Essays on Case Based Belief Formation

Inaugural-Dissertation

zur Erlangung des Grades eines Doktors

der Wirtschaftswissenschaften (Dr. rer. pol.)

an der

Fakultät für Wirtschaftswissenschaften

der Universität Bielefeld

vorgelegt von

Jörg Bleile

Bielefeld, 2014

Gedruckt auf alterungsbeständigem Papier nach DIN-ISO 9706.

Erstgutachter:	Prof. Dr. Frank Riedel
Zweitgutachter:	Prof. Christoph Kuzmics, Ph.D.

Contents

1 General introduction **6**

1.1 Beliefs in the classical Savage framework 6

1.2 Belief formation in a case based framework of information representation 7

1.3 Content of the thesis 8

2 Cautious Belief Formation **12**

Abstract 12

2.1 Introduction and motivation 12

2.2 The model 19

2.2.1 Database framework 19

2.2.2 Induced beliefs 19

2.3 Motivating examples 20

2.3.1 Exemplary development of a belief formation process 20

2.3.2 Leading example 23

2.4 Axioms 24

2.4.1 Retained axioms 24

2.4.2 Different versions of the Concatenation Axiom 26

2.5 Representation result of cautious belief formation 37

2.5.1 Representation Theorem with maximal anchored axioms . . . 37

2.5.2 Comparison to related belief representations 41

2.5.3 Remarks on the similarity function 42

2.5.4 Remarks on relationship to statistical methods 43

2.6 Conclusion 44

2.7 Further remarks and considerations 46

2.7.1 Incompatible objective belief formation 46

2.7.2 Relationship to EG's axiom "Concatenation restricted to databases of equal length" 47

2.7.3 Minimal anchored axiomatization 48

2.8 Proof of Theorem 2.1, necessity part 49

2.9 Proof of Theorem 2.1, sufficiency part 52
2.9.1 General definitions for a frequency framework 52
2.9.2 Beliefs induced by frequency vectors 53
2.9.3 Axioms in the frequency framework 54
2.9.4 Useful observations 55
2.9.5 Theorem 2.1, sufficiency part in frequency version 59
2.9.6 Step 1: $C = \{c_1, c_2, c_3\}$, i.e. $m = 3$ 61
2.9.7 Step 2: Set of basic cases with $|C| = m > 3$ many cases 69

3 Limited Attention in Case Based Belief Formation 73
Abstract 73
3.1 Introduction and motivation 73
3.2 The model 79
3.2.1 Database framework 79
3.2.2 Filter 80
3.2.3 Induced belief 81
3.3 Filter definition and properties 82
3.3.1 Basic properties of a filter 82
3.3.2 Main structural properties of a filter 85
3.4 Axioms on level of belief formation 88
3.5 Representation Theorem 91
3.5.1 Example: Similarity satisficing 93
3.6 Related literature with consideration or elimination stage 94
3.6.1 Relationship to multi-criteria/stage Decision Theory 94
3.6.2 Filter as choice correspondence in multistage procedures . . . 95
3.7 Conclusion 103
3.8 Proof of Theorem 3.1, necessity part 104
3.9 Proof of Theorem 3.1, sufficiency part 104
3.9.1 Important observations 104
3.9.2 Translating the database framework to frequencies 105
3.9.3 Proof of sufficiency part of Theorem 3.1 in frequency terms . . 109

4 Belief Formation Based on Categorization 116
Abstract 116
4.1 Introduction and motivation 116
4.2 The model 122
4.2.1 Database framework 122
4.2.2 Induced belief 123

4.3 Motivating example 124
4.4 Natural evolutionary (optimal) categorization structure 126
4.5 Axiomatization of category based belief formation 127
4.5.1 Specific properties of the categorization procedure 127
4.5.2 Induced category based belief 131
4.5.3 Axioms on the level of belief formation 131
4.5.4 Representation Theorem of category based belief formation . . 133
4.6 Axiomatization of prototype based belief formation 136
4.6.1 Specific properties of the categorization procedure 136
4.6.2 Induced prototype based belief 142
4.6.3 Axioms on the level of belief formation 143
4.6.4 Representation Theorem of prototype based belief formation . 145
4.7 Conclusion 148
4.8 Proof of Theorem 4.1 (Category based belief formation) 149
4.8.1 Translating the database framework into frequencies 149
4.8.2 Theorem 4.1, sufficiency part in frequency terms 151
4.8.3 Proof of Theorem 4.1, sufficiency part in frequency terms . . . 151
4.9 Proof of Theorem 4.2 (Prototype based belief formation) 154
4.9.1 Translating the database framework into frequencies 154
4.9.2 Theorem 4.2, sufficiency part in frequency terms 156
4.9.3 Proof of Theorem 4.2, sufficiency part in frequency terms . . . 157

1 General introduction

1.1 Beliefs in the classical Savage framework

Beliefs over uncertain future consequences are important determinants of individual behavior. However, in economic theory the belief formation process is most of time excluded from the analysis and models since uncertainty is described and implemented by assuming variants of the prominent Savage (1954) and Bayesian approach. This underlying idea is to model uncertainty by a grand state space that is sufficiently rich to describe and resolve all possible sources of uncertainties. It means that all information, knowledge and past experiences are encoded and reflected in the definition of the states. In this way the definition and determination of such an all encompassing state space requires the (immense) cognitive ability to implicitly come up with some "perfect" belief and theories about structures and relationships in the future. However, in many situations states are not naturally given, readily available or easy to imagine and generate. Basically, often environments do not offer sufficient (or too complex) information to suggest a "correct" definition of a grand state space.

Another cornerstone of the state based approach is the representation of a belief as an unique probability over the grand state space that is endogenously deduced from observable behavior of the agents. Such a purely subjective belief does not explicitly process and incorporate (past) information into a belief formation in an objective way. More importantly, the approach inherently lacks an explicit description of the formation of the belief and thus does not help an agent to form the belief that eventually would lead to the behavior. Gilboa et al. (2012) discuss this issue extensively and nicely summarize it (on p.8) "... main difficulty with assigning probability to the Grand State Space is that there is no information on which one can base the choice of prior beliefs. Any information that one may obtain, and that may help in the choice of a prior, should be incorporated into the description of the Grand State Space. This means that the prior on this state has to specify beliefs one had before obtaining the information in question. Thus, the information may help one choose posterior beliefs (presumably according to Bayes's rule), but not prior beliefs." Consequently, employing the state based approach in modeling uncertainty induces by definition an impossibility of an explicit belief formation. If the available information is sufficient to define an all encompassing state space, structurally a state space precludes a belief formation process that takes into account (no "remaining") information directly. Thus, in order to allow for an explicit (objective) belief formation process that enables agents to directly incorporate information into a belief formation an alternative representation or description of the available information and uncertainty is necessary. This applies also to environments that are not appropriate to translate and condense its information into a "perfect " state space (and thus prevents per se a deduction of a subjective belief a la Savage). Usually, information can be interpreted as a list or collection of single pieces of past observations or cases, i.e. so called databases. In the following, we also have in mind that a database might represent an agent's memory. In the following we will adopt this "database or case based"-representation of information and replace the state space as an information aggregation. This appears especially appropriate for situations in which the information is not all encompassing as needed for the correct description of a grand state space. The main advantage of the database representation of information is that it enables an explicit belief formation that relies on the contained information.

1.2 Belief formation in a case based framework of information representation

In contrast to Savage's state based approach a belief will and cannot be derived purely subjectively anymore. A belief based on a database needs to explicitly process and incorporate factual knowledge, properties and theoretical considerations provided by the present database. This task appears to be very close to the goal of statistical inference. However, statistical inference mainly deals with asymptotic considerations (i.e. sufficiently rich databases), but the focus of our analysis lies on environments that are mainly characterized by databases that contain insufficiently rich information (i.e. small or medium sized databases). In particular, we are concerned with a belief formation process and its behavioral foundation (axiomatization) (similar as Billot et al. (Econometrica, 2005)). In contrast to statistical experiments that usually deal with identical observations which are considered equally relevant, small or medium sized databases contain limited and heterogenous information that might be differently relevant. Thus, agents might want to take into account not only (a few) identical but also partially relevant observations for their belief formation. In this sense - differently to statistical experiments - relevance or similarity measures become important, when data sets contain limited heterogenous information. The underlying assumption of this idea is the similarity hypothesis, stating that similar situations/actions induce similar outcomes, e.g. in the spirit of Hume (1748): "Reasoning rests on the principle of analogy".

Case-based Decision Theory (Gilboa and Schmeidler (1995, 2001)) deals with such a framework in decision theoretic contexts. Billot et al. (2005) can be interpreted as an adoption of it to belief formation[1]. Their axiomatized belief describes a generalized (subjective) frequentist in which agents assign different similarity weights to information with different degree of relevance. For a new problem, described by a vector of properties x, and given a database or memory D of past observations/cases $c = (x', r')$, their belief P over possible outcomes r is represented as a similarity weighted average of estimates that are induced by the observed cases, i.e.

$$\text{"BGSS- belief":} \qquad P(x, D) = \frac{\sum_{c \in D} s(x, c) f_D(c) P^c}{\sum_{c \in D} s(x, c) f_D(c)}, \tag{1.1}$$

where $s(x, c)$ measure the degree of similarity of the problem under consideration x with the past case c and P^c is the estimate or prediction over outcomes an agent derives from observing case c.

1.3 Content of the thesis

The thesis mainly consists of the Chapters 2-4, where each chapter is based on a self contained article. Each chapter starts with a detailed introduction and motivation of the particular issue and its relevance for the related existing literature. The three chapters deal with axiomatizations of different belief formation process in the vein of BGSS, i.e. (1.1). Their axiomatic belief formation can be seen as a starting point for this thesis.

Chapter 2[2] deals with an issue that arises when dealing with small databases that consist of limited and heterogenous content. As in controlled statistical experiments, the number of observations may serve as a proxy for its informativeness, precision or accuracy. Thus, small databases with few or insufficiently many observation of cases might appear to be relatively imprecise or unreliable and evoke some concerns about precision. Additional observations of the cases might increase its precision and reliability.

[1]Related also to Gilboa and Schmeidler (2003), Gilboa (2009) and Gilboa et al. (2011).

[2]Based on Bleile (2014a) "Cautious Belief Formation", Working Paper No. 507, Center of Mathematical Economics, Bielefeld.

However, not only the precision per se is important for prediction, but also the related perception and sensation of it in form of confidence and cautiousness (or anxiety) in evaluating or predicting based on differently precise information. Ellsberg (1961) (p.657) summarizes the problem as follows: "What is at issue might be called the ambiguity [or imprecision] of this information, a quality depending on the amount, type, reliability, and "unanimity" of information, and giving rise to ones degree of "confidence" in an estimate of relative likelihoods." In fact increasing the number of observations is likely to influence both the forecast made by agent and his confidence in this forecast, i.e. confidence that the observed frequencies reflect the actual data-generating process. The more additional confirming evidence is observed, the less cautious (anxious) and the more confident are the induced estimates. This intuition is in line with approaches in controlled statistical experiments, in which the set of possible priors shrinks as additional data confirm the evaluation, as well as with the usual Bayesian updating method, in which additional information is used to exclude states that have not occurred.

Chapter 2 tackles exactly the issue that additional confirming observations increases precision and reliability of the information, which leads to more confident and less cautious beliefs. The axiomatized cautious belief formation capture the impact of precision and related cautiousness on the induced belief. The precision and cautiousness concerns are in particular relevant for small databases with differently precise pieces of information, for which the belief formation of BGSS is not suitable, since their axiomatization incorporates some form of irrelevance of growing precision or insensitivity to additional information. Our axiomatization leads to the following representation

$$\text{"Cautious belief":} \qquad P_T(x, D) = \frac{\sum_{c\in C} s(x,c) f_D(c) P^c_{T^*_D}}{\sum_{c\in C} s(x,c) f_D(c)},$$

where $T^*_D := \max_{c\in C} f_D(c)T$ and the estimates $P^c_{T^*_D}$ reflect the impact of precision and related feelings of cautiousness and confidence in estimating based on differently precise pieces of information c in the database, i.e. the estimate based on observing a case T or L times should differ $P^c_T \neq P^c_L$.

