a}, \vec{b})$ leads to the truth table of $A(a, b, c, d)$ with exactly 6 valid valuations. □

Because when $(\vec{a}, \vec{b}) \models (\vec{c}, \vec{d})$, the 5 valuations (0010), $(1110), (0100)$, (0110), (0111) are *forbidden* for each $(a_i\ b_i\ c_i\ d_i)$, this means that

- $(a_i, b_i) = (0, 1) \Rightarrow (c_i, d_i) = (0, 1)$; (a property acquired from $\vec{a}$ to $\vec{b}$ has to be acquired from $\vec{c}$ to $\vec{d}$)

- $a_i = b_i \Rightarrow (c_i, d_i) \neq (1, 0)$ (when there is no acquisition or loss from $\vec{a}$ to $\vec{b}$, there cannot be a loss from $\vec{c}$ to $\vec{d}$)

Similarly, we have $(\vec{c}, \vec{d}) \models (\vec{a}, \vec{b}) \Leftrightarrow \begin{cases} (a_i, b_i) = (1, 0) \Rightarrow (c_i, d_i) = (1, 0) \\ a_i = b_i \Rightarrow (c_i, d_i) \neq (0, 1) \end{cases}$

which forbids the 5 valuations (1000), (1001), (1011), (0001), (1101).

Thus we have, as expected, $(\vec{a}, \vec{b}) \models (\vec{c}, \vec{d})$ and $(\vec{c}, \vec{d}) \models (\vec{a}, \vec{b}) \Leftrightarrow A(\vec{a}, \vec{b}, \vec{c}, \vec{d})$.

Table 6 exhibits the situations where the entailments (defined by (8) $(a, b) \models (c, d)$ and $(c, d) \models (a, b)$ are true. The relation $\models$ is a clear weakening of the analogical proportion when viewed as a relation between pairs. To support the intuition of the entailment, let us consider the case where a, b, c, d are just Boolean values. As previously explained:

- $(0, 1)$ has 1 consequence $(0, 1)$,

- $(0, 0)$ has 3 consequences $(0, 0), (0, 1), (1, 1)$,

- $(1, 1)$ has 3 consequences $(0, 0), (0, 1), (1, 1)$,

a	b	c	d	$(a,b) \vDash (c,d)$	$(c,d) \vDash (a,b)$	$a : b :: c : d$
0	0	0	0	1	1	1
0	0	0	1	1	0	0
0	0	1	0	0	1	0
0	0	1	1	1	1	1
0	1	0	0	0	1	0
0	1	0	1	1	1	1
0	1	1	0	0	1	0
0	1	1	1	0	1	0
1	0	0	0	1	0	0
1	0	0	1	1	0	0
1	0	1	0	1	1	1
1	0	1	1	1	0	0
1	1	0	0	1	1	1
1	1	0	1	1	0	0
1	1	1	0	0	1	0
1	1	1	1	1	1	1
nb	of	values	'true'	**11**	**11**	**6**

Table 6: Entailments defined by (8) vs. analogical proportion

- $(1,0)$ has 4 consequences $(0,0),(0,1),(1,0),(1,1)$.

A compact way to put it would be to order the set of Boolean pairs such as $(1,0) < \{(0,0),(1,1)\} < (0,1)$ and to consider that any pair entails all pairs at the same level and higher.

Thus, if we consider a pair $(\vec{a},\vec{b})$ of vectors of dimension 4 where $\vec{a} = (0,0,1,1)$ and $\vec{b} = (0,1,0,1)$, we see that this pair has $3 \times 1 \times 4 \times 3 - 1 = 35$ distinct logical consequences in the sense of $\vDash$ defined by (8).

5.4 Logical combinations of ordered pairs

One may think of defining conjunctive or disjunctive combinations of ordered pairs, but these combinations should agree with the consequence relation (8) and make sense with respect to the interpretation of pairs. Natural componentwise definitions, including negation, seem to be:

$$(\vec{a},\vec{b}) \wedge (\vec{c},\vec{d}) = (\vec{a} \wedge \vec{c}, \vec{b} \wedge \vec{d});$$

$$(\vec{a},\vec{b}) \vee (\vec{c},\vec{d}) = (\vec{a} \vee \vec{c}, \vec{b} \vee \vec{d}).$$

$$\neg(\vec{a}, \vec{b}) = (\neg\vec{a}, \neg\vec{b})$$

Note that $\neg(\vec{a}, \vec{b}) \neq (\vec{b}, \vec{a})$ in general. However, an involutive operation such as

$$\circlearrowleft(\vec{a}, \vec{b}) = (\vec{b}, \vec{a})$$

looks more interesting, as we shall see, since it reverses the order of comparison in the pair. Besides, as a consequence of the above definitions, we have

$$(\vec{a}, \vec{b}) \wedge (\vec{a}, \vec{b}) = (\vec{a}, \vec{b}) = (\vec{a}, \vec{b}) \vee (\vec{a}, \vec{b})$$

But unfortunately one can check that

$$(\vec{a}, \vec{b}) \wedge (\vec{c}, \vec{d}) \not\models (\vec{a}, \vec{b}) \not\models (\vec{a}, \vec{b}) \vee (\vec{c}, \vec{d}).$$

This failure is simply due to the fact that a feature acquired from $\vec{a} \wedge \vec{c}$ to $\vec{b} \wedge \vec{d}$ may not be a feature acquired from $\vec{a}$ to $\vec{b}$. Indeed starting with $(a_i, b_i, c_i, d_i) = (1, 1, 0, 1)$, we get $(a_i \wedge c_i, b_i \wedge d_i) = (0, 1)$ and $(0, 1) \not\models (1, 1)$.[14]

However, this should not come as a surprise. Indeed, here $\models$ preserves pairs of the form $(0, 1)$, while the conjunction of pairs preserves $(0, 1)$ if it appears in both places of the conjunction, but also when one of the pairs is equal to (1, 1) for some feature. This leads us to introduce a new operation $\wedge\vee$ mixing conjunction and disjunction:

$$(\vec{a}, \vec{b}) \wedge\vee (\vec{c}, \vec{d}) = (\vec{a} \wedge \vec{c}, \vec{b} \vee \vec{d})$$

Obviously, this operator $\wedge\vee$ is commutative and associative by construction. As much as the logical consequence relation between pairs defined by (8) makes sense, the intuition might seem more fragile for the conjunction / disjunction of pairs. However note that $(a_i \wedge c_i, b_i \vee d_i) = (1, 0)$ only if $(a_i, b_i) = (c_i, d_i) = (1, 0)$. By contrast, if (a_i, b_i) or $(c_i, d_i) = (0, 1)$, $(a_i \wedge c_i, b_i \vee d_i) = (0, 1)$.

In a dual manner, one can define

$$(\vec{a}, \vec{b}) \vee\wedge (\vec{c}, \vec{d}) = (\vec{a} \vee \vec{c}, \vec{b} \wedge \vec{d}).$$

Indeed there is a De Morgan duality with respect to the operation $\circlearrowleft$ between $\vee\wedge$ and $\wedge\vee$, namely

$$\circlearrowleft(\circlearrowleft(\vec{a}, \vec{b}) \vee\wedge \circlearrowleft(\vec{c}, \vec{d})) = (\vec{a}, \vec{b}) \wedge\vee(\vec{c}, \vec{d}).$$

[14] There are two other cases of violation when $(a_i, b_i) = (1, 0)$, $(c_i, d_i) = (0, 0)$ or $(c_i, d_i) = (0, 1)$, we get $(a_i \wedge c_i, b_i \wedge d_i) = (0, 0)$, and $(0, 0) \not\models (1, 0)$. Besides, $(\vec{a}, \vec{b}) \not\models (\vec{a}, \vec{b}) \vee (\vec{c}, \vec{d})$ due to three possible situations: i) $(a_i, b_i) = (0, 0)$, $(c_i, d_i) = (1, 0)$ and $(0, 0) \not\models (1, 0)$; ii) & iii) $(a_i, b_i) = (0, 1)$, $(c_i, d_i) = (1, 1)$ or $(c_i, d_i) = (1, 0)$, and $(0, 1) \not\models (1, 1)$.

Note that $(a_i \vee c_i, b_i \wedge d_i) = (0, 1)$ only if $(a_i, b_i) = (c_i, d_i) = (0, 1)$. But, if (a_i, b_i) or $(c_i, d_i) = (1, 0)$, $(a_i \wedge c_i, b_i \vee d_i) = (1, 0)$. Then it can be checked that $\vee\wedge$ behaves like a conjunction, and $\wedge\vee$ like a disjunction, in the sense that:

Proposition 4.

$$(\vec{a}, \vec{b}) \vee\wedge (\vec{c}, \vec{d}) \vDash (\vec{a}, \vec{b}) \vDash (\vec{a}, \vec{b}) \wedge\vee (\vec{c}, \vec{d})$$

where $\vDash$ *is defined by (8).*

Proof. We should first show that $(a \vee c, b \wedge d) \vDash (a, b)$. Indeed it holds since we have 1. $\neg(a \vee c) \wedge b \wedge d \vDash \neg a \wedge b$; 2. $a \wedge \neg b \vDash (a \vee c) \wedge \neg(b \wedge d)$.

It remains to show that $(a, b) \vDash (a \wedge c, b \vee d)$. Indeed it can be checked that we have 1. $\neg a \wedge b \vDash \neg(a \wedge c) \wedge (b \vee d)$; 2. $a \wedge c \wedge \neg(b \vee d) \vDash a \wedge \neg b$. □

Remark. Lines for further research

The conditional events involved in the conditional logical proportions have a tri-valued semantics. From an analogical proportion point of view, a natural way to associate a tri-valuation to an ordered pair of Boolean vectors, is to compute their difference to get a vector belonging to $\{-1, 0, 1\}^n$: $val_A(\vec{a}, \vec{b}) = \vec{a} - \vec{b} = (a_1 - b_1, ..., a_n - b_n) \in \{-1, 0, 1\}^n$.

Then one can check that $A(\vec{a}, \vec{b}, \vec{c}, \vec{d})$ is true if and only if $val_A(\vec{a}, \vec{b}) = val_A(\vec{c}, \vec{d})$. Moreover, if $A(\vec{a}, \vec{b}, \vec{c}, \vec{d})$ is true, we have

$$(\vec{a} \wedge \vec{c}) - (\vec{b} \wedge \vec{d}) = \vec{a} - \vec{b} = \vec{c} - \vec{d} = (\vec{a} \vee \vec{c}) - (\vec{b} \vee \vec{d}).$$

This means that $A(\vec{a}, \vec{b}, \vec{c}, \vec{d})$ entails $A(a \vec{\wedge} \vec{c}, b \overset{\rightarrow}{\wedge} \vec{d}, \vec{a} \vee \vec{c}, \vec{b} \vee \vec{d})$, but the converse is wrong.[15]

While the analogical proportion insists on the identity of the differences existing in each pair, the paralogy expresses rather a parallel between the pairs at the level of shared properties, positively or negatively. This is reflected in the following result, dual to that for analogy:

$$P(\vec{a}, \vec{b}, \vec{c}, \vec{d}) \text{ iff } \begin{cases} Dif(\vec{a}, \vec{b}) = Dif(\vec{c}, \vec{d}) \\ Equ^1(\vec{a}, \vec{b}) = Equ^1(\vec{c}, \vec{d}) \\ Equ^0(\vec{a}, \vec{b}) = Equ^0(\vec{c}, \vec{d}) \end{cases}$$

We could also define an entailment starting from paralogy, such that
$(\vec{a}, \vec{b}) \vDash_P (\vec{c}, \vec{d}) \Leftrightarrow \vec{a} \wedge \vec{b} \vDash \vec{c} \wedge \vec{d}$ and $\neg\vec{c} \wedge \neg\vec{d} \vDash \neg\vec{a} \wedge \neg\vec{b}$,
or alternatively $(\vec{a}, \vec{b}) \vDash'_P (\vec{c}, \vec{d}) \Leftrightarrow \neg\vec{a} \wedge \neg\vec{b} \vDash \neg\vec{c} \wedge \neg\vec{d}$ and $\vec{c} \wedge \vec{d} \vDash \vec{a} \wedge \vec{b}$,
depending if we privilege the persistence of properties shared positively inside the pairs, or shared negatively, when going from the pair $(\vec{a}, \vec{b})$ to the $(\vec{c}, \vec{d})$.

[15] Indeed $A(a \vec{\wedge} \vec{c}, b \overset{\rightarrow}{\wedge} \vec{d}, \vec{a} \vee \vec{c}, \vec{b} \vee \vec{d})$ is also true for $(a\ b\ c\ d) = (0\ 1\ 1\ 0)$ or $(1\ 0\ 0\ 1)$.

Moreover, the tri-valuation naturally associated with a pair, from the point of view of paralogy, would be $val_P(\vec{a}, \vec{b}) = (a_1 + b_1, ..., a_n + b_n) \in \{0, 1, 2\}^n$. Indeed it can be checked that $P(\vec{a}, \vec{b}, \vec{c}, \vec{d})$ holds true if and only if $val_P(\vec{a}, \vec{b}) = val_P(\vec{c}, \vec{d})$.

We leave these entailments associated with the paralogy, and the tri-valued logics associated with the analogical proportion and the paralogical proportion for a further study.

5.5 Creative inference

Given a set $\mathcal{S}$ of existing items, each represented by a set of Boolean attribute values, creativity may amount to produce an item not in $\mathcal{S}$, but described by the same set of attributes. Viewed like that, creativity is an easy game: we have just to choose at random the attribute values and to check if the result is not already in $\mathcal{S}$. However with such a process we have no control on the the attribute values that might be desirable. In the following, we present a *creative inference process* that attempts to improve a particular item or entity, taking advantage of a set of ordered pairs of existing items, using an analogical proportion-based mechanism. However, we certainly do not claim that every form of creative analogical inference, taken in the broadest sense, could be captured by the mechanism we propose.

More precisely, let us suppose we have a sample set $\mathcal{S}$ of items from which a set $\mathcal{P}$ of k ordered pairs $(\vec{a^j}, \vec{b^j})$ with $j \in \{1, \ldots, k\}$ has been extracted where the $\vec{a^j}$'s and $\vec{b^j}$'s are in $\mathcal{S}$. Each vector in $\mathcal{S}$ is a Boolean representation of an object/profile/situation belonging to a real world universe, and then, each pair of vectors $(\vec{a^j}, \vec{b^j})$, all of the same dimension n, represents legitimate, feasible / allowed changes from $\vec{a^j}$ to $\vec{b^j}$.

Then given a current fixed item represented by a vector $\vec{c} \in \mathcal{S}$ one may wonder what new item(s) $\vec{d}$ could be obtained by applying some change existing in the base of pairs. This item could be the representation of a plausible item in the real world.

A first option is to consider the set of solutions (when the solution exists)

$$\vec{d} \in \{\vec{x} \mid \exists(\vec{a^j}, \vec{b^j}) \in \mathcal{P}, j \in \{1, \ldots, k\} \text{ such that } A(\vec{a^j}, \vec{b^j}, \vec{c}, \vec{x}) \text{ holds}\} \qquad (9)$$

This is the approach followed in [1]. When there is no solution or when the values found for $\vec{d}$ are not considered satisfactory enough, we have to consider other options. One idea would be to consider the entailment (8) between pairs associated to the analogical proportion, and then to look for the $\vec{d}$'s such that:

$$\vec{d} \in \{\vec{x} \mid \exists(\vec{a^j}, \vec{b^j}) \in \mathcal{P}, j \in \{1, \ldots, k\} \text{ such that } (\vec{a^j}, \vec{b^j}) \models (\vec{c}, \vec{x})\}$$

But this option has two drawbacks. First, the $\vec{d}$'s obtained depend on a unique pair $(\vec{a^j}, \vec{b^j})$. Second, $\models$ is quite permissive and the number of pairs $(\vec{c}, \vec{x})$ obtained is likely to be rather large as seen in subsection 5.3 and there is a risk of losing control.

What seems to be a better idea is to enlarge the initial base of pairs by computing new pairs belonging to the closure of operation $\wedge\!\!\vee$ introduced in the previous subsection 5.4. This operation has the merit of "cumulating" the acquisition of features[16]. Extending the initial set $\mathcal{P}$ of pairs gives us more chance to find a plausible $\vec{d}$, perhaps with more desirable features. More precisely we apply (9) where $\mathcal{P}$ is replaced by $\mathcal{P}' = \{(a^k, b^k)|\ (a^k, b^k) = (a^i, b^i) \wedge\!\!\vee (a^j, b^j)$ such that $((a^i, b^i), (a^j, b^j)) \in \mathcal{P}^2\}$. We may apply this enlargement of $\mathcal{P}$ recursively to $\mathcal{P} \cup \mathcal{P}'$ and so on. This process ensures that i) the $\vec{d}$ obtained are new, and ii) they are obtained from an existing $\vec{c}$ on the basis of already existing changes, since observed on pairs of existing items. Is $\vec{d}$ thus obtained, valuable ? This a completely different issue.

Note that this way of reasoning parallels non monotonic reasoning with conditional objects, where from a base of default rules "if a^j then b^j" represented by a set of conditional objects $b^j|a^j$, one deduces a new conditional object $d|c$, using entailment (7) and conjunction $\&$, where c corresponds to everything we know in the current context, for which we then conclude d [18].

6 Example and first experiments

Before giving some statistics about the behaviour of our mechanism, we start with a simple example.

6.1 An example freely inspired from a simplified Kaggle dataset

To avoid the creation of a completely artificial dataset, we start by using a Kaggle dataset. Kaggle is a platform renowned for hosting data science competitions, collaborative projects, and educational resources, accessible at https://www.kaggle.com/.

The targeted dataset [16] encompasses the details of 1000 users, characterized by 32 attributes. The final column denotes whether they purchased a bike, forming the basis for a binary classification task. Following the exclusion of rows with missing data, 952 complete rows remain.

In order to facilitate the understanding of our process, we narrow our focus to the first 6 attributes of this dataset, creating a simplified universe where objects are represented as Boolean vectors of dimension 6. In the initial dataset, these 6 first features are Marital Status, Gender, Income, Children, Education, Occupation, but, more generally, each attribute has to be viewed as an individual feature describing an object.

By limiting our analysis to the first 6 attributes, we inevitably encounter duplicates. Depending on the random shuffle of the initial dataset comprising 952 items, we end up

[16] However note that $(0, 0) \wedge\!\!\vee (1, 1) = (1, 1) \wedge\!\!\vee (0, 0) = (0, 1)$, which may create some unfeasible change; in such a case the generated pair(s) should not be considered in the further process.

with less than 2^6 = 64 distinct elements from which we must select pairs.

Consequently, we consider only two pairs for the need of our example, denoted $(\vec{a_1}, \vec{b_1})$, $(\vec{a_2}, \vec{b_2})$ that respect the following constraints:

- The Hamming distance $hamming(\vec{a_i}, \vec{b_i})$ is equal to 2 because a pair should represent a realistic perturbation of $\vec{a}$ into $\vec{b}$.
- We do not loose any option when moving from $\vec{a_i}$ to $\vec{b_i}$, i.e., we forbid to have an attribute j such that $a_j = 1$ and $b_j = 0$. All other combinations are allowed.

At this stage of this experiment, the two pairs are selected in a random manner w.r.t. the previous constraints, from the candidate pairs, since we currently lack specific information about the entire universe. However, in practical scenarios, prior knowledge about the universe could result in more suitable selections. Here is an example of the two pairs constituting $\mathcal{P}$:
- $(\vec{a_1}, \vec{b_1}) = ([0,0,0,0,1,0],[0,1,0,0,1,1])$
- $(\vec{a_2}, \vec{b_2}) = ([0,0,0,0,1,0]),[1,0,0,0,1,1])$
When we extend $\mathcal{P}$ with operator ($\wedge\vee$ but without doing the full closure), we add to P the following pair:
- $(\vec{a_3}, \vec{b_3}) = ([0,0,0,0,1,0],[1,1,0,0,1,1])$
just because $(\vec{a_3}, \vec{b_3}) = (\vec{a_1}, \vec{b_1}) \wedge\vee (\vec{a_2}, \vec{b_2})$
Because, $\vec{a_1} = \vec{a_2}$, obviously $\vec{a_1} = \vec{a_3}$ but this is not necessary. Starting from $\vec{c} = [0,0,0,1,0,0]$, we observe in Table 7 that the 3 corresponding analogical equations are solvable. The solution of the third equation is then a new object, which is distinct from the 5 existing vectors $\vec{a_1}, \vec{b_1}, \vec{b_2}, \vec{x_1}, \vec{x_2}$.

This approach only makes practical sense when considering Boolean representations of relatively large dimensions. That is why we give in the following subsection some figures about what can be expected in higher dimensions.

6.2 Higher dimensions

Indeed, in the context of Boolean vectors with high dimension (let us say larger than 10), the available data $\mathcal{S}$ are generally scarce compared to the whole universe: this is a well-known consequence of the curse of dimensionality. For instance, for vectors of dimension 30, the space of possible profiles is of size $2^{30} \sim 10^9$.

Considering $\mathcal{P}$ as the set of pairs built from two distinct elements from $\mathcal{S}$, we can first inquire about the existence, within $\mathcal{P}$, of pairs $(\vec{c}, \vec{d})$ that are logical consequences of another pair $(\vec{a}, \vec{b})$ (in the sense of (8)) in $\mathcal{P}$.

To answer this question, we conducted experiments in dimension 10 and 30 (with reasonable execution times) while varying the size of the sample $\mathcal{S}$. From a practical perspective, when given a pair $(\vec{a}, \vec{b})$ in $\mathcal{P}$, we determine the quantity of pairs $(\vec{c}, \vec{d})$ in $\mathcal{P}$ that are

	Opt1	Opt2	Opt3	Opt4	Opt5	Opt6
$\vec{a_1}$	0	0	0	0	1	0
$\vec{b_1}$	0	1	0	0	1	1
$\vec{c}$	0	0	0	1	0	0
x_1	0	1	0	1	0	1
	Opt1	Opt2	Opt3	Opt4	Opt5	Opt6
$\vec{a_2}$	0	0	0	0	1	0
$\vec{b_2}$	1	0	0	0	1	1
$\vec{c}$	0	0	0	1	0	0
x_2	1	0	0	1	0	1
	Opt1	Opt2	Opt3	Opt4	Opt5	Opt6
$\vec{a_3}$	0	0	0	0	1	0
$\vec{b_3}$	1	1	0	0	1	1
$\vec{c}$	0	0	0	1	0	0
x_3	1	1	0	1	0	1

Table 7: Example in dimension 6

logical consequences of $(\vec{a}, \vec{b})$. Additionally, we distinguish in this count the pairs that form an analogical proportion $(\vec{a}, \vec{b}) :: (\vec{c} : \vec{d})$. We calculate the average of these two numbers across the total number of pairs $(\vec{a}, \vec{b})$ within $\mathcal{P}$. Finally, we present the average values from these calculations in Table 8 based on 10 tests (or 10 samples $\mathcal{S}$). Obviously, as soon as the random sample set S has a small size w.r.t. the whole universe size, it is very unlikely to get inside $\mathcal{S}$, four vectors $\vec{a}, \vec{b}, \vec{c}, \vec{d}$ such that $\vec{a} : \vec{b} :: \vec{c} : \vec{d}$. Additionally, we notice that as the dimension increases, we also fail to discover any logical consequence within $\mathcal{P}$ that can be seen as a relaxation of analogical proportion. This lack of logical consequences can be attributed to the fact that a global solution is only acceptable if there is a solution for each individual component. When the number of components increases, the number of constraints also increases, thereby decreasing the probability of obtaining a global solution.

Dim	Size $\mathcal{S}$	# pairs	# tests	# analogies	std. dev	# log. cons.	std. dev
10	50	1225	10	0	0	20	4
10	100	4950	10	0	0	113	18
30	100	4950	10	0	0	0	0
30	200	19900	10	0	0	0	0

Table 8: Number of pairs that are logical consequences inside S

And the size of the sample $\mathcal{S}$, from which $\mathcal{P}$ is derived, cannot compensate these increasing constraints.

If we have a sample $\mathcal{S}$ of size 1000, it is natural to be interested in a “reasonable” extension of the sample. This is where the analogy with analogical extension comes into play, which involves completing the set of initial examples, as seen, for instance, in [11]. But if the analogical extension does not provide enough new elements, we could then implement, initially, the logical consequence of pairs, seen as a way to weaken the analogical constraint as follows:

- Every pair $(\vec{a}, \vec{b})$ from the sample represents a potential variation of the profiles.
- Any pair $(\vec{c}, \vec{d})$ such that $(\vec{a}, \vec{b}) \models (\vec{c}, \vec{d})$ can potentially be regarded as the description of a candidate variation of the profiles.

In the absence of an efficient algorithm, the task of generating logical consequences can prove to be very complex. So, instead of trying to generate via brute force the set of logical consequences, another option is to try to solve the equation $(\vec{a}, \vec{b}) \models (\vec{c}, \vec{d})$ where $\vec{a}, \vec{b}, \vec{c}$ are in $\mathcal{S}$: in that context, instead of looking for a pair $(\vec{c}, \vec{d})$ we just look for at least one element $\vec{d}$, if it exists, that is not in $\mathcal{S}$ and satisfying $(\vec{a}, \vec{b}) \models (\vec{c}, \vec{d})$. Once again:

- Every pair $(\vec{a}, \vec{b})$ from the sample represents a potential variation of the profiles.
- Given another profile $\vec{c}$ from $\mathcal{S}$, a profile $\vec{d} \notin \mathcal{S}$ such that $(\vec{a}, \vec{b}) \models (\vec{c}, \vec{d})$ can be considered plausible and added to the initial sample.

First of all, given a pair $(\vec{a}, \vec{b}) \in \mathcal{P}$, we compute the average number over $\vec{c} \in \mathcal{S}$ of $\vec{d} \notin \mathcal{S}$ satisfying $(\vec{a}, \vec{b}) \models (\vec{c}, \vec{d})$. This number tells us the likelihood of creating a new profile $\vec{d}$ when solving the equation $(\vec{a}, \vec{b}) \models (\vec{c}, \vec{d})$ starting from 3 elements $\vec{a}, \vec{b}, \vec{c} \in S$. Then for a given sample S, we can compute the average number of profile $\vec{d}$ that can be generated from pairs in $\mathcal{P} = \mathcal{S} \times \mathcal{S}$ with the help of a third element $\vec{c} \in \mathcal{S}$. Finally, we average this computation on 10 tests and show the result in Table 9 showing the number of $\vec{d} \notin \mathcal{S}$ with the average standard deviation as last column. We conducted experiments in dimensions 10, 30, and 50 with various sample sizes. Table 9 shows that, in general, the equation $(\vec{a}, \vec{b}) \models (\vec{c}, \vec{d})$ where $\vec{d}$ is the unknown, and $\vec{a}, \vec{b}, \vec{c} \in S$ does not have a solution. As it is the case for Table 8, this might be understood because a global solution $\vec{d}$ is acceptable only if there is a solution componentwise. Augmenting the number of components (from 10 to 30 to 50) increases the number of constraints and reduce the likelihood to having a global solution.

The previous experiments suggest that the use of $\wedge\!\!\vee$ operator as a pair creator might be more productive. We will count in Table 10 how many completely new pairs are created when applying the $\wedge\!\!\vee$ operator to all pairs derived from the sample $\mathcal{S}$. We display the

Dim.	Size $\mathcal{S}$	# pairs	# tests	# $\vec{d} \notin \mathcal{S}$	std. dev.
10	50	1225	10	0.3	0.2
10	100	4950	10	0.23	0.21
30	100	4950	10	0.03	0.03
30	200	19900	10	0.03	0.04
50	100	4950	10	0.01	0.02
50	200	19900	10	0.007	0.006

Table 9: Number of vectors $\vec{d}$ solutions of the equation

average value on 10 tests with the standard deviation. At this stage, we do not eliminate pairs where at least one component appears as $(0, 0) \wedge\vee (1, 1)$ or $(1, 1) \wedge\vee (0, 0)$. See footnote number 8.

Dim.	Size S	# pairs	# tests	# new pairs	std dev.
10	50	1225	10	333	34
10	100	4950	10	552	28
30	100	4950	10	9423	65
30	200	19900	10	Not Avail.	Not Avail.

Table 10: Number of deduced pairs built with vectors not in S

Since the resulting pairs are only retained if both constituting vectors are not in S, we have constructed at least $\#newpairs$ new vectors (a new vector may appear in multiple new pairs).

It is widely admitted that analogical reasoning only leads to plausible consequences. Its application to creativity does not escape this rule. It will certainly be useful in practice to verify, in one way or another, the feasibility of the new pairs obtained.

7 Conclusion

This paper has discussed analogical reasoning based on analogical proportions. We have first singled out this inference as a special form of induction, more precisely of transduction, where comparisons between examples take place. After providing a refresher on analogical proportions defined between entities represented by means of Boolean or nominal features, we have emphasized that, in this case, analogical proportions define *equivalence* relations between ordered pairs of entities.

We have then surveyed how analogical proportions-based inference can be used for classification tasks, before contrasting this use with the generation of a novel entity from three known entities under some conditions. Taking advantage of the belonging of analogical proportions to the setting of logical proportions, we have found out that there exist only two small subsets of logical proportions that define equivalence relations between ordered pairs: the analogical proportion together with a related proportion called paralogy on the one hand, and four conditional logical proportions between conditional events on the other hand.

Taking lessons from the logic of conditional events and its key role in nonmonotonic inference, we have defined an entailment relation between ordered pairs (in agreement with analogical proportions) and appropriate conjunction and disjunction of ordered pairs. Then, we have described a creative inference process using these entailment and operations. First experiments with them have been also reported. However, it should be clear that, in our formal setting, we only capture a particular form of "creative" inference, which is not intended to cover every type of creative analogical inference.

It is clear that the new logic of ordered pairs outlined in the second half of this paper is still in its infancy and many aspects remain to be developed, as well as its possible use in a creative machinery for controlling the derivation of new items from a given entity on the basis of a set of ordered pairs reporting feasible changes between entities. Moreover, we have focused on analogical proportions between entities described by means of Boolean or nominal values, the case of numerical values already investigated in classification, remains to be considered in creativity. Finally, the pairs involved in analogical proportions can be seen as describing changes resulting from actions, which suggests studying relationships with the logic of action.

References

[1] S. Afantenos, H. Prade, and L. Cortez Bernardes. Analogical proportions and creativity: A preliminary study, 2023. arXiv:2310.13500 [cs.CL].

[2] S. D. Afantenos, T. Kunze, S. Lim, H. Prade, and G. Richard. Analogies between sentences: Theoretical aspects - preliminary experiments. In J. Vejnarová and N. Wilson, editors, *Proc. 16th Europ. Conf. on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU'21), Prague, Sept. 21-24*, volume 12897 of *LNCS*, pages 3–18. Springer, 2021.

[3] N. Barbot, L. Miclet, and H. Prade. Analogy between concepts. *Artif. Intell.*, 275:487–539, 2019.

[4] N. Barbot, L. Miclet, H. Prade, and G. Richard. Analogical proportions and binary trees. In J.-Y. Béziau, J.-P. Desclés, A. Moktefi, and A. C. Pascu, editors, *Logic in Question. Talks from the Annual Sorbonne Logic Workshop (2011- 2019)*, pages 435–458. Birkhäuser, 2022.

[5] P. F. A. Bartha. *By Parallel Reasoning: The Construction and Evaluation of Analogical Arguments*. Oxford University Press, 2009.

[6] S. Benferhat, D. Dubois, and H. Prade. Nonmonotonic reasoning, conditional objects and possibility theory. *Artificial Intelligence*, 92(1-2):259–276, 1997.

[7] M. Bounhas, K. Mellouli, H. Prade, and M. Serrurier. Possibilistic classifiers for numerical data. *Soft Comput.*, 17(5):733–751, 2013.

[8] M. Bounhas, H. Prade, and G. Richard. Analogy-based classifiers for nominal or numerical data. *Int. J. Approx. Reason.*, 91:36–55, 2017.

[9] M. Bounhas, H. Prade, and G. Richard. Analogy-based classifiers for nominal or numerical data. *Int. J. of Approximate Reasoning*, 91:36 – 55, 2017.

[10] W. Correa Beltran, H. Prade, and G. Richard. Constructive solving of Raven's IQ tests with analogical proportions. *Int. J. Intell. Syst.*, 31(11):1072–1103, 2016.

[11] M. Couceiro, N. Hug, H. Prade, and G. Richard. Analogy-preserving functions: A way to extend Boolean samples. In *Proc. 26th IJCAI Conf*, 1575-1581, 2017.

[12] M. Couceiro, N. Hug, H. Prade, and G. Richard. Behavior of analogical inference w.r.t. boolean functions. In *Proc. 27th IJCAI Stockholm*, pages 2057–2063, 2018.

[13] M. Couceiro, E. Lehtonen, L. Miclet, H. Prade, and G. Richard. When nominal analogical proportions do not fail. In J. Davis and K. Tabia, editors, *Proc. 14th Int. Conf. on Scalable Uncertainty Management (SUM'20), Bozen-Bolzano, Sept. 23-25*, volume 12322 of *LNCS*, pages 68–83. Springer, 2020.

[14] T. Davies and S. Russell. A logical approach to reasoning by analogy. In *Proc. 10th Int. Joint Conf. on Artificial Intelligence (IJCAI-87), Milan, Aug. 23-28*, pages 264–270, 1987.

[15] B. De Finetti. La logique des probabilités. In *Congrès Int. de Philosophie Scientifique, IV. Induction et probabilité*, pages 31–39, Paris,, 1936. Hermann.

[16] H. Dedhia. Bike buyers 1000. `https://www.kaggle.com/datasets/heeraldedhia/bike-buyers`, 2021. Accessed: September 1, 2023.

[17] R. Dreistadt. The use of analogies and incubation in obtaining insights in creative problem solving. *The Journal of Psychology*, 71:159–175, 1969.

[18] D. Dubois and H. Prade. Conditional objects as nonmonotonic consequence relationships. *IEEE Trans. on Syst., Man and Cyber.*, 24:1724–1740, 1994.

[19] D. Dubois and H. Prade. Possibility theory: Qualitative and quantitative aspects. In D. M. Gabbay and Ph. Smets, editors, *Quantified Representation of Uncertainty and Imprecision*, volume 1 of *Handbook of Defeasible Reasoning and Uncertainty Management Systems*, pages 169–226. Kluwer Acad. Publ., 1998.

[20] D. Dubois and H. Prade. An overview of ordinal and numerical approaches to causal diagnostic problem solving. In D. M. Gabbay and R. Kruse, editors, *Handbook of Defeasible Reasoning and Uncertainty Management Systems. Vol. 4. Abductive Reasoning and Learning*, volume 1 of *Handbook of Defeasible Reasoning and Uncertainty Management Systems*, pages 231–280. Kluwer Acad. Publ., 2000.

[21] D. Dubois and H. Prade. Towards a logic-based view of some approaches to classification tasks. In M.-J. Lesot, S. M. Vieira, M. Z. Reformat, J. P. Carvalho, A. Wilbik, B. Bouchon-Meunier,

and R. R. Yager, editors, *Proc. 18th Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU'20), Lisbon, June 15-19, Part III*, volume 1239 of *CCIS*, pages 697–711. Springer, 2020.

[22] W. F. Edwards. *Likelihood*. Cambridge University Press, 1972.