Chapter 3[3] covers a belief formation process that does not rely on all potentially available information, but relies only a somehow filtered subset of information. Evidence and insight from psychology, marketing and recent decision theoretic developments show that individuals want to or are constraint to pay attention to (or

[3]Based on Bleile (2014b) "Limited Attention in Case Based Belief Formation", Working Paper, Center of Mathematical Economics, Bielefeld.

recall) only parts of the in principle available information. Unintentional filtering induced by unawareness, cognitive or psychological limitations (limited attention span, cognitive overload, etc.) are plausible sources, but also intentional rough screening heuristic for reasons of cost/effort efficiency are identified as sources in the literature. We incorporate such a pre-stage of filtering information into the axiomatization of a belief formation and allow for a deviation of the "full attention" hypothesis in BGSS and Chapter 2. This two stage approach enables to cover more realistic and cognitively less demanding belief formation processes, which might, for instance, be of interest for bounded rational agents or if the databases are interpreted as memory and the filtering as a retrieving process. An intuitive special case of the axiomatized belief formation is the following similarity satisficing principle (regarding a threshold level of similarity)

$$\text{"Similarity Satisficing belief":} \qquad P(x, D, \Gamma) = \frac{\sum_{c \in D} s(x,c) 1_{\{s(x,c) \geq s^D\}} P^c}{\sum_{c \in D} s(x,c) 1_{\{s(x,c) \geq s^D\}}},$$

where s^D denotes some appropriate database D specific threshold level and Γ represents the filtering process that selects ((un)intentional) the cases (from the database or memory D, i.e. $\Gamma(D) \subseteq D$), that meet the filtering criterion, i.e. surpasses the similarity threshold s^D. Such an agent considers only cases in her memory, which are sufficiently similar and hence relevant for the problem at hand. If one assumes that those cases are retrieved or recalled first which are most similar to the actual problem, then the above principle would just take into account the cases that comes to her mind first and ignore the experiences that were not retrieved quickly enough. In this vein, the retrieval process stops after some time of brainstorming about relevant past cases.

Chapter 4[4] deals with another characteristic of individual information processing. There is a rich literature in cognitive science that shows that human minds structure, process and store information by grouping (or summarizing) them into subgroups or categories, rather than by treating the information piece by piece. Depending on how the categorized information is activated and employed the literature distinguishes between two main procedures. One approach examines only the single prices of information within the most appropriate (target) categories for a specific problem, whereas the information in the other categories are ignored. Another benchmark approach considers only the summarized representative (prototypical) information given by each category and ignores the detailed information in each

[4]Based on Bleile (2014c) "Belief Formation based on Categorization", Working Paper, Center of Mathematical Economics, Bielefeld.

category. Both procedures of categorizing information reduce the cognitive effort and required ability of processing information.

In this chapter we axiomatize belief formation processes based on information that is categorized according to the two mentioned procedures. For a category based belief formation we obtain the following representation, which is based only on the cases of the database that are also in ("target") categories that are activated by a specific problem x (i.e. $\tilde{C}(x, D)$).

$$\text{"Category based belief":} \qquad P(x, D) = \frac{\sum_{c \in \tilde{C}(x,D)} s(x, x_c) P^c}{\sum_{c \in \tilde{C}(x,D)} s(x, x_c)},$$

where s measures the degree of similarity.
A prototype based belief only relies on the prototypical estimates $P^{\tilde{C}_l^D}$ on R for each of the categories $\{(\tilde{C}_l^D)_l\}$ that partition database D.

$$\text{"Prototype based belief":} \qquad P(x, D) = \frac{\sum_{l \leq L} \tilde{s}(x, \tilde{C}_l^D) P^{\tilde{C}_l^D}}{\sum_{l \leq L} \tilde{s}(x, \tilde{C}_l^D}$$

where $\tilde{s}$ measures the relevance or similarity of the particular category for the actual problem. Basically, belief formations based on categorized and thus somehow organized information might reduce the cognitive effort and simplify the belief formation process.

Thus far we have just introduced the main content of the thesis and briefly mentioned the structure of the framework. However, even though the chapters are closely related, they still differ in the particular matter and axiomatizations, such that we postpone the detailed discussion and placement into the relevant literature to the particular chapters.

2 Cautious Belief Formation

Abstract

We provide an axiomatic approach to a belief formation process in an informational environment characterized by limited, heterogenous and differently precise information. For a list of previously observed cases an agent needs to express her belief by assigning probabilities to possible outcomes. Different numbers of observations of a particular case give rise to varying precision levels associated to the pieces of information. Different precise information affects the cautiousness and confidence with which agents form estimations. We modify the Concatenation Axiom introduced in Billot, Gilboa, Samet and Schmeidler (BGSS) (Econometrica, 2005) in a way to capture the impact of precision and its related perceptional effects, while still keeping its normative appealing spirit. We obtain a representation of a belief as a weighted sum of estimates induced by past cases. The estimates are affected by cautiousness and confidence considerations depending on the precision of the underlying observed information, which generalizes BGSS. The weights are determined by frequencies of the observed cases and their similarities with the problem under consideration.

2.1 Introduction and motivation

Beliefs of agents are important ingredients of many economic models dealing with uncertainties. Belief formation is studied recently in environments with limited and heterogenous information that are not suitable to be modeled in the widely used and accepted state space framework of Savage (1954) and Bayes. Lacking a state space representation of uncertainties an agent needs to form her belief explicitly by directly incorporating available information.

We axiomatize a belief formation process based on limited, differently relevant and precise information. Our main axiom modifies the Concatenation axiom in Billot et al. (2005)(BGSS), which precludes the impact of agents' perceptions and reactions to differently precise information. Their axiom says that for any two information sets the belief induced by their combination can be expressed as a weighted average of the beliefs induced by each information set separately. The averaging of beliefs induced by any arbitrary information sets requires a cognitive challenging tradeoff of identical, but differently precise information contained in the particular information sets. Our axiom says that agents who care about precision of information can only average beliefs (in a normatively reasonable way) induced by specific - almost disjoint - information sets. Thereby, we focus not only on the precision itself, but mainly on its perception and impact in form of cautiousness and confidence feelings.

The most prominent and often used models to describe and analyze uncertainty in economic theory are versions of the approach of Savage and Bayes. The fundamental idea in this approach is to model uncertainty by a grand state space, which is sufficiently rich to describe and resolve all possible sources of uncertainties. In this way a state space implicitly incorporates some (perfect) belief and theories about structures and relationships in the future and thereby requires a large (often unachievable) task of imagination and theorization. In addition, insufficient (or too complex) information may preclude the derivation or definition of a grand state space. Another principle of the state based approach is the representation of a belief as an unique probability over the grand state space. In this framework a purely subjective probability distribution over states can be endogenously deduced from preferences, which inherently lacks an explicit description of the formation of the belief that generated the preference.[1]

There is basically only one way to deal with these two difficulties. Sticking to the grand state space principle, but abandoning the subjective prior approach, would precludes an direct (objective) assignment of probabilities, since the state space already encodes all available information. More promising is to give up the representation of uncertainties by a state space, when an agent is (cognitive) incapable to translate information into states or when the information is not all encompassing as needed for the "correct" description of a grand state space. Usually in real life, our information basis can be represented by a list or collection of pieces of information (databases). Thus, we will replace the state space as an information aggregation by such a database representation of (actually observed) information (data-points or recalled cases).

However, a belief based on a database needs to explicitly incorporate factual objective knowledge, characteristics and theoretical considerations provided by the present database. In general, belief formation based on a database is very close to the goal of statistical inference. In contrast to mainly asymptotic considerations in statistical inference, our focus (as in BGSS and Eichberger and Guerdjikova (2010) (EG)) lies on behavioral foundations (axiomatizations) of a belief formation and in

[1]See Gilboa et al. (2012) for extensive discussion of these issues.

particular in the analysis of small databases containing differently precise information.

Usually statistical experiments are dealing with identical observations that are equally relevant. However, since small or medium sized databases contain limited and heterogenous information agents might want to take into account not only (a few) identical but also partially relevant observations for their belief formation. In this sense - differently to statistical experiments - relevance or similarity measures become important for data sets containing limited heterogenous information.

Case-based Decision Theory (Gilboa and Schmeidler (1995, 2001)) deals with such a framework in decision theoretic contexts. BGSS can be interpreted as an adoption of it to belief formation.[2] Their axiomatized belief describes a generalized (subjective) frequentist, in which agents assign different similarity weights to information with different degree of relevance. For a new problem and given a database of past observations,their belief over possible outcomes is represented as a similarity weighted average of estimates that are induced by the observed cases.

Their main Concatenation Axiom deals with relationships between databases and their induced belief. It requires that for a new problem x the belief P (probability vector over outcomes for x) induced by the combination (concatenation) of any two databases ($D \circ E$), is a weighted average of the belief induced by each single database (D and resp. E) separately, i.e. for all databases D and E, there exists a $\lambda \in (0,1)$ such that

$$P(x, D \circ E) = \lambda P(x, D) + (1 - \lambda) P(x, E)$$

Our paper deals mainly with the modification of this Concatenation Axiom in order to allow for impacts of precision of information and its induced cautiousness and confidence concerns. Additionally, our precision dependent belief formation is suitable for small databases, which is only partially possible and reasonable for BGSS.

The Concatenation Axiom shows some irrelevance of growing precision. The belief induced by a database coincides with the belief induced by arbitrary many replications of the same database, i.e. $P(x, D) = P(x, D^T)$ for all $T \in \mathbb{N}$. Growing precision might not be a concern for sufficiently rich and large databases such that observing additional identical information will not affect her predictions. However, for small database, specifically consisting only of one piece of information c, it is unreasonable that additional observations do not induce some learning and refinement of an already "perfect" estimation, i.e. $P(x, c^T) = P(x, c)$ for all T.

In this way the Concatenation Axiom implies that one observation carries already all that can be inferred by arbitrary many confirming observations. Such an in-

[2]Related also to Gilboa and Schmeidler (2003), Gilboa (2009) and Gilboa et al. (2011).

stantaneous learning in a highly objective (and perfect) way of forecasting appears to be questionable and un-intuitive. For instance consider a situation, in which an agent throws a dice once and the figure six results. A guess of the outcome of the next throw of the dice would differ from the estimation an agent would come up after observing one million times a six in one million throws of that dice. However, roughly speaking, the Concatenation Axiom requires that an agent would infer right after the first dice throw that all sides of the dice show the figure six without any doubt. A procedure to base the estimation on just one observation appears to be in-cautious, hazardous (error-prone) and unrealistic and cannot be considered as an appealing normative advice. In fact, as in controlled statistical experiments, additional (identical) confirming observations may serve as a proxy for its increased informativeness, precision or accuracy, which should be reflected in a dynamic learning and refinement of the estimations.

In addition, increasing precision might affect estimations through its perception in form of altered cautiousness (to wrongly eliminate some outcomes) with which the forecast is made and her changed confidence in this forecast.[3] If information becomes more precise, an agent's decreased cautiousness and increased confidence might allow to specify their prediction more narrowly. After receiving substantial information of disconfirming evidence that makes some outcomes negligible, agents (even) might want to eliminate some (not observed) outcomes. More general, differently precise information should lead to different induced beliefs, i.e. $P(x, c^T) \neq P(x, c^L)$ for different $L, T \in \mathbb{N}$, which contradicts the Concatenation Axiom and requires a modification in order to incorporate precision and cautiousness issues.

In general, the Concatenation Axiom is stated for any kind of databases, but (with regard to potentially induced different precise estimations) it is most appealing and appropriate as a normative advice for disjoint databases. For disjoint databases, the belief induced by the concatenated database can be quite intuitively interpreted as an average of the beliefs induced by the single databases separately, since no pieces of information appear in different precision in different databases and cause conflicting considerations. The average is determined solely by a weighting of the relevances of the concatenating databases.