[23] T. G. Evans. A program for the solution of a class of geometric-analogy intelligence-test questions. In M. L. Minsky, editor, *Semantic Information Processing*, pages 271–353. MIT Press, Cambridge, Ma, 1968.

[24] E. Fix and J. L. Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties. Report USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

[25] B. Fuchs, J. Lieber, L. Miclet, A. Mille, A. Napoli, H. Prade, and G. Richard. Case-based reasoning, analogy and interpolation. In P. Marquis, O. Papini, and H. Prade, editors, *A Guided Tour of Artificial Intelligence Research. 1 - Knowledge Repreentation, Reasoning and Learning*, pages 307–339. Springer, 2020.

[26] A. Gammerman, V. Vovk, and V. Vapnik. Learning by transduction. In *Proc. 14th Conf. on Uncertainty in AI*, pages 148–155. Morgan Kaufmann, 1998.

[27] D. Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170, 1983.

[28] D. Gentner, K. J. Holyoak, and B. N. Kokinov. *The Analogical Mind: Perspectives from Cognitive Science*. Cognitive Science, and Philosophy. MIT Press, Cambridge, MA, 2001.

[29] D. F. Halpern. Analogies as a critical thinking skill. In Dale E. Berger, Kathy Pezdek, and William P. Banks, editors, *Applications of Cognitive Psychology: Problem Solving, Education, and Computing*, pages 75–86. Routledge, 1986.

[30] M. B. Hesse. On defining analogy. *Proceedings of the Aristotelian Society*, 60:79–100, 1959.

[31] S. Klein. Analogy and mysticism and the structure of culture (and Comments & Reply). *Current Anthropology*, 24 (2):151–180, 1983.

[32] S. Klein. The analogical foundations of creativity in language, culture & the arts: the upper paleolithic to 2100 CE. In P. McKevitt, S. O'Nullain, and C. Mulvihill, editors, *Proc. 8th Int. Workshop on the Cognitive Science of Natural Language Processing, Galway, Ireland, Aug. 9-11, 1999*, pages 347–371. Amsterdam: John Benjamin, 2002.

[33] S. Kraus, D. Lehmann, and M. Magidor. Nonmonotonic reasoning, preferential models and cumulative logics. *Artificial Intelligence*, 44:167–207, 1990.

[34] Y. Lepage and C.-L. Goh. Towards automatic acquisition of linguistic features. In K. Jokinen and E. Bick, editors, *Proc. 17th Nordic Conf. of Computational Linguistics, NODALIDA'09, Odense, Denmark, May 14-16*, pages 118–125. Northern European Association for Language Technology (NEALT), 2009.

[35] J. Lieber, E. Nauer, and H. Prade. Improving analogical extrapolation using case pair competence. In K. Bach and C. Marling, editors, *Proc. 27th Int. Conf. on Case-Based Reasoning (ICCBR '19), Otzenhausen, Sept. 8-12*, volume 11680 of *LNCS*, pages 251–265. Springer, 2019.

[36] J. Lieber, E. Nauer, H. Prade, and G. Richard. Making the best of cases by approximation, interpolation and extrapolation. In M. T. Cox, P. Funk, and S. Begum, editors, *Proc. 26th Int.*

Conf. on Case-Based Reasoning (ICCBR '18), Stockholm, July 9-12, volume 11156 of *LNCS*, pages 580–596. Springer, 2018.

[37] S. Lim, H. Prade, and G. Richard. Classifying and completing word analogies by machine learning. *Int. J. Appr. Reas.*, 132:1–25, 2021.

[38] L. Miclet, S. Bayoudh, and A. Delhay. Analogical dissimilarity: Definition, algorithms and two experiments in machine learning. *J. Artif. Intell. Res.*, 32:793–824, 2008.

[39] L. Miclet and H. Prade. Handling analogical proportions in classical logic and fuzzy logics settings. In *Proc. 10th Eur. Conf. on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU'09),Verona*, pages 638–650. Springer, LNCS 5590, 2009.

[40] J. S. Mill. *A System of Logic, Ratiocinative and Inductive, being a connected view of the principles of evidence and the methods of scientific investigation.* Cambridge University Press, 1843. Book III "Of Induction", chap VIII "Of the four method of experimental inquiry", chap. XX: "Of Analogy".

[41] G. Minnameier. Abduction, induction, and analogy. On the compound character of analogical inferences. In L. Magnani, W. Carnielli, and C. Pizzi, editors, *Model-Based Reasoning in Science and Technology: Abduction, Logic, and Computational Discovery*, pages 107–119. Springer Berlin Heidelberg, 2010.

[42] R. Misiewicz. Peirce on analogy. *Trans. of the Charles S. Peirce Society: A Quarterly Journal in American Philosophy*, 56(3):299–325, 2020.

[43] T. M. Mitchell. *Version spaces: An approach to concept learning.* PhD thesis, Stanford University, 1979.

[44] C. S. Peirce. *Philosophical Writings of Peirce.* Dover Publ., New York, selected and edited by J. Buchler, 1955.

[45] C. S. Peirce. *Collected Papers of Charles Sanders Peirce.* Harvard University Press, Cambridge, 1958. Edited by C. Hartshorne & P. Weiss (vol. 1-6), and A. Burks (vol. 7-8).

[46] V. Pirrelli and F. Yvon. Analogy in the lexicon: a probe into analogy-based machine learning of language. In *Proc. 6th Int. Symp. on Human Communication*, 1999. Santiago de Cuba.

[47] G. Polya. *How to Solve It.* Princeton University Press, 2nd ed. 1957, 1945.

[48] H. Prade and G. Richard. From analogical proportion to logical proportions. *Logica Universalis*, 7(4):441–505, 2013.

[49] H. Prade and G. Richard. Homogenous and heterogeneous logical proportions. *IfCoLog J. of Logics and their Applications*, 1(1):1–51, 2014.

[50] H. Prade and G. Richard. Analogical proportions and analogical reasoning - An introduction. In D. W. Aha and J. Lieber, editors, *Proc. 25th Int. Conf. on Case-Based Reasoning (ICCBR'17), Trondheim, June 26-28*, volume 10339 of *LNCS*, pages 16–32. Springer, 2017.

[51] H. Prade and G. Richard. Analogical proportions: From equality to inequality. *Int. J. of Approximate Reasoning*, 101:234 – 254, 2018.

[52] H. Prade and G. Richard. Analogical proportions: Why they are useful in AI. In *Proc. 30th Int. Joint Conf. on Artificial Intelligence (IJCAI-21), (Z.-H. Zhou, ed.) Virtual Event / Montreal, Aug. 19-27*, pages 4568–4576, 2021.

[53] H. Prade and G. Richard. First steps towards a logic of ordered pairs. In Z. Bouraoui and

S. Vesic, editors, *Proc. 17th Europ. Conf. on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU'23), Arras, Sept. 19–22*, volume 14294 of *LNCS*, page to appear. Springer, 2023.

[54] S. J. Russell. *The Use of Knowledge in Analogy and Induction*. Pitman, UK, 1989.

[55] G. Shakhnarovich, T. Darrell, and P. Indyk, editors. *Nearest-Neighbor Methods in Learning and Vision*. MIT Press, 2005.

[56] N. Stroppa and F. Yvon. Analogical learning and formal proportions: Definitions and methodological issues. Technical report, ENST, June 2005.

[57] P. H. Winston. Learning and reasoning by analogy. *Commun. ACM*, 23(12):689–703, 1980.

[58] F. Yvon and N. Stroppa. Formal models of analogical proportions. Technical report, Ecole Nationale Supérieure des Télécommunications, no D008, 2006.

Received June 2023

Inductive Reasoning, Conditionals, and Belief Dynamics

Gabriele Kern-Isberner
Dept. of Computer Science, TU Dortmund, 44221 Dortmund, Germany
gabriele.kern-isberner@cs.tu-dortmund.de

Wolfgang Spohn
Dept. of Philosophy, University of Konstanz, 78457 Konstanz, Germany
wolfgang.spohn@uni-konstanz.de

Abstract

This paper presents a broad view on inductive reasoning by embedding it in theories of epistemic states, conditionals, and belief revision. More precisely, we consider inductive reasoning as a specific case of belief revision on epistemic states which include conditionals as a basic means for representing beliefs. We present a general framework for inductive reasoning from conditional belief bases that also allows for taking background beliefs into account, and illustrate this by probabilistic reasoning based on optimum entropy as well as by ranking-theoretic reasoning based on so-called c-revision. We explain the philosophical perspective behind our approach, and we illustrate its constructive usefulness as well as its integrating power.

1 Introduction

In a familiar sense, inductive reasoning means deriving general knowledge from given examples in a way that completes the example-based information concisely to make it applicable to other situations. In this paper, we take a bit broader view on inductive reasoning: we pursue the idea that inductive reasoning should be able to generate any

The idea for this joint paper originated from the First International Conference on Foundations, Applications and Theory of Inductive Logic at the University of Munich from October 12-14, 2022, supported by the Deutsche Forschungsgemeinschaft (DFG) under grant XXX. We thought that our presentations ("Some Strategic Considerations Concerning Inductive Logic", invited talk by Wolfgang Spohn, and "Inductive reasoning, conditionals, and belief revision" by Gabriele Kern-Isberner) there merged nicely.

kind of new beliefs from given beliefs and, ideally, complete the beliefs of a human being as far as possible. This is a very common and basic problem in the area of knowledge representation in artificial intelligence. Here, it is usually assumed that knowledge and beliefs of a human being, or an agent, respectively, can be represented by a knowledge base, i.e., a finite set of formulas in a suitable logic, and that more knowledge and beliefs can be inferred from this base. In artificial intelligence, the view on the distinction between knowledge and beliefs is a pragmatic one, because its main goal is to model knowledge and behaviour of agents. So, knowledge often means only subjective knowledge which is more or less the same as belief. Here, we avoid discussing the precise nature of knowledge and belief and use the terms "knowledge" and "belief" interchangeably, just as the term "epistemic state", the Greek origin of which refers to knowledge, stands for any kind of belief state.[1]

So, inductive reasoning should be able to extend the beliefs of a belief base in a non-trivial, principled way. Of course, the logic framework in which beliefs are represented plays a crucial role here. In the simple case of propositional logic, deduction, or more generally, a Tarski consequence operator would satisfy the general requirements of an inductive reasoning operator, and similarly for first-order predicate logic. Beyond classical logics, non-monotonic logics using so-called default rules, or rules with exceptions, provide more powerful inference operators, prominent approaches here are Reiter's default logic [49] and answer set programming [19]. Both are symbolic and able to infer formulas from belief bases of facts and rules. In quantitative logical settings, probability theory offers a rich semantic framework for nonmonotonic reasoning, and the *principle of maximum entropy* [26, 40] yields a most powerful inductive inference operator from probabilistic belief bases. There are also popular approaches using qualitative structures like (total) preorders, or semi-quantative methodologies based on Spohn's ordinal conditional functions, also called ranking functions [53, 55], like system Z [21] that allow for reasoning from conditional belief bases.

This paper aims at describing inductive reasoning in a broader context and in a more unified way, elaborating on connections to conditionals and belief change theory distinguishing clearly between background, or generic, beliefs and evidential, or contextual, information, a feature that is listed in [12] as one of three basic requirements a *plausible exception-tolerant inference system* has to meet. We build upon previous works, in particular [29, 33], and elaborate a general vision of inductive reasoning in the context of belief revision. While it has been well known that nonmonotonic reasoning and belief revision are "two sides of the same coin" [18],

[1] Philosophers are used to sharply distinguish knowledge and belief and have intensely discussed this distinction for more than 60 years. They mostly deny that knowledge is merely true belief or even merely justified true belief, see [25].

the focus here is on inductive reasoning as a concept that merges techniques from both areas to bring forth a methodology in which reasoning and revision can interact in various ways and which represents inductive reasoning from different background beliefs and under different contextual information. A core move in this methodology is to equip epistemic states with meta-structures supporting reasoning and revision, and to use conditionals for expressing beliefs in the first place.

Our approach allows for taking plausible propositional beliefs into account as well, namely by identifying a conditional $(A|\top)$, where $\top$ is a tautology, with the plausible belief A. Note that the statement "A is plausible" is not the same as saying "A is a (certain) fact", both from an epistemological and a knowledge representation point of view. While the latter statement is considered as factual evidence and takes only models of A into account, the first one also considers models of $\neg A$, but as less plausible, reflecting a more generic perspective. We show that our framework can indeed make a difference here.

Interestingly, total preorders on possible worlds are meta-structures that provide a solid foundation for reasoning, revision, and conditionals, and indeed, they are a basic requirement for AGM revision [28]. So, we build upon AGM revision but go far beyond that by addressing iterative revision and conditional revision. Ranking functions implement total preorders by assigning natural numbers to the different layers of a total preorder and thus allow for calculating differences as a measure of plausibility which make it possible to reason in a way that is similar to probabilistic reasoning. As a proof of concept, we illustrate our formal framework in a probabilistic environment by the entropy principles, and in a qualitative/semi-quantitative environment by ranking functions and c-revision.

The outline of the paper is as follows: After recalling basic definitions and notations in Section 2, we discuss, in Section 3, the nature of epistemic states and their dynamics or revision and their fundamental connections to argumentation, inductive reasoning, and conditionals. We explain both, the philosophical perspective as well as how this perspective allows a detailed view of the interactions between inductive reasoning and belief revision. Section 4 then specifies our approach in probabilistic terms via the principles of optimum entropy and in ranking-theoretic terms via the method of so-called c-revision. In section 5, we want to exemplify the integrative power of our approach by comparing inductive reasoning as developed here with the so-called method of focusing that says how to apply beliefs to specific situations. Section 6 provides a brief conclusion.

2 Basics and notations

The propositional language $\mathcal{L}$ with formulas A, B is defined in the usual way by virtue of a finite signature Σ with atoms $a, b, \ldots$ and junctors $\wedge, \vee$, and $\neg$ for conjunction, disjunction, and negation, respectively. The $\wedge$-junctor is mostly omitted, so that AB stands for $A \wedge B$, and negation is usually indicated by overlining the corresponding proposition, i.e. $\overline{A}$ means $\neg A$. Literals are positive or negated atoms. The set of all propositional interpretations over Σ is denoted by Ω_Σ. As the ignature will be fixed throughout the paper, we will usually omit the subscript and simply write Ω. Possible worlds are understood as a synonym for interpretations, and are usually represented by a complete conjunction of the corresponding literals, i.e., a conjunction mentioning all atoms of the signature such that exactly those atoms are negated that are evaluated to *false*. Also the satisfaction relation $\models$ between worlds and formulas is defined in the usual way: $\omega \models A$ iff ω evaluates A to be *true*. In this case, we say ω is a model of A. The set of all models of A is denoted by $Mod(A)$. Then, $A \models B$ for two formulas $A, B \in \mathcal{L}$ if $Mod(A) \subseteq Mod(B)$.

$\mathcal{L}$ is extended to a conditional language $(\mathcal{L} \mid \mathcal{L})$ by introducing a conditional operator $|$: $(\mathcal{L} \mid \mathcal{L}) = \{(B|A) \mid A, B \in \mathcal{L}\}$. $(\mathcal{L} \mid \mathcal{L})$ is a flat conditional language, no Boolean combinations or nestings of conditionals are allowed. Conditionals $(B|A)$ with *antecedent* (or *premise*) A and *consequent* B are basically considered as three-valued entities in the sense of de Finetti [9] which can be verified ($\omega \models AB$), falsified ($\omega \models A\overline{B}$), or simply not applicable ($\omega \models \overline{A}$) in a possible world ω. So, they have to be interpreted within richer semantic structures such as *epistemic states* like probability distributions, or ranking functions [53]. In this paper, we choose both of these semantic frameworks to exemplify our approach.

Probability distributions in a logical environment can be identified with probability functions $P : \Omega \to [0,1]$ with $\sum_{\omega \in \Omega} P(\omega) = 1$. The probability of a formula $A \in \mathcal{L}$ is given by $P(A) = \sum_{\omega \models A} P(\omega)$. Since $\mathcal{L}$ is finite, Ω is finite, too, and we only need additivity instead of σ-additivity. Conditionals are interpreted via conditional probabilities, so that $P(B|A) = \frac{P(AB)}{P(A)}$ for $P(A) > 0$, and $P \models (B|A)\,[x]$ iff $P(A) > 0$ and $P(B|A) = x$ ($x \in [0,1]$).

Ranking functions, also known as *ordinal conditional functions* (OCFs), $\kappa : \Omega \to \mathbb{N} \cup \{\infty\}$ with $\kappa^{-1}(0) \neq \emptyset$, were first introduced by Spohn [53]. They express degrees of plausibility of propositional formulas A by specifying degrees of disbeliefs of their negations $\overline{A}$. More formally, we have $\kappa(A) := \min\{\kappa(\omega) \mid \omega \models A\}$, so that $\kappa(A \vee B) = \min\{\kappa(A), \kappa(B)\}$. Hence, due to $\kappa^{-1}(0) \neq \emptyset$, at least one of $\kappa(A), \kappa(\overline{A})$ must be 0. Note that expressing absolutely certain beliefs is also possible by assigning the rank ∞ to all worlds falsifying those beliefs. A proposition A is believed if $\kappa(\overline{A}) > 0$ (which implies $\kappa(A) = 0$). Degrees of plausibility can also be assigned to conditionals by setting $\kappa(B|A) = \kappa(AB) - \kappa(A)$. Moreover, ranking functions can also be conditioned by propositions A via $\kappa|A(\omega) = \kappa(\omega) - \kappa(A)$, yielding a ranking function on the models of A. A conditional $(B|A)$ is accepted in the epistemic state represented by κ, written as $\kappa \models (B|A)$, iff $\kappa(AB) < \kappa(A\overline{B})$, i.e. iff AB is more plausible than $A\overline{B}$.[2] Ranking functions can be considered as qualitative counterparts of probability distributions. Their plausibility degrees may be taken as logarithmic order-of-magnitude abstractions of probabilities (cf. [20, 21]).

In the following, we take the concept of epistemic states for granted and elaborate on general notations that we use throughout this paper. So, in general, let Ψ be any epistemic state, specified by some structure that is found appropriate to express conditional beliefs from a suitable conditional language $(\mathcal{L} \mid \mathcal{L})^*$, in which conditionals may be equipped with quantitative degrees of belief, according to the chosen framework. For instance, for probability functions, $(\mathcal{L} \mid \mathcal{L})^* = (\mathcal{L} \mid \mathcal{L})^{prob} = \{(B|A)[x] \mid A, B \in \mathcal{L}, x \in [0,1]\}$, and in qualitative environments, e.g., for ranking functions, $(\mathcal{L} \mid \mathcal{L})^* = (\mathcal{L} \mid \mathcal{L})$. Moreover, an entailment relation $\models$ is given between epistemic states and conditionals; basically, $\Psi \models (B|A)^*$ means that $(B|A)^*$ is accepted in Ψ, where acceptance is defined suitably. Let $\mathcal{E}^* = \mathcal{E}^*_\Sigma$ denote the set of all such epistemic states using $(\mathcal{L} \mid \mathcal{L})^*$ for representation of (conditional) beliefs. Moreover, epistemic states are considered as (epistemic) models of sets of conditionals $\Delta \subseteq (\mathcal{L} \mid \mathcal{L})^*$: $Mod^*(\Delta) = \{\Psi \in \mathcal{E}^* \mid \Psi \models \Delta\}$. As usual, $\Delta \subseteq (\mathcal{L} \mid \mathcal{L})^*$ is *consistent* iff $Mod^*(\Delta) \neq \emptyset$, i.e., iff there is an epistemic state which is a model of Δ.

3 Inductive reasoning based on epistemic states and their revision

In this section, we develop our general approach to inductive reasoning as a special case of epistemic belief revision. Epistemic states serve as a mediator between reasoning and revision by providing both an epistemic background for reasoning and an ideal outcome of induction from and revision by (conditional) belief bases. So, in Subsection 3.1, we first discuss the general relation between arguments and inductive reasoning on the one hand and the dynamics of epistemic states on the other, emphasizing the crucial role of conditionals in this context. In a kind of digression in Subsection 3.2, we more carefully distinguish between inference and arguments and argue that the latter build on relevance relations, which, however, play no further role in this paper. Subsection 3.3 discusses the relation between epistemic dynamics and inductive reasoning in a bit more detail. Subsection 3.4

[2] The full definition here is $\kappa \models (B|A)$ iff $\kappa(AB) < \kappa(A\overline{B})$, or $A \equiv \bot$. For sake of simplicity, we exclude conditionals including contradictions from consideration here.

forms the constructive core of this section. It spells out the various forms the interaction between inductive reasoning and belief revision may take, with particular emphasis on the background beliefs in the form of conditionals which are previously accepted and on the role of a principle called Coherence guiding iterated change. The discussion of epistemic dynamics or belief revision as a framework for inductive reasoning must include the issue where this dynamics may start from. This refers us back to some initial epistemic state as pondered in Subsection 3.5. So, our philosophical tour is topped off by a discussion how we might conceive of such initial epistemic states. It will turn out that we can do so very much in line with our previous discussion of background beliefs.

3.1 Arguments, reasoning, epistemic states and conditionals

Let us start with looking at what happens in arguing with one another. When we give an argument, we start from some hopefully shared premises and infer a conclusion, which is then hopefully shared as well. Or the argument may be only hypothetical, where the epistemic status of the premises is left open. An inference proceeds in the very same way from premises to a conclusion. Or we may say that we reason from the premises to the conclusion. Then we might call the premises the reasons for the conclusion. These are equivalent ways of describing what is going on in an argument.

We should slightly restrict our topic right away. When we talk about arguments, we refer only to descriptive, factual, or empirical reasoning, where premises as well as conclusion are descriptive or truth-evaluable. However, we believe that everything we discuss here applies *mutatis mutandis* to normative, deontic, or evaluative reasoning, where premises and conclusions may be normative sentences the truth-evaluability of which is at least doubtful. We indeed think that there are close parallels, see [58]. However, this is a different large field we do not enter here.

We know well enough what a deductive argument, inference or reason is. There are computational or syntactic and semantic versions. Their mark is (guaranteed) truth preservation. We know how the versions work and we understand their relation. The problem is: most of our inferences and arguments are not deductive or truth-preserving. They are inductive, nonmonotonic, or defeasible. As mentioned, we use “inductive” as a general term for all these kinds of reasoning. When deductive logic dominated formal philosophy, there was the idea that inductive arguments are simply elliptic. They reduce to deductive arguments when implicit premises are made explicit. However, this idea is definitely misguided. We must engage in theories of inductive reasoning on their own, because inductive reasoning involves also defeasible and tentative, even creative processes. It explores the field of rationality beyond deductive logic.

We emphasize that we use “inductive” as a general term for all these kinds of reasoning. In the 19th century, the ’inductive method’ referred to an inference from the particular to the general, the paradigmatic inference being so-called enumerative induction. However, we need all kinds of non-truth-preserving inferences, we need all possible ways to infer what the world is like beyond of what we observe. These ways often include an inductive inference in the traditional sense, an inference from the particular to the general. But we must not presuppose that it is always so. E.g., our forecast of the results of the next election are not based on any putative generalizations of voters’ behavior. So, we are well advised to accept our broad sense of inductive reasoning.

The traditional offers accounting for inductive reasoning are not so rich. There is a long strand of grasping probabilistic inductive reasoning, culminating perhaps in Carnap’s inductive logic [6, 7]. Of course, Bayesian methods are by now well entrenched in all of our scientific canons. Enumerative induction – it indeed looks simplistic– was rather criticized than explicated (but see [56]). The strongest historic attempt at explicating inductive reasoning beyond probabilistic methods are John Stuart Mill’s methods of induction. However, only with the rise of conditional logic, default logic, etc. do we see attempts at grasping inductive reasoning beyond probabilistic methods.

Alas, in the meantime, there is a plethora of diverging, incompatible, or incommensurable accounts of inductive reasoning, forming a large and confusing field that is very difficult to evaluate. Let us try to state some guidelines helping to steer clear in this field.

How might we approach the topic? We take the following observation to be basic: When we give an argument or provide reasons, we try to convince our interlocutors of the conclusion of the argument or of what the reasons entail, not by talking them into the conclusion, but by appealing to their reasoning capacities that make them hopefully infer the same conclusion as we infer. So, the point of giving arguments or providing reasons is to induce rational belief change. This includes the limiting case of stabilizing or confirming, i.e., not changing the epistemic state. In short, reasoning is about the dynamics of belief or about epistemic dynamics in general.

The social dimension, however, is not really essential in our view. Of course, arguing is a social activity, just as language in general. However, for a theory of inductive reasoning, this dimension seems negligible. Sure, one does not argue with oneself, but one is engaged in reasoning or making inferences by oneself all the time. Giving arguments and reasons to others presupposes to have reasons and to have worked them out by oneself, to be convinced by one’s own arguments. And the latter is about individual rational belief change or revision.

There is also a hypothetical variant. When I argue from some assumed premises, I work out what to rationally infer from them, i.e., what to believe given the premises. So, arguments are about conditional belief or, more neutrally, about conditional epistemic states. Indeed, conditional epistemic states and the dynamics of epistemic states are closely related. In simple conditionalization (which can be stated for various epistemic formats) they are even the same; the posterior epistemic state after an input is the same as the prior epistemic state conditional on this input. (It is not always this simple. Still, any other rule of epistemic change we know of is based on the notion of a conditional epistemic state, e.g., conditional probabilities, conditional ranks, or whatever.)

The Ramsey test utilizes this observation for the semantics of the conditional. Indeed, one might say that the semantics of the conditional is the focal point of all our theories of inductive reasoning. All of this establishes a fundamental and close relation between arguments, reasoning and inference on the one hand and epistemic states, their dynamics, and conditionals on the other hand. The direction of the relation – is one relatum more basic than the other? – is still open. We will discuss this.

In the context of inductive reasoning and belief revision we are discussing, we want to take a pragmatic view on epistemic states. We assume the representation of epistemic states to be equipped with some meta-structures allowing to perform reasoning and belief revision in suitable logical frameworks, and we expect them to be complete in the sense that answers to all possible queries (in the respective) framework can be generated, to the best of the human's beliefs, i.e., no further thinking in the sense of exploiting given beliefs and information more deeply would yield a better result. Note that we use the term "revision" here in a general sense, as a synonym for any kind of epistemic dynamics integrating new information to one's current beliefs, i.e., as a super-concept also including update [28] or focusing [12]. When the specific change operator called revision in the AGM theory [1] is meant, we speak of "AGM revision", or specify this explicitly.

The Ramsey test directly utilizes the tight connection between reasoning and belief dynamics for stating a semantics of the conditional. A conditional is accepted in an epistemic state if after acceptance of the antecedent the consequent is accepted as well. In the meantime there are many variations of the Ramsey test. Thereby, we presuppose that epistemic states can evaluate conditionals to be accepted or not accepted. This is a crucial feature of modelling human's beliefs going beyond classical logic. We avoid saying that a conditional is *true* in an epistemic state, because we have above introduced conditionals not as binary but three-valued, and more importantly, because in the commonsense context of reasoning considered here conditionals do not work truth-functionally at all. Rather, to accept a conditional, humans would expect a meaningful connection between antecedent and consequent. This is crucial for our approach to inductive reasoning because this connection can be used for reasoning in a way that captures human-like thinking.

In this paper, we pursue the following well-established variant of the Ramsey test: A conditional $(B|A)$ is accepted if its verification AB is deemed to be more plausible, or probable, than its falsification $A\overline{B}$. The inherent connection between antecedent and consequent is taken into account by considering A and B resp. A and $\overline{B}$ jointly when assessing plausibility, or probability. Beyond plain comparison, also degrees of plausibility, or probability, can be assigned to verification and falsification, thereby measuring the strength of a conditional, if allowed by the respective semantic framework. (In section 2, we have already provided a notation for this measure.)

All in all, it seems appropriate to say that the semantics of the conditional is the focal point of all our theories of inductive reasoning. To resume, our remarks establish a fundamental and close relation between (i) arguments, reasoning and inference, (ii) epistemic states and their dynamics, and (iii) the logic of conditionals. The direction of the relation – is one relatum more basic than the other? – is still open. We will discuss this.

Formally, the upshot of our informal discussion is that in symbolic resp. qualitative frameworks, the fundamental connection between epistemic states, conditionals, plausibility, (inductive) reasoning, and belief revision on which this paper relies can be roughly expressed by the following equivalences:

$$\Psi \models (B|A) \quad \text{iff} \quad AB \prec_\Psi A\overline{B} \quad \text{iff} \quad A \mathrel{|\!\!\sim}_\Psi B \quad \text{iff} \quad \Psi * A \models B, \tag{1}$$

where Ψ is an epistemic state in $\mathcal{E}^*$, $\preceq_\Psi$ is a suitable relation expressing plausibility (or probability)[3], $\mathrel{|\!\!\sim}_\Psi$ is an inference relation based on Ψ, and $*$ is an epistemic (or iterative) revision operator that takes an epistemic state and a proposition and returns again an epistemic state (in the sense of [8]). In quantitative frameworks, degrees of beliefs must be suitably added. More generally, we assume that $*$ can also deal with more complex beliefs given by sets of conditionals Δ such that $\Psi * \Delta \in \mathcal{E}^*$. We also adopt the success postulate of AGM theory [1], i.e., we presuppose that $\Psi * \Delta \models \Delta$, meaning that $\Psi * \Delta \models \delta$ for all conditionals δ in Δ. This also includes the case of revision by a (plausible) proposition A via identifying A with $(A|\top)$, as assumed above. Equation (1) reveals that both epistemic states and conditionals

[3] Note that in qualitative semantic environments, e.g., ranking functions, lesser means more plausible, and this also complies with the readings in nonmonotonic preferential inference. Hence we stick to this tradition here. Of course, for probabilities and also, e.g., possibilities, the scales are inverted, so $\preceq_\Psi$ must be interpreted via the numerical $>$-relation. Technically, $A \prec_\Psi B$ iff $A \preceq_\Psi B$ and not $B \preceq_\Psi A$.

are also carriers of strategic information that become effective for reasoning and revision.

3.2 Arguments, inference, and relevance

In arriving at equation (1) we have more or less equated arguments and inference or reasoning. But, as a kind of digression, we might be a bit more careful here. The aim of inference is establishing a conclusion from given premises. If the conclusion is not established with certainty given the premises, as it is in deductive inference, it should at least be more plausible as its opposite. This is what guides equation (1). However, at least intuitively, an argument does more. It provides a *reason for* the conclusion.

Let, e.g., B be the proposition that we will not reach the climate target of keeping global warming below 1,5 degrees. B is highly plausible according to our present epistemic state Ψ, much more plausible than its opposite that we do reach this target. Now let $A =$ "Tom Cruise wins a special Oscar award", which is epistemically entirely irrelevant to B and does not influence the plausibility of B. According to (1), we then have $A \mid\!\sim_{\Psi} B$. We might still say that B follows from A, because B holds anyway. But it would be odd to say that A is a reason or an argument for B. Or let $A =$ "the US build fifty solar power plants", which is epistemically negatively relevant to B. It diminishes the plausibility of B a little bit, but certainly not far enough to make it less plausible than its opposite. We think that much stronger measures would be needed to reach the climate target. So, according to (1), we still have $A \mid\!\sim_{\Psi} B$. But now it would be even odder to say that A is an argument or a reason for B. Rather, it is an argument or a reason *against* B, though too weak to undermine B.

Hence, let us define that A is an epistemic *reason for* B iff A is epistemically positively relevant[4] to B iff A raises the plausibility of B iff $(B|A) \prec_{\Psi} (B|\overline{A})$, where $\preceq_{\Psi}$ is suitably lifted to conditionals. And the point is that arguments must provide reasons in this sense. That is, an argument is a structure with a premise or premises and a conclusion such that the premise or the conjunction of the premises is positively relevant to the conclusion.[5]. How does this relate to our basic equation (1)?

[4] Relevance is a multiply ambiguous notion. It is clear that in our context epistemic relevance as defined is the only pertinent kind of relevance.

[5] More precisely, this is the structure of a *single* argument. In order to assess a chain of arguments, we would have to study the conditions under which positive relevance spreads along the chain. And in order to study the interaction of arguments, how they defeat, rebut or undermine one another, we would have to study how positive relevance behaves under the augmentation of premises. Here, we do not pursue this study. However, the remark is to express our skepticism that this interaction can be studied in the abstract, as is done in argumentation theory, e.g., according

There are two ways to respond. First, we might discriminate between inference and arguments and still stick to (1) concerning inductive inference, while assigning this explication of an argument to the realm of argumentation theory. This is what we shall do here. Second we might strengthen our notion of inference and our semantics of conditionals by adding the positive epistemic relevance condition. Then we would define $\Psi \models (B|A)$ and $A \mid\!\sim_{\Psi} B$ as $AB \prec_{\Psi} A\overline{B}$ *and* $(B|A) \prec_{\Psi} (B|\overline{A})$. Thereby we enter the topic of so-called relevance conditionals, which were recently studied in great detail, see, e.g., [51, 47, 50, 48]. However, we do not pursue here this line of thought.

Is positive epistemic relevance a good explication of the notion of a reason? In any case, it is a subjective explication, entirely dependent on the subject's epistemic state. For the majority of philosophers this is insufficient. They seek to gain a more objective notion of a reason, even of a good reason. For them, rational epistemic dynamics is driven then by those preconceived good reasons. However, they were constructively quite poor in specifying what good reasons are. And the history of inductive skepticism teaches that this might not be easy. With the subjective understanding, we have at least a workable precise explication of that notion suitable for theorizing. We approximate "good reason" here in an abstract way by considering logic-based reasoning methodologies that are equipped with qualitative meta-information allowing for expressing what is good and what is not. And it is still true that epistemic reasons drive epistemic dynamics. Indeed, our definition entitles us to reversely say that epistemic reasons are whatever drives rational epistemic dynamics.

3.3 Inductive reasoning and epistemic dynamics

So much about a possible amendment of the basic equivalence (1) by positive relevance considerations. In the sequel, however, let us just develop that equivalence. It still leaves open how precisely to understand the relation between arguments and inductive reasoning on the one hand and epistemic dynamics and belief revision on the other. In particular, it raises an issue of primacy: Are we first to spell out inductive logic and rational reasoning? And are we thereby to ground an account of rational epistemic dynamics? Or is it the other way around?

We are skeptical of substantially implementing the first direction. It has an objectivistic flair: there is a correct inductive logic, and our epistemic states have to follow it. However, in the tradition of inductive skepticism this objectivism is discredited. The chief witness is perhaps the decline of Carnap's program of inductive

to [14].

logic, which became weaker and weaker till it became almost indistinguishable from de Finetti's subjectivism; see [6, 7].[6]

We therefore favor the second direction: a theory of rational doxastic dynamics should be provided first, from which then an account of inductive reasoning should be derived. This is our first important specification of (1). This entails that any account of inductive reasoning must be based on a specific conception of epistemic states and their rational dynamics. This is made explicit here by using the symbol $\mid\!\sim_{\Psi}$ for the inference relation. Not all accounts pay heed to this maxim. Still, a variety of potentially suitable accounts remain.