However, we will explain that for unrestricted non-disjoint databases - with common, but differently precise pieces of information - the normative appealing spirit of averaging beliefs conflicts with a simultaneous care about precision and cautiousness

[3]Ellsberg (1961) (p.657): "What is at issue might be called the ambiguity of this information, a quality depending on the amount, type, reliability, and "unanimity" of information, and giving rise to ones degree of "confidence" in an estimate of relative likelihoods."

in the belief formation.[4]

A first obvious modification deals with the issue that a precision related Concatenation Axiom cannot be formulated for pure (non-disjoint) concatenating database, but would require additional information, such that averaging occurs according to $P(D \circ E) = \lambda P(D^L) + (1-\lambda)P(E^T)$ for some appropriate $L, T \in \mathbb{N}$ (which we will specify later). For example consider the concatenation of the easiest non-disjoint databases $(c^7) = (c^3) \circ (c^4)$. By definition, the beliefs induced by combining databases (e.g. $(c^3), (c^4)$) are based on less precise pieces of information (and hence also their weighted average) than the belief based on the combined database (e.g. (c^7)). Thus, for a cautious agent, who cares about precisions, the information contained in concatenating databases may not be sufficient for a belief formation according to the (unmodified) Concatenation Axiom.[5]

However, in general there are no replications T and L, such that each single pieces of information is captured in equal precision in all involved databases $D \circ E$, D^T and E^L. These differences in precision of single common cases complicate the averaging of the beliefs. Determining the average weight cannot anymore be interpreted as normatively appealing comparison based solely on relative relevance of the particular databases. Rather, it is a result of a cognitive challenging (impossible) interwoven tradeoff, balancing and aggregation of different emphasis an agent assigns to single pieces of differently precise information in the various database. Moreover, the average weight might need to reflect also the compensation for failures of the "compulsory" incorporation (by the axiom) of relative more imprecise estimations (based on the same kind of information) contained in some beliefs than in others. Therefore, a (modified) Concatenation Axiom allowing for all (replicated) databases leads to the serious problem that agents might be cognitively overstrained by averaging beliefs based on several identical information with different precision levels.[6]

As a consequence, we propose a restriction on databases to be admissible for our modified version of the Concatenation Axiom, such that it sustains its normative appealing spirit. Our anchored Concatenation Axiom restricts databases to a specific (almost disjoint) structure consisting of only two cases, where only exactly one of these cases (the anchor) appears in all involved databases and all other cases

[4]Stating the axiom only in terms of disjoint databases does not offer sufficient structure to derive a belief formation.

[5]This problematic issue does not appear in the Concatenation Axiom of BGSS, where precision is (endogenously) neglected and one appearance of a case captures already all information.

[6]Alternatively, if one would stick to general non-disjoint databases, then the only way to "unify" the differently precise information in all involved databases is given by assuming ad hoc some arbitrary (imagined) level of precision, according to which all cases are evaluated independent of their actual observation. This will be discussed in detail in Section 2.5 and 2.7.2.

appear only in one database. The main feature is that this single common (anchor) piece of information is contained in all involved databases in equal precision. This enables an easy averaging of beliefs without cognitively demanding compromising between estimations induced by differently precise observations. In addition, the equal appearance of the anchor case in all involved databases intuitively allows to "neglect" its effect in determining the average weights and to compare only the relative importance of the pieces of information that appear only in exactly one of the involved databases. This facilitates a very straightforward way to find the average weights for the beliefs - almost in the spirit of averaging beliefs induced by disjoint (singleton) databases or distinct cases.

In order to take into account the precision of beliefs induced by databases, our agent focusses on the most precise and hence reliable information in the database. Since it is impossible to capture all information in its actual precision level perfectly in a non-disjoint combination of databases (as explained above), our agents require to cover at least the most precise information objectively in her belief. Consequently, our axiom requires that there exists no distortion of the most reliable information in the process of averaging beliefs. To achieve this, the precision of the most reliable information in the concatenated database must be conveyed by the single beliefs induced by the corresponding admissible (sufficiently replicated) databases used for the concatenation.

Besides BGSS, the closest work to ours is the axiomatization of multi-prior beliefs in Eichberger and Guerdjikova (2010) (EG) "Case based belief formation under ambiguity". Their extension of the framework of BGSS aims to formalize two kinds of ambiguity caused by insufficient information (vanishing ambiguity) and irrelevant information (persistent ambiguity). The focus in their paper lies predominantly on the introduction of a multi-prior setup for an information environment with persistent ambiguity. Whereas our work focuses on the analysis of precision in the sense of vanishing ambiguity (imprecision) and related cautiousness in a single prior belief. EG's modification of the Concatenation Axiom of BGSS is adequate to specify how beliefs over outcomes change in response to additional information and tackles also the mentioned drawback of BGSS regarding irrelevance of growing precision of information. Different to our work, their modification reflects the idea of "controlling for the ambiguity " (p.4) (precision) by restricting the involved databases to equal length. However, as discussed above controlling for precision by equal lengths of the involved databases is not sufficient to control for different precise single pieces of information contained in these databases. As a direct consequence, EG's modification of the Concatenation Axiom assumes (and does not prevent) that agents are (cognitively) able to aggregate and balance information of the same kind, but in different precision. In contrast, the focus of our paper lies exactly on the issue to avoid such cognitive challenging or even impossible tradeoffs in the aggregation of differently precise information and to keep the spirit of a normatively reasonable and easy averaging procedure. Moreover, in general EG's axiom implies that no estimation is based on objectively present information in the database, which would require (in our context) that agents need to imagine the (true) cautiousness feeling evoked by a precision level that is imagined as well. In contrast, our approach implicitly requires only the ability to estimate based on already experienced cautiousness and thus avoids imaginations of unexperienced feelings of cautiousness.

In sum, adopting parts of the axiomatization of BGSS and EG, our anchored Concatenation Axiom will allow for a structural similar belief representation as in BGSS and EG. All three axiomatized representations differ in the way, how they treat information of different qualities of precision. BGSS does not take into account precision at all and EG captures the effect of (persistent) ambiguous information by a set of beliefs, which are based on precision-dependent estimates, where the level of (imagined) precision is according to the total amount of information contained in the entire database. In our representation the cautiousness related estimates are based on the level of precision and cautiousness induced by the most precise information in the database. More precisely, for a new problem and a given database, its induced belief can be represented as a similarity-weighted average of cautious estimates induced by past observations in the database. Thereby the similarities and estimates are endogenously derived.

The remainder of the chapter is organized as follows. In the next section we will outline the model and in Section 2.3 we develop an example to illustrate reasonable belief formations and our leading example. Then, the axioms are stated and discussed, where the central Section 2.4.2 points out the drawbacks and necessary modification of the Concatenation Axiom to incorporate precision, which eventually leads to our version of the Concatenation Axiom. Section 2.5 presents and discusses the main representation result and gives a rough sketch of the main part of the proof. Section 2.6 concludes the chapter and Section 2.7 deals with an objective belief, the relationship to EG's Concatenation Axiom and an alternative axiomatization of a very cautious belief. The last two sections contain both directions of the proof.

2.2 The model

2.2.1 Database framework

A basic case $c = (x, r)$ consists of a description of the environment or problem $x \in X$ and an outcome $r \in R$, where $X = X^1 \times X^2 \times .... \times X^N$ is a finite set of all characteristics of the environment, in which X^j denotes the set of possible values features j can take. R denotes a finite set of potential outcomes, $R = \{r^1, ..., r^n\}$.
The set $C \subseteq X \times R$ consists of all $m \geq 3$ basic cases, i.e. $|C| = m$.
A database D is a sequence or list of basic cases $c \in C$. The set of databases D consisting of L cases, i.e. $D = ((x_1, r_1), ..., (x_L, r_L))$ is denoted by C^L and the set of all databases by $C^* = \cup_{L \geq 1} C^L$. The description of databases as sequence of potentially identical cases allows multiple observation of an identical case to be taken into account and treated as an additional source of information.
For a database $D \in C^*$, $f_D(c)$ denotes the relative frequency of case $c \in C$ in databases D.
The concatenation of two databases $D = (c_1, c_2, ..., c_L) \in C^L$ and $E = (c'_1, c'_2, ..., c'_T) \in C^T$ is denoted by $D \circ E \in C^{L+T}$ and is defined by $D \circ E := (c_1, c_2, ..., c_L, c'_1, c'_2, ..., c'_T)$.
In the following we will abbreviate the concatenation or replication of L-times the identical databases D by D^L. Specifically, c^L represents a database consisting of L-times case c.

Definition 2.1
(i) The $\in$- relation on databases is defined by $c \in D$ if $f_D(c) > 0$.
(ii) Two databases D and E are disjoint if for all $c \in C$: $c \in D$ if and only if $c \notin E$.

2.2.2 Induced beliefs

For a finite set S, $\Delta(S)$ denotes the simplex of probability vectors over S and for $n \in \mathbb{N}$ Δ^n denotes the simplex over the set $\{1, 2, ..., n\}$.
An agent will form a belief over the outcomes $P(x, D) \in \Delta(R)$ for a certain problem characterized by $x \in X$ using her information captured in a database $D \in C^*$, i.e. $P : X \times C^* \rightarrow \Delta(R)$. The restriction to databases of length T is denoted by $P_T(x, D) \in \Delta(R)$ for $D \in C^T$ and $P_T : X \times C^T \rightarrow \Delta(R)$.
One can interpret $P_T(x, D)$ as the belief over outcomes induced by database $D \in C^T$ (given environment or problem $x \in X$).
Throughout the paper the problem x is fixed, therefore x is often suppressed in the following, i.e. $P(x, D) = P(D)$.

2.3 Motivating examples

2.3.1 Exemplary development of a belief formation process

A doctor needs to evaluate the likelihood of potential outcomes of a specific treatment.
Let a patient be described by a vector of characteristics $x \in X$, where X might consist of measures of characteristics like age, gender, weight, height, blood pressure, temperature, blood count, vital signs, medical history, drug tolerability, etc.
The doctor might have observed several outcomes of the treatment in the past, which are collected in a set R, containing for instance feels better/worse/unchanged, measures of side effects like headaches, sleepy, depressive, passed out, giddy, dizzy, etc.
The doctor has acquired some working experience prescribing this treatment and/or has access to some medical record of this treatment. Thus, she is able to base her judgement on past experience or observations collected in a database $D = (c_1, ..., c_T)$, where in each case c_i the characteristic and the observable outcome of patient i is recorded, i.e. $c_i = (x_i, r_i)$. It means that a patient characterized by $x_i \in X$ responded to the treatment with outcome $r_i \in R$.
Given the characteristics $x \in X$ of a current patient and her available information and experience in form of a database D, the doctor derives a probabilistic belief $P(x, D) \in \Delta(R)$ over potential outcomes in R for this treatment. How can she do the prediction?

a) A first intuitive approach for the prediction is to consider only patients in the database that are identical (with respect to the measured characteristics) with the present patient. Based on this sub-sample $D_x := (c \in D | c = (x, r_i) \text{ for some } r_i \in R) \subseteq D$ the doctor might derive a prediction over potential outcomes via **empirical frequencies**:

$$P(x, D) = \frac{\sum_{c_j \in D_x} \delta_{r_j}}{|D_x|},$$

where δ_j is the probability vector on R with mass 1 on the outcome $r_j \in R$. Of course this belief formation process is not practical if the sub-sample D_x contains only few observations, i.e. if there are only a couple of identical (with respect to the measured characteristics) patients.

b) To overcome this problem of limited or insufficiently many identical observations, the doctor might include into her prediction procedure not only identical, but in addition also similar patients. Suppose that she is able to judge how similar patients are, i.e. she is able to employ a function s : $X \times X \to \mathbb{R}$, where $s(x_t, x_j)$ measures the degree of similarity between patients characterized by x_t and x_j. Her belief formation process might run in a **"subjective" frequentist** way:

$$P(x, D) = \frac{\sum_{c_j \in D} s(x, x_j)\delta_{r_j}}{\sum_{c_j \in D} s(x, x_j)}$$

c) In addition, the doctor might infer from a case $c_j = (x_j, r_j) \in X \times R$ not only a point prediction on δ_{r_j}, but a more general induced estimation P^{c_j}.[7] For instance, she attaches also some likelihood to outcomes that are closely or reasonably related to the observed r_j:

$$P(x, D) = \frac{\sum_{c \in D} s(x, x_j)P^c}{\sum_{c \in D} s(x, x_j)}. \tag{2.1}$$

This belief formation process is axiomatized in **BGSS**.

d) Furthermore, the doctor might process the past observations not in an one by one estimation problem as in the approaches above, but might want to sample the database beforehand according to identical cases (i.e. patient-outcome pair). Many observations of the same case might foster some learning and improved understanding of the relationship between characteristics of a patient and corresponding outcomes. Additional confirming observations should affect the judgment of a cautious doctor as well by an increased confidence and decreased cautiousness in predicting the observed outcome. In this way, the doctor might generate different predictions depending on how many observations of this case are present in the database.