A further observation is that the equivalences in (1) presuppose that epistemic states must be conceived as coming in grades that are (partially, weakly, or strictly) ordered. That is, an epistemic state Ψ must provide a plausibility ordering $\preceq_{\Psi}$ in the sense that there are faithful assignments (similar to [28, 8]) that associate suitable meta-structures with an epistemic state. This may exclude further candidate accounts (e.g., the representation of an epistemic state plainly as a set of beliefs or as a propositional knowledge base). Hence, the derivation of an account of inductive reasoning from a conception of epistemic states entails some substantial constraints.

Let us be a bit more specific concerning what we expect from the meta-structures associated with an epistemic state. A purely qualitative preorder might be a suitable meta-structure that is associated with an epistemic state. Of course, there are more sophisticated representation frameworks, such as possibility theory, ranking functions, and probability functions. But also modal logical frameworks seem to be good candidates for representing epistemic states, or heterogeneous structures consisting of different components (with reasonable interactions between them) might prove useful. This is not necessarily a question of numerical or symbolic representation, both types of frameworks can be fine.

But when it comes to numbers, it should be clear that the crucial point here is not just their potential for a richer semantics. Rather, they definitely provide richer structures that computations for information processing might utilize. And this makes them quite distinguished candidates for epistemic states in the context of reasoning and belief change. It is not by accident that probability theory with its two independent arithmetic operators (addition and multiplication, both full group operations) have played a major role here. Although AGM might have marked the beginning of symbolic belief revision and of devising rational postulates for belief change, actually performing belief change has been done for a much longer time within the probabilistic framework. The first belief change operator ever is

[6] However, we should at least point to the efforts of Williamson [60]. See also our discussion in section 3.5.

probabilistic conditioning, and Jeffrey's rule [41] shows a possible way of incorporating even uncertain evidence. So, it is not because of the numbers that we should value probability theory, but because of the rich arithmetic structure that provides a powerful apparatus to express and process information (cf. also [41]). Via the multiplication operator, (conditional) independencies (and hence monotonic inference behaviour) can be expressed, and its inverse operator, division, allows to easily transform one distribution into another at the occurrence of new information via conditioning. Furthermore, the addition operator takes care of disjunctive propositional information, e.g., to allow for reasoning by cases in a way that takes the probabilities of all cases into account. Having once adopted such basic techniques, information processing becomes easy. However, ranking functions show similarly good properties, here we have the (group operation) addition instead of multiplication, and the minimum of ranks instead of addition of probabilities. The minimum is weaker than the addition, it is not a group operation and does not allow for exploiting numerical relationships in a way that addition does. For instance, consider two atoms A, B and the propositions $A \wedge B, A \wedge \neg B, \neg A \wedge B$. In probability theory, if we know $P(A) = P(B)$, we can conclude $P(A \wedge \neg B) = P(\neg A \wedge B)$ because $P(A \wedge B) + P(A \wedge \neg B) = P(A) = P(B) = P(A \wedge B) + P(\neg A \wedge B)$. But from $\min\{\kappa(A \wedge B), \kappa(A \wedge \neg B)\} = \kappa(A) = \kappa(B) = \min\{\kappa(A \wedge B), \kappa(\neg A \wedge B)\}$, we cannot conclude $\kappa(A \wedge \neg B) = \kappa(\neg A \wedge B)$. Technically, this has significant effects on reasoning and revision. On the other hand, evaluating minima is computationally and cognitively less demanding, which might be seen even as an advantage of ranking functions.

We have to clarify the relation between inductive reasoning and the dynamics of epistemic states and thus to specify (1) still further. There is a distinction regarding this dynamics that is often neglected, but seems important to us. On the one hand, there is an internal dynamics which takes place without any external stimulus or input. It consists in thinking, reasoning, calculating, working out consequences, etc. All this in some sense amounts to a temporary or only hypothetical change of an epistemic state. *Inference rules* then tell us how the internal dynamics should proceed. On the other hand, there is an external dynamics which is driven by some external input, information, evidence, experience, not merely in the sense that such external input somehow stimulates the internal dynamics – of course, it does –, but in the sense that the input demands a change of the prior into the posterior epistemic state, however and however incompletely this change is computationally realized. *Rules of doxastic change* then tell us what the posterior state should be depending on the prior state and the input. Note that the internal dynamics belongs to the external statics. Thinking, etc. does not count as doxastic change in the external sense. According to the internal dynamics, an epistemic state records the current state of computation. According to the external dynamics, an epistemic state is an ideal entity which sets a computational goal and may or may not be fully reached by the computations in internal dynamics. Typically, logics, in their many variants, deal with the internal dynamics, by specifying calculi, inference rules, etc., on the syntactic level. By contrast, Bayesianism, belief revision theory, etc. are about the external dynamics. Their dynamic rules specify the relation between prior state, input, and posterior state on a semantic level. Their primary point is not to give computational advice, even if this can often be easily derived.

The connection between the two dynamics is this: The internal dynamics works towards reaching the goal set by the external dynamics, i.e., the posterior state necessitated by the input. Again, the connection may be construed in two opposite ways. Either the input initiates an internal dynamics, the completion of which results in some posterior state, which is then the one the external dynamics aims at. Given the internal dynamics, we can say what completion means (roughly, that any query can be answered in a most informed way so that further thinking or computation does not result in further internal change). Or the input necessitates an external change which then governs the internal dynamics (as being one the completion of which leads to the necessitated result).

We think that the second construal is the one to be preferred. This is our second important specification of (1). For, how could the inference rules be justified within the first construal? By being consistent? By intuition? By some model theory unrelated to epistemic dynamics? No, the justification lies in fixing the goal of computation by specifying a rational external dynamics. The internal dynamics then serves this goal; it is only a means to this end.[7]

We admit that the distinction is often subtle. Conditionalization rules directly tell how to compute the posterior state from the prior state and the evidence to condition on. Or: What is the difference between Rational Monotony (an axiom of conditional logic and nonmonotonic inference) and the postulate K*8 (also called subexpansion and crucial in AGM belief revision theory)? Via the Ramsey Test, they are directly intertranslatable. Still, they have different places in the overall picture of doxastic dynamics.

Let us summarize our two important claims so far: When we want to get a hold on inductive logic, we must start from an account of the rational dynamics of

[7] The work of Pollock [43] is characteristic for this opposition. [42] specifies argument types, and [43] then states rules for the interaction of arguments like the weakest link principle or the no-accrual-of-reasons principle. All of this belongs to the internal dynamics in our sense. From this, an account of belief revision, of the external dynamics, is inferred precisely by running this mechanism of argument types and rules of interaction on the new input till it comes to rest (if it does); see [44]. For a detailed criticism of the entire procedure in the direction indicated see [54].

epistemic states, which in turn presupposes a graded notion of a conditional doxastic state. Indeed, we must start from an account of the rational external dynamics of epistemic states, which sets the goal for the internal dynamics and thus for inference, reasoning, and argumentation.

3.4 The interaction of inductive reasoning and belief revision

If we understand inductive reasoning as completing partial (conditional) beliefs (as specified in a belief base Δ) as best as possible, then its result should be an epistemic state Ψ_Δ:

$$\Psi_\Delta = \mathit{ind}(\Delta), \tag{2}$$

where *ind* is some inductive reasoning mechanism; we also say that Δ is *inductively represented* by Ψ via *ind*, or that Δ *inductively generates* Ψ. For instance, Δ may be a set of conditionals, and *ind* might be specified by system Z [21], or c-representations [32], associating to each consistent set of conditionals a ranking function [53]. Inductive reasoning from Δ is then implemented by reasoning from $\Psi = \mathit{ind}(\Delta)$ via the conditionals being accepted in Ψ. That is, *ind* realises *model-based inductive reasoning.*

But this cannot be the end of the story. The mind of a human being is always evolving and changing by learning, or receiving new information $\mathcal{I}$ in general, where $\mathcal{I}$ can just be a fact, more complex contextual information possibly including conditionals (e.g., when we enter a new country, different compliance rules apply), or even trigger some deeper learning processes.

Starting a new inductive reasoning process each time when we receive new information would make our beliefs incoherent, $\Psi = \mathit{ind}(\Delta)$ and $\Psi' = \mathit{ind}(\mathcal{I})$ might be completely unrelated (except for that they have been built up by the same inductive reasoning formalism). Integrating new information $\mathcal{I}$ into existing beliefs represented by an epistemic state Ψ is exactly the task of (epistemic or iterated) belief revision [8], returning a new epistemic state Ψ' after revising Ψ by $\mathcal{I}$:

$$\Psi_\Delta * \mathcal{I} = \mathit{ind}(\Delta) * \mathcal{I} = \Psi' \tag{3}$$

Note that we use $*$ here in a generic sense as a placeholder for a suitable change operator. What can we say about this change operator $*$, i.e., about the rational dynamics of epistemic states?

The most natural and the most wide-spread picture is that this dynamics is a kind of Markov process: The prior doxastic state and the (total) evidence (in between) determine the posterior doxastic state – according to rules of doxastic change that count as rationally justified. This is Markovian in the general sense that the prior state is supposed to encode a history of changes and hence this history can influence the current change process only through that prior state.

There is a very rich discussion about the rules governing rational epistemic change. In probability and ranking theory, there are rules of conditionalization, simple, generalized Jeffrey, and auto-epistemic conditionalization. There are reflection principles governing doxastic change in both theories. There is minimization of relative entropy, which has an analogue in ranking theory. And so on. We will see that our approach based on equations (1), (2), and (3) leads to formal constraints on doxastic change and its interaction with inductive reasoning that narrows down the range of suitable epistemic frameworks.

Let us add just three general remarks: First, there are many proposals for modelling epistemic states; we have mentioned a few of them in the course of this paper. In probability theory and also in ranking theory the discussion about rules of epistemic change is most elaborate. It would be desirable that it is carefully worked out also within other models, since stating a dynamics is imperative for any representation of epistemic states.

Second, doxastic states are not only about 'eternal' or context-independent propositions, which are usually taken as the only objects of epistemic states, but also, indeed essentially, about indexical or context-dependent propositions, which use, e,g„ "I", "now", and "here", the reference of which can only be determined in context. How do rules of doxastic change apply to them, and how do they interact with rules for 'eternal' propositions? These questions seem to receive only local attention. See, e.g., [57] and [15]. In our approach, indexical information can be part of the contextual beliefs, while 'eternal' information in the sense of generic beliefs would be part of the background beliefs.

Third, all the rules we mentioned concern learning or improving one's epistemic state. This is perhaps the only case that is relevant for the sciences. Still, it is a restriction. There are other kinds of epistemic change, and a theory of rationality should attend to them, too. In particular, we are thinking here of forgetting. As such, forgetting befalls us, there are not rational and irrational ways of forgetting. But there are rational ways of responding to forgetting. Not anything goes after having forgotten something; see, e.g., [57]. Some conceptual considerations and technical results about the role that ranking functions can play in the context of forgetting can be found, e.g., in [34, 35, 3]. However, in the present context which is about the basic connection between epistemic dynamics, induction, and conditionals, this is just a side remark.

So much about the change or revision operator $*$ by itself. Let us return to this basic connection. Given that $\Psi_\Delta = \mathit{ind}(\Delta)$ has been built up inductively from a belief base Δ, and that $\mathcal{I}$ will usually be only partial information about some current context, the following questions arise immediately: How do *ind* and $*$ interact? Which (maybe completely different) roles do Ψ_Δ, Δ and $\mathcal{I}$ play in this scenario?

We first focus on the second question by analysing different qualities of beliefs with respect to the roles they play in the reasoning process. Roughly, we can distinguish between background, or generic, and evidential, or contextual knowledge, as well as between explicit and implicit beliefs. From background or generic knowledge, the agent takes beliefs which hold in general and of which she can make use of in different situations. For instance, the current beliefs of an agent getting up on a usual Monday morning might be different from those on a usual Sunday, but presumably his generic background has not changed much. The evidential resp. contextual information $\mathcal{I}$ she receives might include that it is Monday and raining, and that due to new construction areas she has to take some detours when going to work. We prefer the attribute "contextual" to "evidential" in the following, since this information may relate not only to a specific situation and can be much more complex than some evidential facts. For instance, the temporal scope of context may be one hour or one week, the scope may refer to a specific house or to a whole country, or it may contain information on abstract contexts, such as holidays or working environments.

Let us now look more closely at the first question, the interaction between *ind* and $*$. Assuming that $\Psi_\Delta = ind(\Delta)$ expresses background beliefs, incorporating contextual information cannot be done simply via the "union" of Ψ_Δ and $\mathcal{I}$ (whatever this might be), or by the union of Δ and $\mathcal{I}$ because this would ignore the different natures of background beliefs and contextual information. The agent's new epistemic state should rather arise from the adaptation of Ψ_Δ to contextual information. This is expressed by (3), but only as a base case when we start reasoning from a belief base including our core background beliefs. However, this process must be iterative, i.e., $\Psi = \Psi_\Delta$ may more generally be the result of such a revision $\Psi = \Psi_{prior} * \mathcal{I}_{prior}$, or new information $\mathcal{I}'$ arrives that triggers a new change process $(\Psi_\Delta * \mathcal{I}) * \mathcal{I}'$, so that (3) evolves to the iterative change problem

$$(\Psi_\Delta * \mathcal{I}) * \mathcal{I}' = (ind(\Delta) * \mathcal{I}) * \mathcal{I}'. \tag{4}$$

And here, three essentially different reasoning resp. revision scenarios are possible (note that the $*$-operators are just placeholders to be specified adequately):

- First, the context to which $\mathcal{I}$ refers has evolved, and $\mathcal{I}'$ is information on this new context for which, however, $\mathcal{I}$ is still relevant. This scenario is often referred to as *updating*. Then the two $*$-operators in (4) would be of the same type, and $\Psi_\Delta * \mathcal{I}$ would be changed to $(\Psi_\Delta * \mathcal{I}) * \mathcal{I}'$. A modification of this scenario applies if the contexts to which $\mathcal{I}$ and $\mathcal{I}'$ refer are completely unrelated, but the agent uses the same background beliefs Ψ_Δ for reasoning, then we would end up with $\Psi_\Delta * \mathcal{I}'$.

- Second, $\mathcal{I}'$ refers to the same context as $\mathcal{I}$. In this case, $\mathcal{I}$ and $\mathcal{I}'$ should be considered to be on the same level, and we would obtain $\Psi_\Delta * (\mathcal{I} \cup \mathcal{I}')$. This is a typical case of *belief revision* in the AGM-sense that we will call *conservative revision* because more prior information (i.e., $\mathcal{I}$) is preserved. Note that conservative revision generalizes Jeffrey's rule [27] to the case where several observations are processed at the same time, without presupposing that the observations are exclusive.

- Third, $\mathcal{I}'$ enriches or modifies background beliefs, i.e., it affects the basis from which reasoning with the information $\mathcal{I}$ is performed. This is what happens in *learning*. In the first case, if $\mathcal{I}'$ is fully compatible with Δ, $\mathit{ind}(\Delta \cup \mathcal{I}') * \mathcal{I}$ would be a proper solution. If $\mathcal{I}'$ contradicts (parts of) Δ, then $\Psi_\Delta * \mathcal{I}' = \mathit{ind}(\Delta) * \mathcal{I}'$ would provide suitable background beliefs, and $(\mathit{ind}(\Delta) * \mathcal{I}') * \mathcal{I}$ would be the result of the revision problem.

Therefore, we argue that the distinction between revision and update [28], and also the relation between belief change and learning is not just a technical issue, but has to be made on a conceptual and modelling level. The involved revision operators $*$ might respect such differences, but from the discussion above it becomes clear that one might also discriminate different ways of applying one and the same revision operator $*$ in different scenarios, also involving inductive reasoning. While (3) claims that involving belief revision is necessary for a coherent perspective of inductive reasoning, the third of the cases elaborated above shows how inductive reasoning can affect belief revision: Changing $\mathit{ind}(\Delta)$ to $\mathit{ind}(\Delta \cup \mathcal{I}')$ makes the revision of background beliefs possible. For more formal investigations of the differences between conservative revision and update, and for a reconciliation with AGM theory, see [33].

So, starting from an induction perspective, we developed scenarios which are similar to the ones considered in [10] for belief revision: Belief Revision as Defeasible Inference (BRDI), considering a specific case at hand, can be realized as conditioning in our framework, or more generally, via an update where the context has changed. Belief Revision as Prioritized Merging (BRPM), which collects several pieces of uncertain evidence about a case, is realized via conservative revision; please note that it is also possible to apply merging operators instead of simple set union if one wishes to do so. And finally, Revision of Background Knowledge by Generic Information (RBKGI), where the background knowledge is modified by new pieces of (generic) information often in the form of conditionals is also dealt with extensively in Section 3.4.

However, and in contrast to that paper, a main point of our approach is that these "scenarios of belief revision" are not unrelated, but can be realized coherently and naturally (with many interactions) in a rich framework of epistemic revision (see Section 3.4). Our approach is not about technical artefacts that coincidentally bring forth useful results but is grounded on philosophical considerations that clearly show that (inductive) nonmonotonic reasoning and belief revision are not just "two sides of the same coin", but that inductive reasoning is an integral part of epistemic revision in a conditional framework where principles of inductive reasoning follow more general principles of revising epistemic states by conditional beliefs.

Elaborating further on this intimate connection between inductive reasoning and belief revision, we might even envisage inductive reasoning involving background beliefs expressed by an epistemic state Ψ_{bk}, i.e., $\Psi = \mathit{ind}_{\Psi_{bk}}(\Delta)$, and then inductive reasoning from Δ might be realised by revision:

$$\Psi = \mathit{ind}_{\Psi_{bk}}(\Delta) = \Psi_{bk} * \Delta. \tag{5}$$

And when no background beliefs are available or relevant, we might assume some uniform epistemic state Ψ_u as a starting point (but see also the discussion on prior and initial states in Section 3.5 below):

$$\mathit{ind} = \mathit{ind}_{\Psi_u}. \tag{6}$$

This implements inductive reasoning from epistemic states thoroughly via epistemic belief revision because this approach yields

$$\Psi_\Delta = \mathit{ind}(\Delta) = \Psi_u * \Delta. \tag{7}$$

This means that each epistemic revision operator that is able to handle complex information Δ induces an inductive inference operator. This makes inductive reasoning perfectly coherent with the revision operator and allows us to embed inductive reasoning in a richer methodology.

This embedding has two further important advantages: First, revision methodologies may immediately yield mechanisms of inductive reasoning and suitable quality criteria. Second, splitting up inductive reasoning clearly into its inductive mechanism, its involved background beliefs, and context-based beliefs makes formalisms more explicit and more broadly (and flexibly) applicable. However, only very few approaches to epistemic revision with sets of conditionals exist; in Section 4, we briefly present the principle of minimum cross-entropy for probabilities and, a bit more extensively, the c-revisions for ranking functions as suitable methodologies on the base of which inductive reasoning in the respective semantic frameworks can be realised in a straightforward way.

Our approach to inductive reasoning via belief revision sketched above also distinguishes between explicit beliefs in a belief base, and implicit beliefs derivable in an epistemic state. The necessity of such a distinction is quite obvious in a belief change scenario, since implicit resp. derived beliefs are more easily changed than explicit beliefs. Having to give up explicit beliefs not only needs more effort, but it is quite a different thing. Formally, if $\Psi_\Delta = \mathit{ind}(\Delta)$, and the new information $\mathcal{I}$ is in conflict with Δ, e.g., $\Delta \cup \mathcal{I}$ is inconsistent, then we are still able to perform revision in the sense of updating via $\Psi_\Delta * \mathcal{I} = \mathit{ind}(\Delta) * \mathcal{I}$, whereas conservative revision via $\mathit{ind}(\Delta \cup \mathcal{I})$ would not be possible. If the agent comes to know that an explicit belief is (presumably) false, she might react more reluctant to incorporate it, trying perhaps to collect more evidence etc. If finally, she is ready to believe the new information, there are three possibilities: In the first case, the new information $\mathcal{I}$ might contradict the derived beliefs in Ψ_Δ but is nevertheless consistent with Δ, conservative revision $\mathit{ind}(\Delta \cup \mathcal{I})$ would be a suitable option. In the second case, the agent acknowledges that her previous explicit beliefs were erroneous before, in which case she has to perform a proper belief base change by applying merging techniques which are able to resolve conflicts, i.e., we would have $\Psi_\Delta * (\mathcal{I} \circ \mathcal{I}')$ with a merging operator $\circ$. This would give rise to a variant of conservative revision which is neither truly conservative nor prioritized, we leave this for future work. In the third case, the agent admits that the current context has changed, and she has to adapt her beliefs to these changes, in which case one would find some updating process appropriate. Summarizing, our approach to inductive reasoning is able to deal with (and properly distinguish between) generic, background and contextual beliefs, on the one hand side, and explicit and implicit beliefs, on the other. This is made possible by considering inductive reasoning within belief revision frameworks, and provides perfect grounds for a richer methodology that ensures coherence over different reasoning scenarios.

Furthermore, we mention an axiom for iterated revision that is particularly suitable to express coherence in the above sense, but which was considered only in very few of the current belief revision frameworks and introduced under the name *Coherence* in [29], where it plays a crucial role for characterizing the principle of minimum cross entropy, but actually goes back to [52]

(Coherence) $\Psi * (\Delta_1 \cup \Delta_2) = (\Psi * \Delta_1) * (\Delta_1 \cup \Delta_2)$.[8]

[8] Coherence of revision corresponds to path independence of contraction, which was introduced by [24]. Both postulates deal with the iterated revision resp. contraction by sets of propositions where one set is a subset of the other. They express how the overall result can be computed from

(Coherence) demands that adjusting any intermediate epistemic state $\Psi * \Delta_1$ to the full information $\Delta_1 \cup \Delta_2$ should result in the same epistemic state as adjusting Ψ by $\Delta_1 \cup \Delta_2$ in one step. The rationale behind this axiom is that if the new information drops in in parts, changing any intermediate state of belief by the full information should result unambigously in a final belief state. So, it guarantees the change process to be *coherent.*

Note that (Coherence) does not claim that $(\Psi * \Delta_1) * \Delta_2$ and $(\Psi * \Delta_1) * (\Delta_1 \cup \Delta_2)$ are the same. On the contrary, these two revised epistemic states will usually differ in general, because the first is not supposed to maintain prior contextual information, Δ_1, whereas the second should do so, according to success. However, (Coherence) can help ensuring independence of parts of the history that serves as background beliefs for inductive reasoning. In the situation described by (5) where we reason inductively from Δ with background (or prior) beliefs Ψ_{bk}, imagine that we still are aware of the last conditional information Δ_0 that shaped Ψ_{bk}, i.e., $\Psi_{bk} = \Psi_1 * \Delta_0$, which would be mandatory to be able to distinguish among the different scenarios sketched above. But in general, it will be the case that Ψ_{bk} and Δ_0 do not determine Ψ_1 uniquely, so that there may be a different Ψ_2 satifying also $\Psi_{bk} = \Psi_1 * \Delta_0 = \Psi_2 * \Delta_0$. For updating Ψ_{bk}, this is irrelevant because only Ψ_{bk} matters. However, for conservative revision, we would like to compute $\Psi_{bk} * \Delta = \Psi_1 * (\Delta_0 \cup \Delta)$, but also $\Psi_2 * (\Delta_0 \cup \Delta)$ would be a suitable candidate. Here (Coherence) guarantees that the resulting epistemic state would be the same:

$$\Psi_1 * (\Delta_0 \cup \Delta) \quad = \quad (\Psi_1 * \Delta_0) * (\Delta_0 \cup \Delta) = (\Psi_2 * \Delta_0) * (\Delta_0 \cup \Delta) = \Psi_2 * (\Delta_0 \cup \Delta).$$

This makes clear that in our conceptual framework of inductive reasoning in the context of belief revision, integrating background beliefs and different pieces of information can be done in different, but coherent ways. This means, having to deal with different pieces of information, the crucial question is not whether one information is more recent than others, but which pieces of information should be considered to be on the same level, i.e., belonging to the same type of belief (background vs. contextual), or referring to the same context (which may, but is not restricted to be, of temporal type). Basically, pieces of information on the same level are assumed to be compatible with one another, so simple set union will return a consistent set of formulas (please see also our remarks on merging on p. 20). Pieces of information on different levels need not be consistent; here later or more reliable ones may override those on previous levels.

intermediate results, i.e., from intermediate points on the revision resp. contraction path.

3.5 Prior, initial and a priori epistemic states

The above picture of doxastic dynamics does not only appeal to rules of change, it also generates a regress to earlier doxastic states. Even allowing background and context to do their work only reiterates the regress. We reason forward from a certain background and in a certain context. But background and context generate an epistemic state from a still more prior state. So, where does this regress lead to? After all, the epistemic states do not come from an infinite past. Let us say that the regress comes to an end at an *initial* epistemic state from where the dynamics starts. But this is only a label. The question is whether we can characterize the initial state in some reasonable way.

The initial state is crucially important, because it and the course of experience fixes all further doxastic states, at least when the learning rules are deterministic. All our rational learning strategies are already encapsulated in this initial state. So, what can we say about it? This is an old and intriguing philosophical issue. In an absolute sense, 17th century philosophy spoke of innate ideas. This was the kind of preformation of our mind discussed between empiricism and rationalism in those times. The talk of innate ideas can certainly not be taken literally. They did not refer to the newborn baby's doxastic state. The discussion advanced with Kant. He may be taken to suggest that the initial state in an absolute sense consists of a priori knowledge. For him, apriority was an epistemological, not a genealogical category. In today's terminology, we may call the initial state conceived in this absolute way the all-embracive *ur-prior*. Alternatively, we may have a low-key relative understanding of the initial state. Then it is just posited to be initial for a given application at hand, where we are at the beginning of an invesigation, and not intended as an all-purpose ur-prior. Let us call this an application-relative conception of initiality. It is definitely closer to current practice, but perhaps not the foundational response we are looking for.

There is a philosophical debate whether the initial state is rationally unique or whether various initial states may be rationally permitted.[9] The opposition is not designed for an application-relative understanding of the initial state. Prima facie, any kind of constraints on the initial state may be imposed depending on the application at hand, and so the issue of Uniqueness does not really arise. Philosophers rather discuss the issue regarding the absolute ur-prior. No doubt, Uniqueness may look attractive. Sometimes, the debate between Uniqueness and Permissiveness is taken to be a symmetric one. Each side has to advance arguments for its claim. In our view, however, the burden of proof is only on the defenders of Uniqueness. It's not that the defenders of Permissiveness have to positively show that various

[9] See, e.g, [23] and [37].

ur-priors are equally rational. They only see no reason to be presently convinced of Uniqueness. Given the fate of Carnap's inductive logic [6, 7], the constructive outlook of Uniqueness is indeed dim. Carnap started as a defender of Uniqueness. However, he immediately recognized that his so-called Wittgenstein function was a total failure. Then he came up with his so-called λ-continuum of inductive methods, and in [6, 7], he ended up with some symmetry principles, which were still extremely permissive.

Note also that Uniqueness would entail an alternative picture of the doxastic dynamics. Any doxastic state is then only a function of the total evidence since the initial time. We only need to take stock of the accumulating evidence. References to any intermediate doxastic states are no longer required. This picture is no longer Markovian. This is how objective Bayesianism as propounded by Williamson [60, 38] conceives of inductive logic. He certainly pursues only a modest application-relative understanding of the initial state, but he thoroughly applies maximum-entropy reasoning both to the initial state as well as to how the total evidence changes the initial state. (The evidence need not be propositional, but can provide any kind of constraints on the posterior state.) Objective Bayesianism is certainly the most constructive attempt to establish Uniqueness. However, let's not further discuss its prospects.

Let us rather look a bit more closely at the initial state by ourselves. We mentioned its relation to the a priori, if taken in the absolute sense. However, so far we referred only to the Kantian a priori. In Kantian terms it is characterized by absolute necessity and generality. We better do not engage into Kant exegesis. A better and indeed fitting characterization is in the present terms of belief dynamics: A priori propositions are just those believed under any circumstances, *whatever the evidence*. Therefore, it is apt to call them unrevisably a priori. Certainly, every initial state must respect this kind of apriority. However, it cannot fully characterize initiality, since it is inductively barren. It cannot tell anything about inductive inference and thus misses what we are after here.

There are strong suggestions in the literature that there are also weaker relative or contextual notions of apriority (see, e.g., [45, 16]). One idea is to relativize apriority to the concepts we have. E.g., Kant may have been right about the apriority of Euclidean geometry, but only as long as there was no other conception of physical space.[10] This kind of apriority may be aptly called defeasible. Such are the beliefs or, in general, the features of initial doxastic states *before any evidence*, which may change afterwards. Apriority in this sense does not entail truth, such beliefs may turn out to be false. The historically first clear example for defeasible apriority is the

[10] This is the suggestion of Putnam (1962, pp. 372f.)

so-called principle of ignorance dictating symmetric or uniform probabilities, which are defeasible, sensitive to experience. Of course, we know by now how elusive the principle of ignorance is. If uniformity, as assumed in (6), is a feature of the initial epistemic state, it may need heavy qualification.

So far, though, defeasible apriority is just another label for the initial epistemic state. Let us at least give some hints how we might say a bit more about it. Above we said that a background may contain, or a context may provide, a lot of conditional information, which is then used for inductive inference. We have indicated how this may technically work in Section 3.4, where we also showed that the property of (Coherence) allows for reducing the impact of the concrete form of the initial state on future revisions significantly. We suggest that the same is true for the initial state. The idea is this:

A subject's doxastic state, however modelled, operates on a given algebra of propositions which are generated from a given conceptual field. This field need not consist of all concepts the subject possesses. It may be a small field just grasping the application at hand. But the subject must master those concepts; she cannot have doxastic attitudes towards propositions she does not understand. Then we might conceive of an initial doxastic state about (an algebra of propositions generated by) a given conceptual field as consisting just of what is required to master this field, but without any further information or evidence concerning those propositions. This would explicate the phrase that the initial state (concerning this field) is one the subject is in before acquiring any (relevant) evidence.

Of course, the explication can't mean that the subject has had no experience whatsoever in such an initial state. She needs a lot of experience in order to acquire any concept at all. However, it is hard to say how much information exactly she must have gathered in order to count as possessing a certain concept. Therefore, the explication is bound to be vague. Still, we think that the notion of a stereotype introduced by Putnam [46] is useful here. One must have learned the relevant stereotype in order to be said to possess a concept. One masters the concept of a dog only if one believes that some (ostensive) paradigms are dogs, that dogs have a certain variable size and shape, that they bark, that they have four legs and a tail, and so on. All this is not unrevisably a priori, it may turn out false in specific cases. But it comes along with the concept of a dog and may thus be called defeasibly a priori.

Such stereotypes are ubiquitous. Our prime examples are dispositional concepts. Reduction sentences, e.g., "an object, when put in water, is soluble if and only if it dissolves", are stereotypes. A disposition typically shows its manifestation. But it may be present, while the manifestation fails, and the other way around.

Thus, reduction sentences are defeasibly a priori conditionals[11]. This is now not just an application-relative a priori, but more strictly a concept-relative a priori. And it promises to answer our quest for a characterization of initiality. Formally, however, it works like any presupposed background or contextual information, be it in conditional or unconditional form. This is how the present point connects up with the explanations in the previous sections.

4 Proofs of concept: Reasoning on optimum entropy and with ranking functions

In a purely qualitative setting, epistemic states can be represented by systems of spheres [39], or simply by a preorder on $\mathcal{L}$ (which is mostly induced by a preorder on worlds). However, for this type of epistemic states, no methodologies are available to date which can handle the complex scenarios of inductive reasoning and belief revision that we sketched in Section 3. Therefore, we choose probabilities and ranking functions as illustrations of our general concept. We briefly describe two well-known revision methodologies in these two frameworks which induce approaches to inductive reasoning from conditional belief bases that have also attracted much attention: The probabilistic principle of minimum cross-entropy (with the principle of maximum entropy as the method for inductive reasoning), and c-revisions of ranking functions (with c-representations allowing for inductive reasoning). Since there is already a vast literature on the entropy principles while ranking functions and c-revisions are less well-known, we focus on the latter approach here. Note that c-revisions have been introduced in a more general form in [30, 32], for the sake of ease of notation, we only use a simplified version of c-revisions here which is nevertheless able to capture all aspects of our approach.

Both revision operators are linked by the property that they both satisfy the *principle of conditional preservation*, as specified e.g. in [30, 32]. This principle makes use of the arithmetic structures underlying probabilities and rankings and allows a very accurate and precise handling of conditional information under belief change. We do not go into technical details here, but the structural similarity between the operators is obvious (see equations (8) and (9) below, keeping in mind that ranks can be understood as logarithmic order-of-magnitude abstractions of probabilities, hence exponents become factors, and products turn into sums when going from probabilities to rankings).

[11] The point is elaborated in [55], sect. 13.3.

4.1 Reasoning on optimum entropy

The principles of maximum entropy and minimum cross-entropy are powerful methodologies for inductive reasoning and belief revision in probabilistics. Due to lack of space, we cannot rehearse them fully here but refer in particular to [40, 29, 32]. For a (consistent) set of probabilistic conditionals Δ, the principle of maximum entropy selects the unique probability distribution $ME(\Delta)$ with maximum entropy, and if prior information P is given, then the principle of minimum cross-entropy selects (under mild consistency conditions) a unique probability distribution $P *_{ME} \Delta$ that is a model of Δ and has minimal information distance to P, thus realizing probabilistic belief revision. We refer to both principles as the *ME-principles.* The crucial equation for understanding and analyzing ME-revision is given by

$$P *_{ME} \Delta(\omega) = \alpha_0 P(\omega) \prod_{\substack{1 \leqslant i \leqslant n \\ \omega \models A_i B_i}} \alpha_i^{1-x_i} \prod_{\substack{1 \leqslant i \leqslant n \\ \omega \models A_i \overline{B_i}}} \alpha_i^{-x_i}, \tag{8}$$

with the α_i's being exponentials of the Lagrange multipliers, one for each conditional in Δ, and have to be chosen properly to ensure that $P *_{ME} \Delta$ satisfies all conditionals in Δ with the associated probabilities. α_0 is simply a normalizing factor. For a complete axiomatization of the principle of minimum cross-entropy within the scope of probabilistic revision by conditional-logical postulates, see [29]. If P_u is a suitable uniform distribution, both ME-principles are related via $ME(\Delta) = P_u *_{ME} \Delta$. This means that ME is an inductive reasoning mechanism derived from a belief revision operator in the sense of (7), and $*_{ME}$ realises inductive reasoning from general background beliefs P in the sense of (5). Let us further note that ME-revision also satisfies (Coherence) [52]. Hence the ME-methodology is quite a perfect example to illustrate all concepts and relationships presented in this paper in a probabilistic framework.

4.2 Ordinal c-revision

Transferring the basic ideas underlying the ME-principles to the framework of ranking functions brings us to *c-revisions* and *c-representations* [32].

A(n *ordinal*) *c-revision operator* $*_c$ returns for each ranking function κ and each consistent set Δ of conditionals a ranking function $\kappa *_c \Delta$ that satisfies the principle of conditional preservation, as specified in [30, 32].