For instance, suppose there is a generally observed side-effect of many different medicines, then the doctor might still assign a positive likelihood to this side-effect if the doctor has observed just a few (identical) patients not suffering from this side effect under the specific treatment. However, if the treatment is well established and many identical patients did not feel this side-effect, then she might not consider this side-effect as a potential hazard anymore. This intuition can be modeled by incorporating precision into a cautious belief formation, where the number of observation can be interpreted as a proxy for the precision of the information:

$$P(x, D) = \frac{\sum_{c \in D} s(x, c) f_D(c) P^c_{T_D(c)}}{\sum_{c \in D} s(x, c) f_D(c)},$$

[7] More precisely, actually $P^c = P^{(x,c)}$ represents an (conditional) estimate induced by c given the current patient x, i.e. if c is totally unrelated to the current patient, it might be that P^c is uniform on R.

where P_L^c represents the precision dependent estimation on R induced by L observation of case c, where usually $P_T^c \neq P_L^c$ for $T \neq L$. Hence $T_D(c) \in \mathbb{N}$ denotes a database dependent precision (and induced cautiousness) level, according to which a doctor will estimate the outcomes based on observation of case c.

Interestingly, the already mentioned belief formations of BGSS and EG are special cases of this representation:
(i) BGSS's axiomatization implies $T_D(c) = \infty$. Hence, their agent learns instantaneously the "correct" distribution $P_\infty^c = P^c$ induced by case c (see representation (2.1)).
(ii) The axiomatic derivation in EG results in $T_D(c) = T$ for $D \in C^T$.

(e) A natural interpretation of $T_D(c)$ in (d) is $T_D(c) = f_D(c)T$ for all $c \in D \in C^T$:

$$P_T(x, D) = \frac{\sum_{c\in D} s(x,c) f_D(c) P^c_{f_D(c)T}}{\sum_{c\in D} s(x,c) f_D(c)}, \tag{2.2}$$

where $f_D(c)T$ gives the actual number of appearance of case c in database D. Such a representation is very objective by incorporating only actually available and observed information. This representation is (unfortunately) irreconcilable (see Section 2.7.1) with any generalized version of a Concatenation Axiom (in the sense of not only combining disjoint databases), which we consider as an important behavioral component of a belief formation.

However, unless its appealing objective character, this belief formation might entail the following problem. Obviously, the belief employs (in general) an aggregation of estimates $(P^c_{f_D(c)T})_{c\in D}$ based on different precise single pieces of information c, which carry different deficits in their "correctness" of prediction. This might overcomplicate the evaluation since the doctor might want to accompany the fact that some of her predictions are more reliable than others and should receive more weight independent from the similarity and frequency weighting. In this sense, she might want to include an additional weighting scheme taking into account the precision or reliability of the estimates.[8]

f) However, a doctor might not only rely on objective precisions of the estimations, but also wants to capture her perception of its precision, i.e. the influence of how cautious and confident she feels while estimating $P^c_{T_D(c)}$. In this vein, we prefer a

[8]Alternatively, these considerations might be incorporated into the weights s, which prevents an desirable independent interpretation in similarity terms and requires the function s to depend also on the databases directly, which will be precluded (later) by our Constant Similarity Axiom. In addition, it conflicts with the easy averaging intuition

different choice for $T_D(c)$ in order to take the doctor's cautiousness and confidence concerns into account.

The underlying intuition is that she does not change or adjust constantly her cautiousness and confidence attitude in response to each differently precise information. Rather, after the doctor has experienced an (extreme) level of cautiousness and confidence by estimating based on objectively available (unimagined) information, she might keep it and adopt it to other estimations. Basically she attained an "appropriate" sustainable attitude regarding her cautiousness sensation or learned how to confidently estimate sufficiently cautious and applies it to all remaining estimations. This also overcomes the mentioned potential disfavor of aggregating different precise estimations emerging in the objective belief formation (2.2).

The most intuitive choices for a cautiousness attitude are the two extreme perceptions, i.e. the experience of minimal and maximal cautiousness, which are induced by the most or least precise information in the database. A minimal cautious attitude might distract from any other more cautious perceptions, since the doctor learned how to handle information in an appropriate cautious way. A maximal cautious agent might be intimidated by the largest experienced imprecision and can not be convinced to leave her skeptical mood to adopt a more confident attitude for estimating according to the available more accurate information.

The following **cautious belief formation** captures these ideas (for a attitude of minimal cautiousness) and will be axiomatized in this chapter:

$$P_T(x, D) = \frac{\sum_{c \in D} s(x, c) f_D(c) P^c_{T^*_D}}{\sum_{c \in D} s(x, c) f_D(c)},$$

where $T^*_D := \max_{c \in C} f_D(c)T$.

The above examples were intended to clarify the framework and demonstrate a meaningful evolution of a belief formation taking into account subjective and precision concerns. However, in the following we will use a reduced version as our leading example.

2.3.2 Leading example

Assume that the patients are not anymore described by a large vector of their personal characteristics, but just according to their symptoms or diagnosed sickness. In particular, each patient is characterized by just a single symptom and the outcome of a treatment is only roughly distinguishable between w(orse), n(ot affected) or b(etter), i.e. $R = \{w, n, b\}$.

So basically a doctor has prescribed a certain medicine to many patients with different symptoms or illnesses and observed the outcome of this drug, i.e. a case is described as a pair of symptom and outcome of the treatment. For example the drug improved the state of patients suffering from sore throat, but was harmful for most patients suffering from stomach problems.

2.4 Axioms

In the first part we state modified versions of the axioms in BGSS that we will adopt. The second main part discusses in detail the Concatenation Axiom and its drawbacks in a precision dependent framework that eventually leads to our new anchored Concatenation Axiom.

2.4.1 Retained axioms

Invariance Axiom

For every $T \geq 1$, every database $D = (c_1, ..., c_T) \in C^T$ and any permutation π on $\{1, ..., T\}$

$$P_T((c_1, ..., c_T)) = P_T((c_{\pi(1)}, ..., c_{\pi(T)})).$$

The Invariance Axiom states that an induced belief over outcomes depends only on the content of that database and is insensitive to the sequence or order in which data arrives.

However, the order in which information is provided or obtained can influence the judgment strongly and may carry information by itself (e.g. see Rubinstein and Salant (2006)). For example, first and last impressions or reference effect demonstrate the different impacts of cases depending on their positions. One way to cope with these order effects is to describe the cases informative enough. E.g. if one wants to capture the position or time of occurrence of a case in a database, one could implement this information into the description of the cases itself. Put differently, if one challenges the Invariance Axiom, then there must be some criteria which distinguish the cases at different positions in a database and paying attention explicitly to this difference in the description of the cases may lead the agent to reconcile with the invariance assumption.

Hence, we will base our belief formation only on the content of the database D, which allows to characterize each D by the pair of its frequency vector and length.

Learning Axiom

For every $c \in C$ the limit of $(P_T(c^T))_T$ exists, i.e. the sequence converges to P^c_∞.

In the context of precision dependent beliefs the axiom can be interpreted as a stable learning process. For instance, an agent starts out with an initial prior (like a uniform as in the principle of insufficient reason) that will be adjusted in the process of observing additional information. Increasing the number of confirming observations will lead to vanishing imprecision and cautiousness in estimating. Basically, the estimate will become less sensitive to new additional confirming information and will eventually converge to a limit distribution. This intuition is as in Bayesian updating, where additional (confirming) information may render the prior beliefs more precise, but differently to Bayesian updating the support might change here. For instance, it is reasonable to assume that finally the agent will learn the true distribution of a case $c = (x, r) \in C$ given the problem $x \in X$, i.e. $\lim_{T\to\infty} P(x, c^T) = \delta_r$, where δ_r is again the Dirac measure. However, for a problem $x' \neq x \in X$ the belief might just converge to a general uniform-like distribution on R, since the observed case might not give relevant information for the current problem at all. Hence, we require only that a limit estimation exists.

Another intuition, that we mentioned already, runs as follows. T many observations a case $c = (x, r)$ might not make a cautious agent feel confident to reliably rule out a non-observed outcomes completely, but she wants to assign at least some positive likelihood to it, i.e. $P_T((x, r)^T)(r') > 0$. However, observing further confirming cases might carry sufficient evidence, such that an agent would feel confident not to make a mistake or act incautious in excluding some outcomes, i.e. $P_L((x, r)^L)(r') = 0$ for $L \gg T$.

Alternatively, one can apply a (accordingly adjusted) learning procedure as in Epstein and Schneider (2007), where an agent might start out with a uniform estimation and after observing new information keep only the most plausible estimates. Plausible estimations in their sense are those that survive a maximum likelihood test (according to some strictness parameter, which might correspond with a cautiousness measure in our setup) against the belief that best explains the observations, i.e. the Dirac measure on the observed outcome.

Diversity Axiom
There exist $T^* \in \mathbb{N}$, such that for all $T \geq T^*$, no three of $\{P_T(c^T)\}_{c\in C}$ are collinear.

From a technical point of view this axiom allows to derive an unique similarity function, but it also carries an appealing intuition. Roughly it states that sufficiently many observations induce always estimations that are informative (or diverse) in the sense that no combination of two other sufficiently often observed cases can deliver the same estimation. Hence, no sufficiently precise case can be "replaced" by sufficient observations of two other cases in this sense. The reason to base the diversity of induced estimation on a precision threshold T^* is the following. In order to derive unique similarity values one could also require non-collinearity for every value of T, but this would exclude learning as mentioned in the description of the Learning Axiom. If an agent would start out with an uniform-like prior for databases containing few observations, it might happen that different cases induce very similar estimations, which are likely to be collinear. The axiom just rules out that after a sufficient learning period any three estimations are still collinear.

2.4.2 Different versions of the Concatenation Axiom

Concatenation Axiom of BGSS
For every database $D, E \in C^*$ there exists some $\lambda \in (0,1)$ such that:

$$P(D \circ E) = \lambda P(D) + (1-\lambda) P(E).$$

In the following we will call the database that emerges from concatenations of other databases the **combined or concatenated** database, whereas the databases used for the concatenation will be called **combining or concatenating** databases. We call the weights $\lambda, (1-\lambda)$ **average weights.**

The Concatenation Axiom states that the belief induced by a combined database is a weighted average of the beliefs induced by its combining databases. It captures the idea that a belief based on the combined database cannot lie outside the interval spanned by the beliefs induced by each combining database separately. Intuitively it can be interpreted in the following way (from an exclusion point of view). If the information in any database induces an belief that does not exclude an outcome r, then the outcome r cannot be excluded by the belief induced by the combination of all these databases.[9] Alternatively, if a certain conclusion is reached given two databases, the same conclusion should be reached given their union.

The normatively appealing spirit of the axiom is that the average weights are determined by relative relevances or importance of the combining databases for its combination.

As already mentioned, the Concatenation Axiom implies an irrelevance of growing precision or insensitivity to additional information in the beliefs, i.e. $P(D) = P(D^Z)$

[9] Of course the axiom is stronger in the sense that it not only requires that the probability of such an r is positive, but it should lie between the minimal and maximal assigned probabilities induced by the combining databases.

for all $D \in C^*$ and $Z \in \mathbb{N}$, which might be appropriate for sufficiently rich and large databases. However, already BGSS admit that it "... might be unreasonable when the entire database is very small ...." (BGSS (2005), p. 1129).[10] Indeed, the axiom induces some sort of perfect objectivity and instantaneously learning. Estimation based on one observation $c = (x, r)$ needs to coincide with the estimation induced by arbitrarily many observation, which can be identified in some sense with the "true" limiting distribution, i.e. $P(x, (x, r)) = P(x, (x, r)^\infty)$. For our leading example it would mean that a doctor would predict after one unsuccessful treatment of a sore throat that this treatment is worthless for (identical) patients suffering from sore throat. However, this appears very unrealistic and un-intuitive, since a database $c = (x, r)$ might be considered more imprecise and might induce a more cautious belief than $c^T = (x.r)^T$ for sufficiently many observations T, i.e. $P(c) \neq P(c^T)$.

In order to incorporate precision and cautiousness aspects into the belief formation process, the Concatenation Axiom needs to be modified in various ways to maintain its normative appeal in a modified framework.