Again, by applying the c-revision approach to the uniform prior, i.e., the ranking function κ_u with $\kappa_u(\omega) = 0$ for all $\omega \in \Omega$, we obtain quite easily very well-behaved inductive inference operations on default (or conditional) bases called *c-representations.*

Formally, the c-revision methodology[12] provides approaches to revision of ranking functions by sets of conditionals, and inductive reasoning from conditional belief bases (also by taking background beliefs into account) according to (7) and (5) via the following schemata:

$$\kappa *_c \Delta(\omega) = \kappa_0 + \kappa(\omega) + \sum_{\substack{1\leqslant i\leqslant n\\ \omega\models A_i\overline{B_i}}} \kappa_i^- \tag{9}$$

such that

$$\kappa_i^- > \min_{\omega\models A_iB_i}(\kappa(\omega) + \sum_{\substack{j\neq i\\ \omega\models A_j\overline{B}_j}} \kappa_j^-) - \min_{\omega\models A_i\overline{B}_i}(\kappa(\omega) + \sum_{\substack{j\neq i\\ \omega\models A_j\overline{B}_j}} \kappa_j^-) \tag{10}$$

for revision and inductive reasoning with background beliefs, and

$$\kappa_\Delta(\omega) = \sum_{\substack{1\leqslant i\leqslant n\\ \omega\models A_i\overline{B_i}}} \kappa_i^- \tag{11}$$

such that

$$\kappa_i^- > \min_{\omega\models A_iB_i}(\sum_{\substack{j\neq i\\ \omega\models A_j\overline{B}_j}} \kappa_j^-) - \min_{\omega\models A_i\overline{B}_i}(\sum_{\substack{j\neq i\\ \omega\models A_j\overline{B}_j}} \kappa_j^-) \tag{12}$$

for inductive reasoning *ind*$(\Delta) = \kappa_\Delta$. Because the principle of conditional preservation constitutes the main building principle for both the ME-principles and the c-revision methodology, c-representations resp. c-revisions can be considered as suitable translations of the ME-principles to the framework of ranking functions (see also [20, 5]). Note that c-revisions also satisfy Coherence, when considering the whole family of c-revisions of a specific revision problem; for technical details, please see [36]. The paper [36] also elaborates on relations to other properties of iterated revision, in particular to the Darwiche-Pearl postulates [8].

For representing an epistemic state or revising it, any c-representation resp. c-revision of a conditional belief base may be chosen, all of them share the same good properties, see, e.g., [32, 2]. For practical applications, c-representations resp. c-revisions with pareto-minimal parameters κ_i^- are usually chosen, which however are not uniquely determined in general. C-representations generalize the approach of system Z^* [20] which is defined only for so-called minimal-core knowledge bases, without relying on a probabilistic foundation. For minimal-core knowledge bases, minimal c-representations are unique and coincide with system Z^*.

[12] We consider a simplified version here which is sufficient for the purposes of this paper.

This framework of c-revisions resp. inductive c-representations usually takes plausible beliefs in the form of conditionals (which also cover plausible facts by identifying $(A|\top)$ with a plausible fact A) but can also deal with certain information by assigning ∞ to all falsifying worlds. We illustrate this for the case of c-revising κ with a certain fact which we denote by A^∞. According to (9), we obtain

$$\kappa *_c A^\infty(\omega) = \kappa_0 + \kappa(\omega) + \begin{cases} 0 & \text{if } \omega \models A \\ \infty & \text{if } \omega \not\models A \end{cases}$$

with $\kappa_0 = -\min\{\min_{\omega \models A} \kappa(\omega), \infty\} = -\min\{\kappa(A), \infty\} = -\kappa(A)$, hence

$$\kappa *_c A^\infty(\omega) = \begin{cases} \kappa|A(\omega) & \text{if } \omega \models A \\ \infty & \text{if } \omega \not\models A. \end{cases} \tag{13}$$

Therefore, $\kappa *_c A^\infty$ can be considered as an extension of Spohn's conditioning of ranking functions [53]. If a certain fact A^∞ is part of a belief base, this can be handled similarly by the general approach via (9) and (10) by assigning ∞ to all models of $\overline{A}$. Note that this might influence the minima in (10) and lead to a different revision result, please see also Section 5.

The following example illustrates inductive reasoning and all three change scenarios from Section 3.4 – update, conservative revision, and learning – via the c-change methodology in the framework of ranking functions and qualitative conditionals.

Example 1. *Suppose we have the propositional atoms* f - flying*,* b - birds*,* p - penguins*,* w - winged *animals,* k - kiwis*,* d - doves*. Let the set* Δ *consist of the following conditionals:*

Δ :	r_1 :	$(f\|b)$	birds fly
	r_2 :	$(b\|p)$	penguins are birds
	r_3 :	$(\overline{f}\|p)$	penguins do not fly
	r_4 :	$(w\|b)$	birds have wings
	r_5 :	$(b\|k)$	kiwis are birds
	r_6 :	$(b\|d)$	doves are birds

Moreover, we assume strict knowledge, i.e., absolute certainty of the fact that penguins, kiwis, and doves are pairwise exclusive*, which amounts to considering only those worlds as possible that make at most one of* $\{p, k, d\}$ *true; all other worlds have infinite rank.*

As initial epistemic state, we represent inductively Δ *via a c-representation (11) obtaining* $\kappa_\Delta = \kappa_u *_c \Delta$ *as current epistemic state (cf. Figure 1). The calculation of the parameters* κ_i^- *according to (12) is straightforward: For each* r_i *with* $i \in$

ω	$\kappa_\Delta(\omega)$	ω	$\kappa_\Delta(\omega)$	ω	$\kappa_\Delta(\omega)$	ω	$\kappa_\Delta(\omega)$
$p\overline{k}\,\overline{d}bfw$	2	$\overline{p}k\overline{d}bfw$	0	$\overline{p}\,\overline{k}dbfw$	0	$\overline{p}\,\overline{k}\,\overline{d}bfw$	0
$p\overline{k}\,\overline{d}bf\overline{w}$	3	$\overline{p}k\overline{d}bf\overline{w}$	1	$\overline{p}\,\overline{k}dbf\overline{w}$	1	$\overline{p}\,\overline{k}\,\overline{d}bf\overline{w}$	1
$p\overline{k}\,\overline{d}b\overline{f}w$	1	$\overline{p}k\overline{d}b\overline{f}w$	1	$\overline{p}\,\overline{k}db\overline{f}w$	1	$\overline{p}\,\overline{k}\,\overline{d}b\overline{f}w$	1
$p\overline{k}\,\overline{d}b\overline{f}\,\overline{w}$	2	$\overline{p}k\overline{d}b\overline{f}\,\overline{w}$	2	$\overline{p}\,\overline{k}db\overline{f}\,\overline{w}$	2	$\overline{p}\,\overline{k}\,\overline{d}b\overline{f}\,\overline{w}$	2
$p\overline{k}\,\overline{d}\,\overline{b}fw$	4	$\overline{p}k\overline{d}\,\overline{b}fw$	1	$\overline{p}\,\overline{k}d\overline{b}fw$	1	$\overline{p}\,\overline{k}\,\overline{d}\,\overline{b}fw$	0
$p\overline{k}\,\overline{d}\,\overline{b}f\overline{w}$	4	$\overline{p}k\overline{d}\,\overline{b}f\overline{w}$	1	$\overline{p}\,\overline{k}d\overline{b}f\overline{w}$	1	$\overline{p}\,\overline{k}\,\overline{d}\,\overline{b}f\overline{w}$	0
$p\overline{k}\,\overline{d}\,\overline{b}\,\overline{f}w$	2	$\overline{p}k\overline{d}\,\overline{b}\,\overline{f}w$	1	$\overline{p}\,\overline{k}d\overline{b}\,\overline{f}w$	1	$\overline{p}\,\overline{k}\,\overline{d}\,\overline{b}\,\overline{f}w$	0
$p\overline{k}\,\overline{d}\,\overline{b}\,\overline{f}\,\overline{w}$	2	$\overline{p}k\overline{d}\,\overline{b}\,\overline{f}\,\overline{w}$	1	$\overline{p}\,\overline{k}d\overline{b}\,\overline{f}\,\overline{w}$	1	$\overline{p}\,\overline{k}\,\overline{d}\,\overline{b}\,\overline{f}\,\overline{w}$	0

Figure 1: Epistemic state κ_Δ as result of inductive reasoning from Δ in Example 1

$\{1,4,5,6\}$, *and for each of the two minima occurring in (12), respectively, we can choose worlds that do not falsify any (other) conditional from* Δ*; e.g., for* r_1*, choose* $\overline{p}\,\overline{k}dbfw$ *for the first minimum over the models of* bf*, and* $\overline{p}\,\overline{k}\,\overline{d}b\overline{f}w$ *for the second minimum over the models of* $b\overline{f}$*. So for each of these* κ_i^-*, both minima are evaluated to* 0*, and we have* $\kappa_i^- > 0$*. We choose all parameters minimally, so we obtain*

$$\kappa_i^- = 1 \textit{ for } i \in \{1,4,5,6\}.$$

The calculation of κ_2^-, κ_3^- *is a bit more complicated. First note that due to* p, k, d *being exclusive, only the models in the leftmost column of Figure 1, i.e., the penguin worlds, are relevant for this calculation. For* κ_2^-*, we compute*

$$\begin{aligned} \kappa_2^- &> \min\{\kappa_3^-, \kappa_3^- + \kappa_4^-, \kappa_1^-, \kappa_1^- + \kappa_4^-\} - \min\{\kappa_3^-, 0\} \\ &= \min\{\kappa_3^-, \kappa_1^-\} - 0, \end{aligned}$$

and because we set $\kappa_1^- = 1$*, we have* $\kappa_2^- > \min\{\kappa_3^-, 1\}$*. Similarly, for* κ_3^- *we obtain* $\kappa_3^- > \min\{\kappa_2^-, 1\}$*. From both inequalities, we can conclude that each of* κ_2^-, κ_3^- *must be at least* 1*, and choosing them minimally yields*

$$\kappa_2^- = \kappa_3^- = 2.$$

Using these parameters defines κ_Δ *according to (11).*

It can be checked easily that κ_Δ *yields the conditional beliefs that penguin-birds do not fly (*$\kappa_\Delta \models (\overline{f}|pb)$ *because of* $\kappa_\Delta(pb\overline{f}) = 1 < 2 = \kappa_\Delta(pbf)$*), and that also penguins are expected to have wings (*$\kappa_\Delta \models (w|p)$ *because of* $\kappa_\Delta(pw) = 1 < 2 = \kappa_\Delta(p\overline{w})$*). So, c-representations do not suffer from the so-called* drowning problem*, particularly a problem of system Z [21]. Moreover, also both kiwis and doves inherit the property* *of having wings from their superclass* birds*, due to* $\kappa_\Delta(kw) = 0 < 1 = \kappa_\Delta(k\overline{w})$ *and* $\kappa_\Delta(dw) = 0 < 1 = \kappa_\Delta(d\overline{w})$*.*

Suppose now that the agent gets to know that this is false for kiwis - kiwis do not *possess wings - and we want the agent to adopt this new information which has escaped her beliefs before. So, the agent wants to change her beliefs about the world, but the world itself has not changed. Hence conservative revision is the proper belief change operation and amounts to computing a new inductive representation for the set* $\Delta' = \{(f|b), (b|p), (\overline{f}|p), (w|b), (b|k),\ (b|d)\} \cup \{(\overline{w}|k)\}$*, i.e. a c-revision of* κ_u *by* Δ' *has to be computed:* $\kappa_{\Delta'} = \kappa_u *_c \Delta'$*. Note that the new information* $(\overline{w}|k)$ *is not consistent with the prior epistemic state* κ_Δ *but with the context information* Δ *which is refined by* $(\overline{w}|k)$*.*

Alternatively, let us suppose that the agent (with current epistemic state κ_Δ*) learned from the news, that, due to some mysterious illness that has occurred recently among doves, the wings of newborn doves are nearly completely mutilated. She wants to adopt her beliefs to the new information* $(\overline{w}|d)$*. Obviously, the proper change operation in this case is an update operation as the world under consideration has changed by some event (the occurrence of the mysterious illness).*

The updated epistemic state $\kappa^* = \kappa_\Delta *_c \{(\overline{w}|d)\}$ *is a c-revision of* κ_Δ *by* $\{(\overline{w}|d)\}$ *and can be obtained from* κ_Δ *via (11) by setting* $\kappa^*(\omega) = \kappa_\Delta(\omega) + 2$ *for any* ω *with* $\omega \models dw$ *and setting* $\kappa^*(\omega) = \kappa_\Delta(\omega)$ *otherwise.*

While the conservatively revised state $\kappa_{\Delta'}$*, by construction, still represents the six conditionals that have been known before (and, of course, the new conditional), it can be verified easily that the updated state* κ^* *only represents the five conditionals* $(f|b)$*,* $(b|p)$*,* $(\overline{f}|p)$*, and* $(w|b)$*,* $(b|k)$*, but it no longer satisfies* $(b|d)$ *because* $\kappa^*(bd) = \kappa^*(\overline{b}d) = 1$ *- since* birds *and* wings *have been plausibly related by the conditional* $(w|b)$*, the property of not having wings casts (reasonably) doubt on doves being birds. Moreover, the agent is now also uncertain about the ability of doves to fly, as also* $\kappa^*(fd) = \kappa^*(\overline{f}d) = 1$*. This illustrates that explicitly stated prior beliefs are kept under conservative revision, but might be given up under update. It can be easily checked that these effects would have been the same when c-revising* $\kappa_{\Delta'}$ *instead of* κ_Δ*, since the agent's beliefs on kiwis and doves do not interfere. Moreover, note that if the agent became aware of having missed to represent the conditional belief* $(\overline{w}|k)$ *in the new world after the occurrence of the mysterious illness, still* $\kappa_{\Delta'} *_c (w|b)$ *would be the most adequate result, because here background beliefs are affected, the agent has* learned $(\overline{w}|k)$ *by conservatively revising* κ_Δ*.*

5 Focusing and Conditioning

Focusing means applying generic knowledge to a reference class appropriate to describe the context of interest (cf. [12]). As this reference class is assumed to be specified by factual information and indicates a shift in context (to that reference class), focusing should be performed by updating the current epistemic state to *factual* information which is certain. It can easily be shown that both for ME-change and for ordinal c-change, updating with such information results in conditioning the prior epistemic state (see Propositions 2 and 3 below), and indeed, conditioning is usually considered to be the proper operation for focusing. We share this view in this paper, i.e., in our framework, focusing is done via conditioning resp. updating with certain facts.

However, conditioning has been used for revision, too [17, 12]. So revision and focusing are often supposed to coincide though they differ conceptually: revision is not only *applying knowledge*, but means incorporating a new constraint so as to *change knowledge.* Due to this conceptual mismatch, paradoxes have been observed. Gärdenfors investigated *imaging* as another proper probabilistic change operation [17]. Dubois and Prade argued that the assumption of having a uniquely determined probability distribution to represent the available knowledge at best is responsible for that flaw, and they recommend to use upper and lower probabilities to permit a proper distinction (cf. [12]).

However, we will show that in our framework, it is easily possible to treat revision as different from focusing without giving up the assumption of having a single, distinguished epistemic state as a result of revision and a base for inferences. Making use of ME-revision for probabilities, and c-revisions for ranking functions, respectively, it is indeed possible to realize this conceptual difference appropriately. To make this clear, we have to consider belief changes induced by some certain information A, that is, we learn proposition A with certainty. For probabilities, this means that we assign probability 1 to A, while for ranking functions, we assign rank ∞ to $\overline{A}$, see (13) (which implies particularly that A has rank 0). The following two propositions reveal the difference between revision by a certain information A, as realized according to (8) resp. (9) and (10), and focusing to A by conditioning; the proofs are straightforward but tedious, using the mentioned equations.

Proposition 2. *Let P be a distribution, $\Delta \subseteq (\mathcal{L} \mid \mathcal{L})^{prob}$ a (P-consistent[13]) set of probabilistic conditionals, and suppose $A[1]$ to be a certain probabilistic fact.*

(i) Focussing on A, i.e., updating P with $A[1]$ via ME-revision is done by conditioning and yields $P *_{ME} \{A[1]\} = P(\cdot|A)$; in particular, $(P *_{ME} \Delta) *_{ME} A[1] =$

[13] Δ is P-consistent if there is a distribution Q with $Q \models \Delta$ and $Q(\omega) = 0$ whenever $P(\omega) = 0$.

$(P *_{ME} \Delta)(\cdot|A)$.

(ii) Conservatively revising $P *_{ME} \Delta$ with $A[1]$ yields $P *_{ME} (\Delta \cup \{\mathrm{A}[1]\}) = P(\cdot|\mathrm{A}) *_{ME} \Delta$.

An analogical statement holds for focussing and conservative revision for ranking functions.

Proposition 3. *Let κ be a ranking function, $\Delta \subseteq (\mathcal{L} \mid \mathcal{L})$ a (κ-consistent[14]) set of conditionals, and suppose A to be a certain fact.*

(i) Focussing κ on A, i.e., updating κ with the certain fact A via c-revision is done by conditioning and yields $\kappa *_c A^\infty(\omega) = \kappa|A(\omega)$ for models ω of A; in particular, $(\kappa *_c \Delta) *_c A^\infty = (\kappa *_c \Delta)|A$ on the models of A.

(ii) Conservatively revising $\kappa *_c \Delta$ with the certain fact A yields $\kappa *_c (\Delta \cup \{\mathrm{A}^\infty\}) = (\kappa *_c \mathrm{A}^\infty) *_c \Delta$ (which coincides with $(\kappa|A) *_c \Delta$ on the models of A) if the same parameters κ_i^- are chosen for both c-revisions.

We present a ranking function adaptation of a probabilistic example from [33].

Example 4. *A psychologist has been working with addicted people for a couple of years. His experiences concerning the propositions*

a : addicted to alcohol
d : addicted to drugs
y : being young

can be summarized by the ranking function κ as given in (14), serving as the initial epistemic state of the psychologist here.