For this purpose an immediate modification concerns the issue that an agents can not rely on beliefs induced by the concatenating databases directly, but requires appropriately replicated concatenating database[11], i.e.

$$P(D \circ E) = \lambda P(D^T) + (1 - \lambda) P(E^L) \text{ for appropriate } T, L \in \mathbb{N} \tag{2.3}$$

The reason for this is that the information contained in non-disjoint concatenating databases appears by definition in less precision as in their concatenation, e.g. consider $c^{2Z} = c^Z \circ c^Z$. However, for a cautious agent caring about precision, $P(c^Z)(r) > 0$ does not necessarily imply $P(c^{2Z})(r) > 0$. For example our doctor might not want to rule out a successful treatment of a coughing agent after observing 20 or 30 unsuccessful treatments according to her perceived cautiousness, i.e. $P(c^T)(r) > 0$ for $T = 20, 30$. However, the combined information of 50 unsuccessful treatments on coughs might make her feel confident and convinced to evaluate the treatment as useless for curing a cough without violating her cautiousness feeling, i.e. $P(c^{50})(r) = 0$. Thus, non-disjoint concatenating databases do not carry sufficient information to capture refinements of a cautious belief implied by the concatenated database and as stated in (2.3) more precise (i.e. appropriately replicated) concatenating information is required.

[10] From this perspective, our modification can be interpreted as an extension of BGSS to derive a belief formation also for relatively small databases, which is only partially possible and reasonable given their Concatenation Axiom.

[11] This problem emerges only if the databases are non-disjoint. However, to allow only disjoint databases in the Concatenation Axiom does not offer enough structure to derive a belief.

Furthermore, in general there exist no replications T and L that ensure that each case in $D \circ E$ is captured in identical precision for unrestricted non-disjoint D^T and E^L.[12] Depending on its precision the same case might induce differently cautious estimations. This leads to the difficulties that agents need to balance the differently cautious estimations induced by the same case appearing in different precisions in the replicated concatenating databases. Such a compromising between estimates is necessary for all cases that are observed in more than one database. For instance, our doctor compares the (replicated) databases $D^2 = (c_1^4, c_2^6, c_3^4)$ and $E^2 = (c_1^4, c_2^8, c_3^2)$ (and eventually average its induced beliefs), where each (replicated) database contains differently many observations of harmfully treated colds c_2 (6 vs. 8), neutrally treated colds c_3 (4 vs. 2) and at least the successfully treated sore throats c_1 are observed identically often (4 vs. 4) (by replicating D and E, with the focus on unifying according to c_1). Hence each induced beliefs rely on different precision with regard to observations of cases c_2 and c_3. How could an objective doctor compare and average the differently precise information incorporated in these databases? Intuitively, the doctor should use the most precise available information contained in these databases. Information c_2 is contained in the belief induced by E^2 in a more precise fashion than in database D^2 and hence the doctor would like to rely predominately on (i.e. assign high weight to) E^2 regarding c_2 (since P_8^{c2} vs P_6^{c2}) and to ignore the less precise estimation wrt. c_2 in D^2. However, the opposite is true for the precision of information c_3, for which she relies predominately on D^2 and ignores E^2.

Anyway, such a reasonable behavior is not admissible in any version of a Concatenation Axiom, in which an agent is forced to assign exactly one (non-zero) average weight to the beliefs induced by the entire databases D^2 and E^2 and not many different weights to the estimates induced by the single pieces of information contained in the databases.[13] In order to reach one "aggregated" average weight, these single weights would need to be balanced, traded off and aggregated somehow. In particular, since the beliefs induced by D^2 and E^2 contain induced estimates that are too imprecise and cautious in comparison to other available ones, our doctor needs to offset and capture these imprecisions and mistakes by adjusting the average weights accordingly. However, a determination of the average weight as a result of difficult balancing and interwoven compromising appears to be even in this easy

[12] This follows from the general non-existence of solutions T, L to the system of equations resulting from $f_D(c)|D|T \in \{0, f_{D\circ E}(c)|D \circ E|\}$ and $f_E(c)|E|L \in \{0, f_{D\circ E}(c)|D \circ E|\}$ for all $c \in D \circ E$.

[13] However, this potential cognitive difficulties are not an issue in the way the Concatenation Axiom of BGSS processes information, where information is additive in the sense that L observations in one database and T observations in another is equivalent to observing $T + L$. Since one observation caries all information.

example rather cognitively challenging and becomes impossible for more complex (decompositions of) databases. Further and most importantly, it conflicts with the normatively appealing spirit of the Concatenation Axiom to average beliefs by an easy comparison of relevances of the particular underlying databases.

Our modification of the Concatenation Axiom will deal with this problem by restricting it to specifically structured database such that balancing and compromising due to differently precise information is avoided and the cognitively simple averaging intuition sustains.

However, interestingly, following the idea of a Concatenation Axiom that still allows for unrestricted non-disjoint concatenating databases would eventually arrive at an (not yet given) intuition and explanation for the modification of the Concatenation Axiom followed in EG. The basic idea is that agents tackle the immense compromising considerations of different cautious estimations by assuming or choosing a common arbitrary level of precision, according to which all cases are estimated - independent of their true objective precision. Since objective or imagined precisions might evoke different feelings of cautiousness such an approach would interfere with our purpose to seriously take into account objective precision and its related concerns. A more detailed discussion on that and on EG's modification can be found in Section 2.7.2.

Anchored Concatenation Axiom

The above discussion shows that a Concatenation Axiom for unrestricted non-disjoint concatenating databases might destroy the underlying normatively appealing idea of an easy averaging, when agents care about precision and its perceptional consequences. In order to keep the normative appealing spirit, we will restrict the involved databases to a specific reasonable structure. These databases will contain sufficiently precise information (in the sense of (2.3)) and allow an cognitively easy averaging. We have seen that an agent will run into a difficult balancing process to determine the average weights when she is faced with concatenating databases containing common cases. For this reason, our anchored databases are as disjoint as possible, but still sharing a specific (exploitable) structure to facilitate an easy comparison (and in the end a straightforward averaging of its induced beliefs). In particular, the anchored databases consist of only two different cases, where all anchored databases admissible for the concatenation contain a common anchor (reference) case with identical frequency and one additional mutually different case in

each of the databases.[14] Apart from the desire to employ databases that are almost disjoint, their structure is also driven by the general observation that agents can compare items easier if they consist of less features (here: only two) and if they contain common features in the same fashion as a reference (here: anchor case).
Recall that $m \in \mathbb{N}$ denotes the number of basic cases, i.e. $|C| = m$.

Definition 2.2
Let $k \in [0,1]$ and $T_j \in \mathbb{N}$ be s.th. $kT_j \in \mathbb{N}$ for all $1 \leq j \leq m$ and let $T := \sum_{j \neq i \leq m} T_j$.
Let $c_i, c_j \in C$ for all $j \neq i \leq m$
(i) For all $j \neq i \leq m$ a database $D_i^j(k, T_j) \in C^{T_j}$ defined by

$$D_i^j(k, T_j) : \;=\; (c_j^{(1-k)T_j}, c_i^{kT_j})$$

is called an ***anchored database*** *of length T_j with* ***non-anchor case*** *c_j (for all $j \neq i$) and* ***anchor (case)*** *c_i, which appears in the database with frequency k.*
(ii) An ***anchored chain*** *$F \in C^T$ (wrt. to case c_i) is defined as a concatenation of anchored databases $D_i^j(k, T_j) \in C^{T_j}$ for all $j \neq i$ (with common anchor case $c_i \in C$)*

$$F = \circ_{j \neq i \leq m} D_i^j(k, T_j) = (c_1^{(1-k)T_1}, ..., c_{i-1}^{(1-k)T_{i-1}}, c_i^{kT}, c_{i+1}^{(1-k)T_{i+1}}, ..., c_m^{(1-k)T_m})$$

Note hat not all databases can be interpreted as an anchored chain, since it requires to be a result of a concatenation of specifically structured anchored databases.

In order to illustrate the anchor-framework, we use our leading example of a doctor that forms a belief over the outcomes of a treatment -worse, no effect, better- $\{w, n, b\}$.
For our doctor anchored databases and chains might look as follows. Each involved (anchored) database consist of only two different cases (patient groups or symptom-outcome pairs), where one of these groups (the anchor case) needs to be observed in all involved database, e.g. patients with a successful treatment (b) of their cough (c) might be the anchor group (i.e. $c_1 = (c, b)$). The other patient group observed in each database is different in all involved databases, for instance the different non-anchor groups might be patients with a neutral treatment (n) of their sore throats (st) (i.e. $c_2 = (st, n)$) or stomachache problems (i.e. $c_3 = (s, n)$) or harmful treatment (w) of patients suffering from sore throats (st) (i.e. $c_4 = (st, w)$).
To simplify the comparison of the databases (by providing a systematical structural guideline) the anchored database contain the (anchor) group c_1 in a specific proportion k (e.g. $k = \frac{2}{3}$) of the databases' total length. E.g. each database consisting of

[14] In some sense, one can interpret the restriction to such database by agents feeling to be cognitively skilled or capable to confidently compare only such easily structured databases.

two thirds of successfully treated coughing patients and one third of patients with any other mutually different (symptom,outcome)-pair.

- For example a anchored database consists of 20 successfully treated coughs (i.e. c_1^{20}) and 10 neutrally treated sore throats (i.e. c_2^{10}), which results in the anchored database with 30 patients $D_1^2(\frac{2}{3}, 30) = (c_1^{20}, c_2^{10})$.
- Another database might contain 40 successfully treated coughs (i.e. c_1^{40}) and 20 neutrally treated stomachaches (i.e. c_3^{20}), i.e. anchored database $D_1^3(\frac{2}{3}, 60) = (c_1^{40}, c_2^{20})$ with 60 patients.
- Another anchored database consist of 16 successfully and 8 harmfully treated coughs (i.e. c_1^{16} and c_4^{8}), i.e. $D_1^4(\frac{2}{3}, 24) = (c_1^{16}, c_4^{8})$ with 24 patients.

The corresponding anchored chain based on $D_1^j(\frac{2}{3}, T_j)$ (for $j = 2, .., 4$ and $T_2 = 30, T_3 = 60,\ T_4 = 24$ and $T := \sum_{j=2}^{5} T_j = 114$) reads $F = \circ_{j=2}^{4} D_1^j(\frac{2}{3}, T_j) = (c_1^{76}, c_2^{10}, c_3^{20} c_4^{8})$.

However, within the anchor structure the comparison of the almost disjoint anchored databases $(D_i^j(k, T_j))_{j \neq i}$ is still not directly straightforward, since the precision of the anchor case c_i in each of the database varies with the corresponding lengths T_j, i.e. c_i is contained in $D_i^j(k, T_j)$ in the amount of kT_j. In our leading example reflected by the different numbers of successfully treated coughing patients. These difference in the precision would again cause the already extensively discussed difficulties in determining the average weights. In order to avoid this problem and as well to respond to the issue of insufficiently precise information in non-disjoint concatenating database (see equation (2.3) and its derivation), we need to replicate some of the anchored database to attain a common level of precision for the anchor case. Due to the identical structure of the databases, enforcing a common precision for anchor case is equivalent to obtain a specific common length L for all involved anchored databases.[15] More precisely, for an anchored chain F of $(D_i^j(k, T_j))_{j \neq i \leq m}$ a belief induced by F should rely on an average of the beliefs induced by anchored databases $(D_i^j(k, L))_{j \neq i \leq m}$. Obviously, this enables an agent to easily compare the involved databases $D_i^j(k, L)$ since their only common case -the anchor case c_i - appears in identical amounts kL in all databases. Therefore, in comparing the anchored database (and determining the average weights) the agent can concentrate on the single and mutually different non-anchor cases.

[15] This seem to be close to the EG approach in fixing the lengths of the databases. However, here it is a consequence of fixing a common precision for a single case. The two approaches use different incompatible restrictions on the databases involved in the modifications of the Concatenation Axiom.

It remains to specify and motivate a choice for a common precision level of the anchor case and (indirect) the common length L. We will introduce it in close relationship to our notion of the precision of an induced belief. As already discussed in general for non-disjoint databases in the last section (see discussion after equation (2.3)), there exists no replication for anchored concatenating databases, such that all single cases appear in equally precision in $(D_i^j(k,L))_{j\neq i\leq m}$ and in the related anchored chain $F = \circ_{j\neq i\leq m} D_i^j(K,T_j)$.[16] Obviously, this leaves the freedom to reason for a specific piece of information that should be captured in equal precision in all involved induced beliefs. A very intuitive (and from our point of view most reasonable) choice to control for precision (and related confidence and cautiousness) is to ensure that the most precise and hence reliable piece of information in the anchored chain is captured in the identical precision in the beliefs induced by the corresponding replicated combining databases.[17] The focus and reliance on the most precise case can be justified by interpreting it as the driving factor of the precision of the belief. Focussing on another, less precise information would imply a less precise belief, since the most precise information would not be captured objectively anymore (in all involved databases).[18] Hence it appears reasonable to require that at least the most reliable information is incorporated in the belief without any distortions, which requires that it is also contained unbiased its generating (averaging) beliefs. More technically, this can be achieved by requiring a particular adjusted length of the combining anchored databases, which is given in the following definition.