ω	$\kappa(\omega)$	$\kappa_1^*(\omega)$	$\kappa *_c y^\infty(\omega)$	$\kappa_2^*(\omega)$	$\kappa_3^*(\omega)$
ady	4	4	3	2	2
$ad\overline{y}$	4	4	∞	∞	∞
$a\overline{d}y$	3	3	2	1	1
$a\overline{d}\overline{y}$	0	0	∞	∞	∞
$\overline{a}dy$	1	6	0	3	4
$\overline{a}d\overline{y}$	4	9	∞	∞	∞
$\overline{a}\overline{d}y$	2	2	1	0	0
$\overline{a}\overline{d}\overline{y}$	3	3	∞	∞	∞

(14)

[14] Δ is κ-consistent if there is a ranking function κ' with $\kappa' \models \Delta$ and $\kappa'(\omega) = \infty$ whenever $\kappa(\omega) = \infty$.

The following conditionals can be entailed from κ:

$$(\overline{d}|a), (\overline{a}|d), (\overline{a}|y), (a|\overline{y}), (d|y), (\overline{d}|\overline{y}).$$

These conditionals express that when focussing on drugs and/or alcohol, usually, people are not addicted to both, and that young people are usually addicted to drugs but not to alcohol, while for older people, it is the other way round.

Now the psychologist is going to change his job: He will be working in a clinic where addictions to both alcohol and drugs are not uncommon, more precisely, people being addicted to drugs tend to also being addicted to alcohol. So, when starting to work in the new environment, the psychologist c-revises his initial epistemic state κ by $\Delta_1 = \{(a|d)\}$, yielding $\kappa_1^* = \kappa *_c \Delta_1$ (with minimal parameter).

After having spent a couple of days in the new clinic, the psychologist realized that this clinic is for young people only, i.e., he has overlooked the certain fact y that only young people are present in his new working context. He conservatively revises κ_1^* by y^∞, yielding $\kappa_2^* = \kappa *_c \Delta_2$ with $\Delta_2 = \{(a|d), y^\infty\}$. Hence, according to Proposition 3, he obtains $\kappa_2^* = (\kappa *_c y^\infty) *_c \Delta_1$.

Note that κ_2^* is different from the ranking function that the psychologist would have obtained by focusing *his beliefs represented by κ_1^* on a young person; in that case, he would have updated κ_1^* by y^∞, yielding $\kappa_3^* = \kappa_1^* *_c y^\infty$ (which coincides with $\kappa_1^*|y$ on the models of y). Clearly, κ_2^* and κ_3^* are different (though the differences are only small).*

Propositions 2 and 3, as well as Example 4 show that, in a (generalized) framework of inductive reasoning including belief revision, a proper distinction between focusing and revision is not only possible, but even mandatory. This difference is akin to the one between "conditioning" and "constraining" elaborated by Voorbraak [59] for classes of probability functions (for a criticism of conditioning *sets* of probability measures, cf. [22]). It is interesting to note that this difference between focusing and (genuine) revision that is possible in our framework also allows for making a basic difference between revising by factual evidence vs. revising by more generic pieces of information. Example 4 nicely illustrates this difference when incorporating the information "all people are young" (κ_2^*) vs. "a specific person is young" (κ_3^*).

However, a proper distinction between focusing and (general) revision is hard to make in most frameworks. For instance, in probabilistics, conditioning is often perceived as the main operation for adjusting to new information (revision) in Bayesian approaches, and hence coincides with focusing, which may lead to unintuitive results. We illustrate this by discussing an example from [13] below. In that paper, a Baysian analysis was performed with conditioning as the major (and only) probabilistic change operation, and it was argued that modelling ignorance via uniform distributions can lead to counter-intuitive results. From that example, we extract the main points for modelling it in the frameworks considered here to highlight the erroneous effects that conditioning may have when applied too plainly.

Example 5 (adapted from [13]). *Peter, Paul, and Mary are killers one of whom has been hired by Big Boss to commit a murder. Police Inspector Smith knows that Big Boss has first tossed a coin to decide whether it should be a man (Peter or Paul), or a lady (Mary), but he does not know about the outcome of the tossing. So, initially, the explicit beliefs of Smith are given by* $\Delta_1 = \{(Peter \vee Paul)[0.5], Mary[0.5]\}$*, and his initial epistemic state can be calculated via the principle of maximum entropy:* $P_1 = ME(\Delta_1)$*. It is straightforward to see that* $P_1(Mary) = 0.5, P_1(Paul) = P_1(Peter) = 0.25$.

Now Smith comes to know that Peter has been arrested right before the murder, so he could not have committed the crime. This piece of information can be encoded by $R_2 = \{\neg Peter[1]\}$*. When incorporating* Δ_2 *by conditioning (which corresponds to the usual Bayesian update), the new epistemic state would be* $P_2 = P_1(\cdot|\neg Peter)$*, and hence the new beliefs concerning Paul and Mary would be* $P_2(Mary) = \frac{2}{3}$*, and* $P_2(Paul) = \frac{1}{3}$*. This seems to be unintuitive, as it gives undue precedence to Mary.*

However, this (admittedly) unintuitive result is neither an argument against uniform priors, nor against maximum entropy or probability theory in general, but caused by the confusion between focusing and revision. Incorporating Δ_2 *by conditioning would be seen as focusing in our framework, which seems unappropriate because we do not focus on (the reference class of) Peter not being the culprit, but should understand* Δ_2 *as an additional piece of information on the same level as* Δ_1 *because both refer to exactly the same context, namely, the murder, and deliberating on possible delinquents. This is exactly what conservative revision does. So, the correct change operation here would be conservative revision instead of conditioning, which results in computing* $P_3 = ME(\Delta_1 \cup \Delta_2)$*. Now, in fact, we obtain* $P_3(Mary) = P_3(Paul) = 0.5$*, as expected.*

Therefore, our approach shows that the problem addressed in the paper [13] is not with uniform priors but with reducing probabilistic belief revision to Bayesian conditioning, and that the problem can be solved within probabilistic reasoning if the context of information (not necessarily only temporal meta-information) is properly taken into a account and if a richer epistemic framework of revision is used where inductive reasoning is an integral part of.

6 Conclusion

The central claim of this paper was that inductive reasoning can be considered as a special case of epistemic belief revision. We should study the latter in order to deliver an account of the former. This allowed us to present a general, abstract framework based on epistemic states and conditionals and to show how a coherent and homogeneous approach to inductive reasoning is possible realizing different forms of inductive reasoning via conservative revision, updating, and focusing, where all change operations are realized via the same revision operator, but applied in different ways. In particular, we could describe inductive reasoning from conditional belief bases in a rich epistemic framework that takes epistemic states and conditionals as basic encodings of information. We could thus capture how to inductively reason from background beliefs in the form of belief bases or epistemic states. We illustrated our ideas both for ordinal and probabilistic environments and finally showed how commonly known paradoxes can be avoided in our framework.

We presented two semantical frameworks that allow for implementing these ideas as a proof of concept: the principles of optimum entropy in probabilistics, and c-representations/c-revisions for ranking functions. For future work, it would be interesting to see what other semantical frameworks can be used resp. extended to realize the cornerstones of our framework as described in Section 3, in particular in Section 3.4. Possibility theory [11] seems to be a most promising candidate here because it is similar to ranking functions, at least in its product-based form [4]. First steps towards elaborating this have been taken in [31, 30] but more needs to be done to fill out the complete framework. Moreover, we mentioned Gärdenfors' *imaging* [17] as another probabilistic change operation that has interesting applications. However, it is still not clear how imaging can be integrated in our approach, this is also part of our ongoing work.

References

[1] C. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50(2):510–530, 1985.

[2] C. Beierle, C. Eichhorn, G. Kern-Isberner, and S. Kutsch. Properties and interrelationships of skeptical, weakly skeptical, and credulous inference induced by classes of minimal models. *Artif. Intell.*, 297:103489, 2021.

[3] C. Beierle, G. Kern-Isberner, K. Sauerwald, T. Bock, and M. Ragni. Towards a general framework for kinds of forgetting in common-sense belief management. *KI – Zeitschrift für Künstliche Intelligenz: Special Issue on Intentional Forgetting*, 33(1):57–68, 2019.

[4] S. Benferhat, D. Dubois, and H. Prade. Representing default rules in possibilistic logic. In *Proceedings 3th International Conference on Principles of Knowledge Representation and Reasoning KR'92*, pages 673–684, 1992.

[5] R. Bourne and S. Parsons. Maximum entropy and variable strength defaults. In *Proceedings Sixteenth International Joint Conference on Artificial Intelligence, IJCAI'99*, pages 50–55, 1999.

[6] R. Carnap. A basic system of inductive logic. In R. Carnap and R. C. Jeffrey, editors, *Studies in Inductive Logic and Probability*, volume I, pages 33–165. University of California Press, Berkeley, 1971/1980.

[7] R. Carnap. A basic system of inductive logic. In R. C. Jeffrey, editor, *Studies in Inductive Logic and Probability*, volume II, pages 7–155. University of California Press, Berkeley, 1971/1980.

[8] A. Darwiche and J. Pearl. On the logic of iterated belief revision. *Artificial Intelligence*, 89:1–29, 1997.

[9] B. de Finetti. La prévision, ses lois logiques et ses sources subjectives. In *Ann. Inst. H. Poincaré*, volume 7. 1937. English translation in *Studies in Subjective Probability*, ed. H. Kyburg and H.E. Smokler, 1964, 93-158. New York: Wiley.

[10] D. Dubois. Three scenarios for the revision of epistemic states. *Journal of Logic and Computation*, 18(5):721–738, 2008.

[11] D. Dubois, J. Lang, and H. Prade. Possibilistic logic. In D. Gabbay, C. Hogger, and J. Robinson, editors, *Handbook of Logic in Artificial Intelligence and Logic Programming*, volume 3. Oxford University Press, 1994.

[12] D. Dubois and H. Prade. Non-standard theories of uncertainty in plausible reasoning. In G. Brewka, editor, *Principles of Knowledge Representation.* CSLI Publications, 1996.

[13] D. Dubois, H. Prade, and P. Smets. Representing partial ignorance. *IEEE Trans. Syst. Man Cybern. Part A*, 26(3):361–377, 1996.

[14] P. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77:321–357, 1995.

[15] A. Egan and M. G. Titelbaum. Self-locating beliefs. In E. N. Zalta and U. Nodelman, editors, *The Stanford Encyclopedia of Philosophy.* Winter 2022 edition, 2022.

[16] H. Field. The a prioricity of logic. *Proceedings of the Aristotelian Society*, 96(1):359–379, 1996.

[17] P. Gärdenfors. *Knowledge in Flux: Modeling the Dynamics of Epistemic States.* MIT Press, Cambridge, Mass., 1988.

[18] P. Gärdenfors. Belief revision and nonmonotonic logic: Two sides of the same coin? In *Proceedings European Conference on Artificial Intelligence, ECAI'92*, pages 768–773. Pitman Publishing, 1992.

[19] M. Gelfond and N. Leone. Logic programming and knowledge representation – the A-prolog perspective. *Artificial Intelligence*, 138:3–38, 2002.

[20] M. Goldszmidt, P. Morris, and J. Pearl. A maximum entropy approach to nonmonotonic reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(3):220–232, 1993.

[21] M. Goldszmidt and J. Pearl. Qualitative probabilities for default reasoning, belief revision, and causal modeling. *Artificial Intelligence*, 84:57–112, 1996.

[22] A. Grove and J. Halpern. Updating sets of probabilities. In *Proceedings Fourteenth Conference on Uncertainty in AI*, pages 173–182, 1998.

[23] B. Hedden. *Reasons Without Persons: Rationality, Identity, and Time.* Oxford University Press, Oxford, 2015.

[24] M. Hild and W. Spohn. The measurement of ranks and the laws of iterated contraction. *Artificial Intelligence*, 172(10):1195–1218, 2008.

[25] J. J. Ichikawa and M. Steup. The analysis of knowledge. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy.* Summer 2018 edition, 2018.

[26] E. Jaynes. *Papers on Probability, Statistics and Statistical Physics.* D. Reidel Publishing Company, Dordrecht, Holland, 1983.

[27] R. Jeffrey. *The logic of decision.* University of Chicago Press, Chicago, IL, 1983.

[28] H. Katsuno and A. Mendelzon. On the difference between updating a knowledge base and revising it. In *Proceedings Second International Conference on Principles of Knowledge Representation and Reasoning, KR'91*, pages 387–394, San Mateo, Ca., 1991. Morgan Kaufmann.

[29] G. Kern-Isberner. Characterizing the principle of minimum cross-entropy within a conditional-logical framework. *Artificial Intelligence*, 98:169–208, 1998.

[30] G. Kern-Isberner. *Conditionals in nonmonotonic reasoning and belief revision.* Springer, Lecture Notes in Artificial Intelligence LNAI 2087, 2001.

[31] G. Kern-Isberner. Representing and learning conditional information in possibility theory. In *Proceedings 7th Fuzzy Days, Dortmund, Germany*, pages 194–217. Springer LNCS 2206, 2001.

[32] G. Kern-Isberner. A thorough axiomatization of a principle of conditional preservation in belief revision. *Annals of Mathematics and Artificial Intelligence*, 40(1-2):127–164, 2004.

[33] G. Kern-Isberner. Linking iterated belief change operations to nonmonotonic reasoning. In G. Brewka and J. Lang, editors, *Proceedings 11th International Conference on*

Knowledge Representation and Reasoning, KR'2008, pages 166–176, Menlo Park, CA, 2008. AAAI Press.

[34] G. Kern-Isberner, T. Bock, C. Beierle, and K. Sauerwald. Axiomatic evaluation of epistemic forgetting operators. In R. Bartak and K. Brawner, editors, *Proceedings of the 32nd International FLAIRS Conference, FLAIRS-32*, pages 470–475, Palo Alto, CA, 2019. AAAI Press.

[35] G. Kern-Isberner, T. Bock, K. Sauerwald, and C. Beierle. Belief change properties of forgetting operations over ranking functions. In A. Nayak and A. Sharma, editors, *Proceedings of the 16th Pacific Rim International Conference on Artificial Intelligence PRICAI 2019*, number 11670 in Lecture Notes in Artificial Intelligence, pages 459–472. Springer, 2019.

[36] G. Kern-Isberner and D. Huvermann. What kind of independence do we need for multiple iterated belief change? *J. Applied Logic*, 22:91–119, 2017.

[37] M. Kopec and M. G. Titelbaum. The uniqueness thesis. *Philosophy Compass*, 11(4):189–200, 2016.

[38] J. Landes, S. Rafiee Rad, and J. Williamson. Determining maximal entropy functions for objective bayesian inductive logic. *Journal of Philosophical Logic*, 52(2):555–608, 2023.

[39] D. Lewis. *Counterfactuals.* Harvard University Press, Cambridge, Mass., 1973.

[40] J. Paris. *The uncertain reasoner's companion – A mathematical perspective.* Cambridge University Press, 1994.

[41] J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufmann, San Mateo, Ca., 1988.

[42] J. L. Pollock. *Nomic Probability and the Foundations of Induction.* Oxford University Press, Oxford, 1990.

[43] J. L. Pollock. *Cognitive Carpentry: A Blueprint for How to Build a Person.* MIT Press, Cambridge, MA, 1995.

[44] J. L. Pollock and A. S. Gillies. Belief revision and epistemology. *Synthese*, 122:69–92, 2000.

[45] H. Putnam. The analytic and the synthetic. In H. Feigl and G. Maxwell, editors, *Scientific Explanation, Space, and Time*, volume 3, pages 358–397. University of Minnesota Press, Minneapolis, 1962.

[46] H. Putnam. The meaning of 'meaning'. In K. Gunderson, editor, *Language, Mind, and Knowledge*, volume 7, pages 131–193. University of Minnesota Press, Minneapolis, 1975.

[47] E. Raidl. Definable conditionals. *Topoi*, 40(1):87–105, 2021.

[48] E. Raidl and H. Rott. Towards a logic for 'because'. *Philosophical Studies*, 2023.

[49] R. Reiter. A logic for default reasoning. *Artificial Intelligence*, 13:81–132, 1980.

[50] H. Rott. Difference-making conditionals and the relevant Ramsey test. *Review of Symbolic Logic*, 15(1):133–164, 2022.

[51] M. Sezgin, G. Kern-Isberner, and H. Rott. Inductive reasoning with difference-making

conditionals. In M. Martinez and I. Varcinczak, editors, *Proceedings of the 18th International Workshop on Non-Monotonic Reasoning, NMR 2020*, 2020.

[52] J. Shore and R. Johnson. Properties of cross-entropy minimization. *IEEE Transactions on Information Theory*, IT-27:472–482, 1981.

[53] W. Spohn. Ordinal conditional functions: a dynamic theory of epistemic states. In W. Harper and B. Skyrms, editors, *Causation in Decision, Belief Change, and Statistics, II*, pages 105–134. Kluwer Academic Publishers, 1988.

[54] W. Spohn. A brief comparison of Pollock's defeasible reasoning and ranking functions. *Synthese*, 131:39–56, 2002.

[55] W. Spohn. *The Laws of Belief: Ranking Theory and Its Philosophical Applications.* Oxford University Press, 2012.

[56] W. Spohn. Enumerative induction. In C. Beierle, G. Brewka, and M. Thimm, editors, *Computational Models of Rationality: Essays Dedicated to Gabriele Kern-Isberner on the Occasion of Her 60th Birthday*, pages 96–114. College Publications, London, 2016.

[57] W. Spohn. The epistemology and auto-epistemology of temporal self-location and forgetfulness. *Ergo*, 4(13):359–148, 2017.

[58] W. Spohn. Defeasible normative reasoning. *Synthese*, 197:1391–1428, 2020.

[59] F. Voorbraak. Probabilistic belief expansion and conditioning. Technical Report LP-96-07, Institute for Logic, Language and Computation, University of Amsterdam, 1996.

[60] J. Williamson. *In Defence of Objective Bayesianism.* Oxford University Press, Oxford, 2010.

Received August 2023