Definition 2.3
Let $F \in C^T$ be an anchored chain of $(D_i^j(k,T_j))_{j\neq i\leq m}$.
*A length $L \in \mathbb{N}$ is called the **adjusted (maximal) length**, denoted by $L(k,(T_j)_{j\neq i\leq m})$, if it is such that the number of observations of the most frequent case in an anchored chain $F \in C^T$ is identical to the number of observations of the most frequent case in the anchored databases $D_i^j(k,L)$ (for all $j \neq i$), (i.e. $\max_{c\in C} f_F(c)T = \max_{c\in C} f_{D_i^j(k,L)}(c)L$).*[19]

Our leading example will clarify the relationships and intuition of the adjusted

[16]This is due to the different appearances of the cases, i.e. for the anchored chain the appearance of a non-anchor case c_j is $(1-k)T_j$ (for all $j \neq i \leq m$) in contrast to $(1-k)L$ in the (replicated) anchored databases $D_i^j(K,T_j)$ and similar for the anchor case c_i, there exists the difference between $kT = k\sum_{j\neq i\leq m} T_j$ and kL.

[17]Another reasonable choice is the least precise information that a very cautious agent might adopt (see Section 2.7.3).

[18]Section 2.5.1 discusses another interpretation in terms of an induced persistent cautiousness attitude that is evoked by the most precise information in the database and serves as basis for all other estimations.

[19]In Section 2.7.3 the maximum is replaced by a minimum to focus on least precise information.

length.

Example:

(i) Our doctor considers the records of different patient groups collected in two studies, i.e. $D_1^2(\frac{2}{3}, 30) = (c_1^{20}, c_2^{10})$ and $D_1^3(\frac{2}{3}, 60) = (c_1^{40}, c_2^{20})$ with common patient group c_1. Patient group c_1 is also the most precise information (with 60 observations) in the corresponding anchored chain $F = (c_1^{60}, c_2^{10}, c_3^{20}) \in C^{90}$. Thus the doctor requires it be matched equally precise in appropriate replications (of the study results) of the anchored databases $D_1^2(\frac{2}{3}, 30)$ and $D_1^3(\frac{2}{3}, 60)$. The adjusted length L such that for $j = 2, 3$

$$60 = \max_{c \in F} f_F(c) 90 = \max_{c \in D_1^j} f_{D_1^j}(c) L = \max\{\frac{2}{3}, \frac{1}{3}\} L = \frac{2}{3} L,$$

is given by $L = 90$, i.e. $D_1^2(\frac{2}{3}, 90) = (c_1^{60}, c_2^{30})$ and $D_1^3(\frac{2}{3}, 90) = (c_1^{60}, c_2^{30})$. Obviously, the most precise case c_1 is capture in identical precision (60) in all three databases $F, D_1^2(\frac{2}{3}, 90)$ and $D_1^3(\frac{2}{3}, 90)$. This allows an easy averaging of beliefs induced by $D_1^j(\frac{2}{3}, 90)$.

(ii) Similarly, let there be two public studies of the treatment for some specific patient groups summarized in the following anchored chain

$$F = (c_1^{30}, c_2^{40}, c_3^{80}) = (c_1^{10}, c_2^{40}) \circ (c_1^{20}, c_3^{80}) = D_1^2(\frac{1}{5}, 50) \circ D_1^3(\frac{1}{5}, 100),$$

where again the anchor patient group is "successfully treated coughs" c_1. The most precise case in F is c_3 (with 80 observations). This implies that an adjusted length $L = 100$ is determined by $80 = \max\{\frac{1}{5}, \frac{4}{5}\} L$. Again, the most precise case c_3 is capture in identical precision (80) in the relevant databases F and $D_1^2(\frac{1}{5}, 100) = (c_1^{20}, c_3^{80})$, $D_1^3(\frac{1}{5}, 100) = (c_1^{20}, c_2^{80})$.

With these definitions at hand we can state our anchored Concatenation Axiom, where a modified version focussing on least precise information can be found in Section 2.7.3.

Recall that the length T of a database in an induced belief P becomes visible via the restriction to P_T. In particular for anchored databases $D_i^j(k, T_j)$, we can skip the length T_j in the induced belief, i.e. $P_{T_j}(D_i^j(k, T_j)) = P_{T_j}(D_i^j(k))$.

Maximal Anchored Concatenation Axiom:
(i) Let $F \in C^T$ be an anchored chain of $(D_i^j(k,T_j))_{j \neq i \leq m}$, i.e. $F = \circ_{j \neq i}^m D_i^j(k,T_j)$ and let $L = L(k,(T_i^j)_{j \neq i}) \in \mathbb{N}$ be the corresponding adjusted (maximal) length. Then there exists $\lambda \in \Delta^m$ (where $\lambda_j = 0$ for all $j \leq m$ s. th. $T_j = 0$) such that

$$P_T(F) = \sum_{j \neq i \leq m} \lambda_j P_L(D_i^j(k))$$

(ii) Let for three distinct $i,j,l \leq m$ and any $V,W \in \mathbb{N}$: $D_i^j(1,V) = (c_i^V) \in C^V$ and $D_j^l(1/2,2W) = (c_j^W, c_l^W) \in C^{2W}$. Let $F = D_i^j(1,V) \circ D_j^l(1/2,2W)$, then there exist $\lambda \in int(\Delta^2)$ such that

$$P_{V+2W}(F) = \lambda P_{\max\{V,W\}}(D_i^j(1)) + (1-\lambda) P_{\max\{2V,2W\}}(D_j^l(1/2)).$$

Part (i) states that the belief induced by an anchored chain is a weighted average of the beliefs induced by the related (replicated) anchored databases. The very similar and almost disjoint databases allow a simple averaging, which keeps the normative appealing spirit of the Concatenation Axiom. The databases share only one identical precise piece of information (the anchor case in kL-many observations). Hence its induced identical estimate is contained in all their induced beliefs. This allows to "neglect" its impact for the determination of the average weights. Since in addition the mutually different non-anchor cases appear only in one of the anchored databases, there emerge no difficulties in (cognitively challenging (interwoven)) balancing of differently cautious estimations based on identical, but differently precise observations in various databases. Thus, the anchored-agent can basically determine the average weights based on judging the relative importance and relevance of the mutual different non-anchor cases.[20] In this way, an anchored agent can find the average weights in a very simple case by case comparison.

The particular (maximal adjusted) length of the related corresponding concatenating databases ensures that the most precise case in an anchored chain is captured objectively in the average of their induced beliefs. An cautious agent does not accept an average of beliefs induced by databases that evoke less precise estimations regarding this information, since this would directly imply a distortion of the precision of the belief induced by the anchored chain.

We **continue the Examples** to illustrate the anchored Concatenation Axiom.
(i) continued: The belief induced by $F = (c_1^{60}, c_2^{10}, c_3^{20})$ is an average of the beliefs

[20] Of course, the estimation based on the anchor case is not contained in the same weight in each belief, but this is directly adjusted for by assigning the desired weights to the beliefs induced by the particular databases.

induced by $D_1^2(\frac{2}{3}, 90) = (c_1^{60}, c_2^{30})$ and $D_1^3(\frac{2}{3}, 90) = (c_1^{60}, c_2^{30})$. Since by construction the estimate based on the anchor case c_1 is identically contained in all beliefs, the doctor can neglect its influence of the anchor case for determining the average weight. Hence the weights can be easily determined by just comparing the relative (a frequency-weighted) importance of c_2^{30} and c_3^{30} for evaluating the remaining parts of the anchored chain (c_2^{10}, c_3^{20}). Intuitively, the discrepancies in the precisions for c_2 and c_3 are negligible, since the focus lies predominantly on capturing perfectly the impact of the most precise case c_1. This is directly achieved in this example, since the most precise case c_1 is also the anchor case, and hence appears equally often in all databases.
(ii) continued: The belief induced by $F = (c_1^{30}, c_2^{40}, c_3^{80})$ is an average of the beliefs induced by $D_1^2(\frac{1}{5}, 100) = (c_1^{20}, c_2^{80})$ and $D_1^3(\frac{1}{5}, 100) = (c_1^{20}, c_3^{80})$. Again, the anchor case c_1 appears equally in both replicated anchored databases, i.e. c_1^{20}, which enables to neglect it for finding the average weight. The agent only needs to weight the amount and relevance of c_2^{80} and c_3^{80} for judging (c_2^{40}, c_3^{80}). Thereby it is essential that the most precise case (in the anchored chain F) c_3 is captured perfectly. The discrepancies in the objective precisions for the cases c_1 (20 in D_1^j vs 30 in F) and c_2 are negligible, since the focus lies on capturing the most precise information c_3 objectively.

A straightforward consequence of the agent's focus on the most precise case and the specific structure of the anchored databases is that the estimations based on minor precise pieces of information are not made in their objective precision, but in the precision of the most precise case. This can be seen directly by the recursive application of the anchored Concatenation Axiom, i.e. $P_T(D) = \sum_{c \in D} \lambda_c P^c_{\max_c f_D(c) \cdot T}$ for appropriate $\lambda_c \in (0, 1)$.
Since this structure (obviously) reappears in our representation theorem, we will postpone the discussion of its plausibility and reasonability to Section 2.5.1.

Part (ii) of the anchored Concatenation Axiom describes just a restriction to the very intuitive requirement that a belief induced by a combination of two disjoint databases should lie in between the induced beliefs of the disjoint databases separately. Averaging beliefs based on disjoint database are at the heart of the axiom, since there are no interdependencies between the information (and their precision) in the different databases. Furthermore, the axiom requires averaging only for very specific databases, i.e. a database consisting only of observations of one case and a database containing (potentially different, but) equally many observations of two other cases. The main assumption concerns the condition on the lengths that is again driven by the agent's focus on the most precise cases, in the sense that the most precise information should be captured equally in all averaging beliefs induced by the respective databases.

Constant Similarity Axiom (for maximal anchored version)
(i) Let $F \in C^T$ be an anchored chain of $(D_i^j(k,T_j))_{j\neq i\leq m}$, i.e. $F = \circ_{j\neq i}^m D_i^j(k,T_j)$ and let $L \in \mathbb{N}$ be the corresponding adjusted (maximal) length, i.e. $L = L(k,(T_i^j)_{j\neq i})$. If there exist some vector $\lambda \in \Delta^m$, (where $\lambda_j = 0$ for all $j \leq m$ such that $T_j = 0$) such that for some $Z \in \mathbb{N}$ the following equation holds:

$$P_{ZT}(F^Z) = \sum_{j\neq i}^m \lambda_j P_{LZ}(D_i^j(k)),$$

then this equation holds for all $Z \in \mathbb{N}$.
(ii) Let for three distinct $i,j,l \leq m$ and any $V,W \in \mathbb{N}$ $F = D_i^j(1,V) \circ D_j^l(1/2,2W)$. If there exist $\lambda \in int(\Delta^2)$ for some $Z \in \mathbb{N}$ such that the following equation holds:

$$P_{Z(V+2W)}(F^Z) = \lambda P_{Z\max\{V,W\}}(D_i^j(1)) + (1-\lambda)P_{Z\max\{2V,2W\}}(D_j^l(1/2)),$$

then this equation holds for all $Z \in \mathbb{N}$.

The average weights $(\lambda_j)_j$ are related to (frequency weighted) relevance or similarity weights, which could in principle depend on the length of the database. However, the Constant Similarity Axiom allows to identify the similarity function independent of the content and the size of the databases. It is reasonable to require a length-independent similarity if the similarity values are determined by some primitive or prior knowledge about the environment, which can not be learned, influenced or based on the information contained in the database. Of course, the axiom is questionable if an agent uses the databases not solely for evaluation of the outcome distribution, but also to learn something about structural (causal) relationship of particular features in the cases. However, the approach taken in this work excludes such deductive reasoning in deriving and updating the similarities from underlying databases.[21]

[21]For deductive reasoning see also Section 2.5.4 about the relationship to statistical methods.

2.5 Representation result of cautious belief formation

2.5.1 Representation Theorem with maximal anchored axioms

Theorem 2.1

Let there be given a function $P : C^* \to \Delta(R)$. Let P_T be the restriction of P to C^T for $T \in \mathbb{N}$. Let P satisfies the Learning and the Diversity Axiom.
Then the following are equivalent.
(i) The function P satisfies the Invariance, the maximal anchored Concatenation and the Constant Similarity Axiom
(ii) There exists for each $(T, c) \in \mathbb{N} \times C$ a unique $P_T^c \in \Delta(R)$, and a unique -up to multiplication by a strictly positive number- strictly positive function $s : C \to \mathbb{R}_+$, such that for all T and any $D \in C^T$:

$$P_T(D) = \frac{\sum_{c \in D} s(c) f_D(c) P^c_{T^*_D}}{\sum_{c \in D} s(c) f_D(c)}, \tag{2.4}$$

where $T^*_D \in \mathbb{N}_+$ is defined by $T^*_D := T \cdot \max_{c \in C} f_D(c)$.

Rough sketch of the proof:

The necessity part is straightforward calculations. The sufficiency part follows the rough structure of the proof of BGSS and EG, but differs in the crucial arguments. The idea is to transform the framework from the space of databases to the space of frequency vectors that is structurally more tractable, i.e. the belief based on databases $P_T(D) = \frac{\sum_{c \in D} s(c) f_D(c) P^c_{T^*_D}}{\sum_{c \in D} s(c) f_D(c)}$ for $D \in C^T$ translates to frequency vectors $P_T(f) = \frac{\sum_{j \le m} s_j f(j) P^j_{T^*_f}}{\sum_{j \le m} s_j f(j)}$ for f that represents $D \in C^T$ via $f = f_D \in \Delta(C)$.
The essential part of the proof is to derive the similarity weights $(s_i)_{i \le m}$. This will be down inductively over $m = |C|$.
Step 1: Base case for the induction, i.e. $|C| = m = 3$, w.l.o.g. $C = \{c_1, c_2, c_3\}$, i.e. aim to find s_1, s_2, s_3.
Step 1.1: A function P satisfying the anchored Concatenation and Constant Similarity Axiom can be written as $P_T(f) = \sum_{j \le m} \lambda_i P_{T^*_f}(f^j)$ for some appropriate $\lambda \in \Delta^3$ and j-th unit vectors $(f^j)_j$. Plugging $\overline{f} := \frac{1}{3}(f^1 + f^2 + f^3)$ into this equation and in the representation given by the theorem yields (with the Diversity Axiom) the similarity values in terms of $\lambda \in int(\Delta^3)$. Using these derived similarity values, we will define $P_T^s(f) := \frac{\sum_{j \le 3} s_j f(j) P_{T^*_f}(f^j)}{\sum_{j \le 3} s_j f(j)}$ for all $f \in \Delta(C)$ and $T \in \mathbb{N}$. The aim is to show that

$$P_T(f) = P_T^s(f) \tag{2.5}$$

for all $f \in \Delta(C)$. Of course $(f^j)_{j\leq 3}$ and $\overline{f}$ do already satisfy this equality.

Step 1.2: Partition the simplex $\Delta(C)$ into so called simplicial triangles recursively and show that $P_T(f) = P_T^s(f)$ for all simplicial points. Simplicial partitions are defined as follows (see Figure 2.1 in Section 2.9.6). The 0-th simplicial partition is exactly the simplex, i.e. consists of vertices $q_0^j \in \Delta(C)$, which are exactly the unit vectors f^j for $j = 1, 2, 3$. The first simplicial partition of $\Delta(C)$ is a partition to four triangles separated by the segments connecting the middle points between the two of the three unit frequency vectors, i.e. $q_1^1 := (\frac{1}{2}f^1 + \frac{1}{2}f^2)$, $q_1^2 := (\frac{1}{2}f^2 + \frac{1}{2}f^3)$ and $q_1^3 := (\frac{1}{2}f^3 + \frac{1}{2}f^1)$. The second simplicial partition is obtained by similarly partitioning each of the four triangles to four smaller triangles, and the l-th simplicial partition is defined recursively. The simplicial points of the l-th simplicial partition are all the vertices of triangles of this partition.
Based on the fact that the points of the 0-th simplicial partition and $\overline{f}$ already satisfy (2.5), the idea of the proof is to find a recursive procedure to cover all points of any l-th simplicial partition.
The underlying intuition and reason that allows such a procedure is the following observation. For any four specifically structured frequency vectors (namely anchored frequency vectors or the pairs of vectors appearing in the anchored Concatenation Axiom (ii)) that fulfill equation (2.5), also the intersection of the lines between two of these vectors satisfies (2.5). The crucial step in the proof is to apply his fact in a appropriate recursive way. Our "algorithm" -which is different than the one in BGSS and EG- ensures that all simplicial points of any l-th partition satisfies equation (2.5).

Step 1.3: Show that $P_T(f) = P_T^s(f)$ for all frequency vectors $f \in \Delta(C)$ for $|C| = 3$. One can show that the beliefs P and P^s induced by a sequence of simplicial points that approximates f converge to the belief of P and P^s induced by its limit f. Thereby the Learning Axiom ensures the existence of such a limit belief. Thus the base case for the induction is shown and we need to proceed with

Step 2: $|C| = m > 3$.
Step 2.1: Determination of $s_1, .., s_m$.
One can show that the similarity weights derived in Step 1 for any set of basic cases $C = \{c^i, c^j, c^k\}$ are independent of the triplet $\{i, j, k\}$ and thus we can define $P_T^s(f) := \frac{\sum_{j\leq m} s_j f(j) P_{T_f^*}(f^j)}{\sum_{j\leq m} s_j f(j)}$ for all $f \in \Delta(C)$ and $T \in \mathbb{N}$.

Step 2.2: Show $P_T(f) = P_T^s(f)$ for all $f \in \Delta(C)$ via induction over m and using Step 1 ($m = 3$) as base case.
Any $f \in \Delta(C)$ can be written in m different ways as anchored chain with respect to m different anchors. Applying the maximal anchored Concatenation Axiom to these m different anchored chains delivers m-many different hyperplanes. Each of these hyperplanes is spanned by $m-1$- many beliefs induced by anchored frequency vectors, for which already the equation (2.5) holds by induction assumption. Since both $P_T^s(f)$ and $P_T(f)$ are elements of all these m many hyperplanes and one can show that their intersection is unique, we have also (2.5) for all $f \in \Delta(C)$, which concludes the proof.

Although the rough structure of the proof is similar to BGSS and EG, our proof needs different arguments to complete the different parts of the proof. In particular the anchored version version of the Concatenation Axiom requires a different recursive approach/algorithm to cover all simplicial points (Step 1.2), which is the crucial step of the proof. Namely, in BGSS the combination of any databases or frequency vector is allowed. Also in EG the combination of any frequency vectors (by taking care about the lengths and the Constant Similarity Axiom) is basically possible. However, we can combine only specific anchored databases or frequency vectors. Also the induction step (Step 2.2) requires a different reasoning as in BGSS or EG. As in EG, the Constant Similarity Axiom is an essential ingredient to facilitate the proof.

Interpretation of Theorem

The induced belief is a frequency and similarity weighted average of the estimations based on past observations. All estimations $(P^c_{\max_c f_D(c)T})_{c\in D}$ are made according to the level of cautiousness implied by the most precise case. This means that only the most precise piece of information is captured objectively in its estimation. Hence, the axiomatized belief formation does not achieve a perfectly objective representation (as mentioned in (2.2)) without any imagination effort. However, such a "perfect imagination-free representation " is impossible for a sufficiently rich Concatenation Axiom (see Section 2.7.3) and also carries some drawbacks (see the discussion after (2.2)). In any case, we are not concerned with imagining additional information to take into account objective precision.

In fact, in first place we are interested in capturing the perception of precision in form of the induced psychological effects on cautiousness and confidence. This is essential for small database containing relatively few information and is manifested in the way how estimations $P^{c}_{T^*_D}$ are made. From this perspective, the seemingly undesirable imagination in the axiomatized belief delivers the following intuitive and reasonable interpretation. The underlying intuition is that an agent does not adjust constantly her cautiousness and confidence attitude in response to each differently precise information she encounters in a database. Rather, once an agent has experienced a (extreme) cautiousness and confidence feelings while estimating based on objectively available information, she keeps, adopts and transmits her developed feeling to other estimation situations. A fixed level of cautiousness according to which all estimates are made can be interpreted as an gained attitude regarding cautiousness or as a learned skill or ability to confidently estimate sufficiently cautiously. In this way, it is a sustainable reference or state of mind, which does not vanish and change for each new estimation.

For instance, an agent gained a feeling of cautiousness in the spirit of eliminating un-reasonable outcomes. Suppose she feels confident and considers herself cautious enough to assign only a small probability ϵ to non observed outcomes $\tilde{r} \neq r$ in estimating based on $c = (x, r)^L$. Separately, her estimation induced by $c' = (x', r')^T$ with $T > L$ assigns a slightly lower likelihood $\epsilon' < \epsilon$ to the not observed outcome $\tilde{r} \neq r'$ according to her lower cautiousness and higher confidence. Assume now that in the past she has only estimated according to a precision level lower than L and someone tells her that $T - L$ pieces of information c were lost and she should better estimate according to T many imagined observations. Without having experienced estimating according to higher precision T (i.e. how far she can narrow down the estimation) and being unable to imagine how she would feel if this information would be objectively available, she might still stick to her already made estimation based on objective information c^L. However, if the agent would have estimated based on case $(c')^T$ in the past, then she has experienced her feeling of estimating according to the objective precision in $(c')^T$ and might adopt and apply the "learned" procedure how to eliminate and assign the likelihoods confidently for c^T without concerns about being too in-cautious.

The most intuitive choices for adopting a specific attitude towards cautiousness are the two extreme situations, i.e. the least and most cautious (and confident) experiences. The most precise case might come directly to her mind, because it has been observed most frequently in the database and induces an attitude of (least) cautiousness (and highest confidence) that is the basis for all estimation. In some sense the most confident and least cautious feeling outshines and distracts from any other more cautious perceptions. In contrast, the least precise information might

intimidate or scare an agent and leaves a very cautious impression. She cannot be persuaded to leave her skeptical mood for a less cautious attitude that might be more appropriate for the remaining more accurate information. In our representation we focus on the optimistic view, i.e. our agent estimates according to the confidence and cautiousness gained and experienced by estimating the most precise information in the database.
In this way it is reasonable and natural to interpret the imagination of additional information in the sense of estimating according to an experienced cautiousness level or as gained skill to estimate cautiously.[22]

Differences in imagined information and its imagined perception

In fact, the imagination of further additional information or more precise cases is not the cognitive difficult or challenging part in estimating based on imagined information. Think about our doctor, who just needs to imagine that the same patient enters her office again and shows the same outcome after being treated identically. Hence, the difficult part is to imagine the "correct" feeling, which would be induced by objective precision, but which is actually only existing in imagined precision. Put differently, usually the implied perception of imagined (non existing) precision differs from the perception based on objective precision. The beliefs (EG and partly our) require that agents are able to ignore this difference, which might be fine if agents have experienced already a situation in which they actually estimated according to that objective precision and know her induced perception of that precision (as in our work). However, if an agent has never experienced such a situation before (as in EG), the requirement to imagine her feeling "correctly" (i.e. ignore the differences) is cognitively challenging and psychologically confusing and can be interpreted as intentionally lying to yourself without noticing. Does our doctor judge the treatment less cautiously after adding an imagined patient to her record?

2.5.2 Comparison to related belief representations

The initial motivation of EG and our paper is to modify the Concatenation Axiom of BGSS to capture variations in the precision of data. A related and implied issue concerns the way how an agent is capable to deal with the problem of combining

[22]From that perspective, our representation is even more convincing than the perfectly objective imagination-free representation (2.2), in which the cautiousness and confidence is altered for each case, putting the agent in different moods of cautiousness and confidence for each piece of information.

beliefs that might be based on identical, but different precise information and thus contain induced differently cautious estimates.

BGSS, EG and our work share the property that eventually the estimations involved in the final representation of a belief are subject to an unique level of precision.[23] By that, technically speaking the aggregation of different precise information is eventually not an issue. However, from an interpretational perspective, there are important differences in the motivation and reasonability of the corresponding Concatenation Axioms.

Consider for example the database $D = (c_1^3, c_2^4, c_3^2)$ for which a purely objective agent forms a belief according to $P(D) \in conv(\{P(c_1^3), P(c_2^4), P(c_3^2)\})$. In BGSS, the induced belief is given by $P(D) \in conv(\{P(c_1), P(c_2), P(c_3)\})$, which neglects precision and cautiousness completely. EG gives $P(D) \in conv(\{P(c_1^9), P(c_2^9), P(c_3^9)\})$, where no involved estimation is made according to its objective precision. Apart from the (unproblematic) imagination of additional pieces of observation for all cases, the main problematic point is the imagination on how this imagined precision is perceived, since the estimation is based on a never (not yet) experienced cautiousness level 9 (see also the discussion above). In our paper, the belief would be based on the most precise information, i.e. $P(D) \in conv(\{P(c_1^4), P(c_2^4), P(c_3^4)\})$, which also would require some (unproblematic) imagination of additional observations with respect to objective precision. However, the perception of this precision needs not to be imagined, since the agent estimates according to an already experienced precision and cautiousness level 4 (experienced for c_2).

Arad and Gayer (2012) analyze beliefs based on datasets containing imprecise pieces of information in the sense that "it is not entirely clear what occurred in them". Roughly speaking, their approach models this sort of imprecision (ambiguity) by assuming subjective capacities. The rough relationship to the approaches discussed above is that these capacities would play the role of the probabilistic estimations occurring in the axiomatized representations of BGSS, EG and ours.

2.5.3 Remarks on the similarity function

One could be tempted to perceive and interpret the belief formation approaches as a translation of the question from which probability to assign to which similarity to employ. This is not completely misleading since the axiomatizations do not provide help in choosing the similarity function. This problem occurs in a similar spirit for the choice of a prior in the Bayesian approach. In the axiomatizations the similarity function is derived from presumably observable probability assignments given

[23] in BGSS: P_∞^c for all $D \in C^*$, in EG: P_T^c for all $D \in C^T$ and here $P_{max_c f_D(c) T}$ for all $D \in C^T$.

various databases. Fortunately, the similarity values need not satisfy any particular properties (even no symmetry) and hence can be derived also objectively or empirically. For example, Gilboa et al. (2006) estimate an empirical similarity function from the data by asking which similarity function best explains the observed data in a similarity-weighted frequency formula. Billot et al. (2004) axiomatized an exponential similarity function. Moreover, assigning similarities appears to be cognitively easier than stating explicit probabilities and many models in the psychology and computer-science literature deal with determination of similarity measures (e.g. Tversky (1977), Schank (1986), Heit, Heit and Rubinstein (1994), Goldstone and Son (2005)).

2.5.4 Remarks on relationship to statistical methods

In the introduction we mentioned already the relationship between the axiomatic approaches to belief formation in the data-based information structure and statistical approaches like inferences. In this section we want to discuss shortly similarities and differences to existing statistical methods. Obviously, the versions of the Concatenation Axioms and the derived representations satisfies the following special cases of frequentism. For $s(x_i, x_t) = 1$, our belief formation coincides with the simple average or frequentist approach if we identify with P^c a Dirac measure on the actually observed outcome. However, the conditional frequentist cannot be covered since the corresponding $s(x_i, x_t) = 1_{\{x_t = x_i\}}(x_i)$ is not strictly positive (as required), but Bleile (2014b) (or Chapter 3) offers a modification that captures it. Gilboa et al. (2010, 2011) and EG show the compatibility with other statistical methods, like kernel estimation and classification (e.g. assign x to either class a or b: define $s(a, (x_c, a_c)) = k(x_c, x)1_{\{a=a_c\}}$ using a kernel function k). As discussed in more detail in Gilboa et al. (2010) p. 16f, the framework can be also employed in contexts, where the observations (e.g. cases) and the prediction (e.g. possible theories) are structurally disjoint. For instance ranking theories by log likelihood methods $s(t, c) = log(p(c|t))$ is also possible where t represent a theory and $p(c|t)$ denotes the conditional likelihood of case c if theory t is true.

However, the main difference to statistical inference is that the axiomatic approaches are concerned with inductive reasoning and do not allow for deductive reasoning, which is the issue of traditional statistical regression approaches. Let there be a database consisting of observation $D = ((x_i, r_i)_{i \leq n})$ and a new problem x_t. A regression approach would try to learn the (empirical) similarity weights $(s(x_i, x_t))_i$ that best explains the database by best fitting an estimate of r_j for all $j \leq n$ and $r_j^s = \frac{\sum_{i \neq j} s(x_i, x_j) r_i}{\sum_i s(x_i, x_j)}$ (see also Gilboa et al. (2006)). Hence in a statistical regression context the weights s are deduced endogenously via the observed data and are updated with new observation, i.e. the weights would be database dependent. Put differently, linear regression analysis (and empirical similarities) use deductive reasoning to derive the weights and then apply them inductively to infer the prediction. In contrast, the Constant Similarity (and the Concatenation) Axiom requires that the weights are fixed and database independent, i.e. there is no updating or learning of the weights.

However, the axiomatization of a belief formation (in close relationship to statistical methods) is still meaning- and insightful, since it allows to inspect, how plausible, consistent and sensible (in the sense of normative appealing axioms) asymptotic statistical methods are also for small database and its implied precision related concerns. From this perspective, axiomatizations suitable for small databases (as done here) play an important role in order to find a sound foundation of statistical methods in non-asymptotic contexts.

Basically, our paper tries to capture exactly the environment of small databases, whereas the framework of implied instantaneously perfect learning of BGSS and EG are rather embeddable in an asymptotic setup. Interpreting EG in asymptotic terms requires some explanation, since EG is intended to deal with different ("non-asymptotic") precisions. From our point of view, when dealing with small and less precise databases, not only the objective precision is important for the belief formation, but also an agent's feeling regarding the precision (e.g. in terms of its implied cautiousness). In this context, EG's axiomatized agent is able to imagine the true feeling induced by any imagined level of precision, as if she actually observed the precision objectively and experienced that feeling. Endowed with this skill (of "perfect" imagination of feelings), it appears un-intuitive that an agent limits herself to use it only until an arbitrary pre-specified level of precision (given by the entire amount of observations in the database, i.e. unrelated to any former experience). Rather one would expect her to continue (unboundedly) with the imaginations such that she will estimate according to an imagined full precision level P^c_∞ for all $c \in D$ and does not need to deal with imprecisions at all. Thus, from our perspective the approaches of BGSS and EG coincide for an interpretation in terms of precision related perception and feelings.

2.6 Conclusion

Chapter 2 deals with the question how agents form beliefs explicitly in an environment with limited, heterogenous and differently precise information that cannot

be condensed into a widely used (perfect) state space a la Savage. We axiomatize a belief formation that can be interpreted as a generalized subjective frequentist approach that incorporates subjective perceptions regarding the relevance and precision of the information in the database. We identify increasing precision of information by additionally observed pieces of confirming information.
Our work is based on the axiomatization of a belief in BGSS that neglects the potential impacts of differently precise information. Thereby, their belief formation is most suitable for sufficiently large databases and less reasonable for small databases, which are captured by our approach. Their belief formation implies that an agent is able to perfectly learn from observations in a very objective and instantaneously way, without displaying any sense of cautiousness and concerns about being potentially mistaken. Our axiomatized cautious belief focusses on precision related cautiousness and confidence in the predictions. The different versions of the main Concatenation Axiom in the approaches of BGSS, EG and ours describe the relationships between databases and their induced beliefs. In the context of caring for precisions in a cautious belief formation an agent following the Concatenation Axiom of BGSS and EG's version would be faced by immense cognitive problems to handle and compare differently precise pieces of information contained in different databases. Our modification and restriction of the axiom takes into account these precision related cognitive problems in describing the relationships. This is achieved by requiring that agents only need to be capable to determine the relationship between databases and their induced beliefs for specifically structured (almost disjoint) databases that allow an cognitively easy comparison (without precision and cautiousness concerns). Moreover, it states that an agent controls for precision and its perceptional impacts in a cautious belief by capturing the most precise (and hence reliable) information objectively in its induced belief.
The resulting cautious belief is a weighted sum of cautious estimates induced by past observed information. The weights are determined by frequencies of the observed cases and their similarities with the problem under consideration. The induced estimates depend on a cautiousness level implied by the most precise case, which can be interpreted as the appropriate (gained) attitude regarding cautiousness in this database.

2.7 Further remarks and considerations

2.7.1 Incompatible objective belief formation

An objective belief without any imagination effort for each case reads

$$P_T(D) = \frac{\sum_{j\leq C} s(c_j) f_D(c_j) P^{c_j}_{f_D(c_j)\cdot T}}{\sum_{j\leq C} s(c_j) f_D(c_j)}.$$

Is there a modification of the Concatenation Axiom that is necessary for an objective representation and which database are admissible?

(Specifically) Modified Version of the Concatenation Axiom
For two databases $D \in \mathcal{D}^{T_1}$ and $E \in \mathcal{D}^{T_2}$, $T := T_1 + T_2$ and two numbers $G, H \in \mathbb{N}_+$ such that $f_D \cdot G \in \mathbb{N}^m$ and $f_E \cdot H \in \mathbb{N}^m$, there exists a $\lambda \in (0,1)$ such that

$$P_T(D \circ E) = \lambda P_L(D^G) + (1-\lambda) P_H(E^H).$$

Applying the objective belief to the modified Concatenation Axiom yields:

$$\begin{aligned}\frac{\sum_{j\leq m} s(c_j) f_{D\circ E}(c_j) P^{c_j}_{f_D(c_j)T_1 + f_E(c_j)T_2}}{\sum_{j\leq m} s(c_j) f_{D\circ E}(c_j)} &= \lambda \frac{\sum_{j\leq m} s(c_j) f_D(c_j) P^{c_j}_{f_D(c_j)L}}{\sum_{j\leq m} s(c_j) f_D(c_j)} \\ &\quad +(1-\lambda) \frac{\sum_{j\leq m} s(c_j) f_E(c_j) P^{c_j}_{f_E(c_j)H}}{\sum_{j\leq m} s(c_j) f_E(c_j)}\end{aligned}$$

We do not consider a law of dynamics for the probabilities $P^{c_j}_T$ (which is also not reasonable), e.g. like some function Y of $P^{c_j}_T = Y(P^{c_j}_L, P^{c_j}_H)$. Thus, we directly need to equalize the precision level for the estimations for each single case, i.e. for all $j \leq m$

Situation 1
Assume that there exists $c \in D$ such that $f_D(c) = 0$, then we get directly from

$$f_D(c)T_1 + f_E(D)(c)T_2 = f_E(c)H$$

that $H = T_2$. Assume that there is another $c \in D \cap E$, then we have that $f_D(c)T_1 + f_E(c)T_2 = f_E(c)T_2 = f_D(c)G$, which is impossible. Hence, a databases D with $f_D(c) = 0$ for some $c \in D$ allows only concatenations of disjoint databases.

Situation 2
We consider only databases that share the same support, i.e. $f_D(c) > 0$ iff $f_E(c) > 0$. Thus we need to have $f_D(c)T_1 + f_D(c)T_2 = f_D(c)G = f_E(c)H$. Summing over all

$c \in D \cap E$ leads to $T_1 + T_2 = G = H$. This also implies that $f_D(c) = f_E(c)$ for all $c \in D \cap E$. Thus, the only non-disjoint concatenating database a modified Concatenation Axiom allows for are replicated identical databases, which is naturally true for all $\lambda \in (0, 1)$, i.e. $P_{T_1+T_2}(A) = \lambda P_{T_1+T_2}(A) + (1 - \lambda)P_{T_1+T_2}(A)$.

In sum, a objective belief satisfies a modified Concatenation Axiom only for concatenations of disjoint databases or replicated identical databases. However, it will become clear in the proof that a restriction to disjoint concatenating databases offers not sufficient structure to derive the desired objective belief.

2.7.2 Relationship to EG's axiom "Concatenation restricted to databases of equal length"

As mentioned at the end of Section 2.4.2, a Concatenation Axiom that allows for concatenations of any unrestricted databases faces immense compromising between differently precise (or cautious) estimates. This can only be avoided by assuming a common arbitrary level of precision according to which all cases are estimated, independent of the objective precision of the information. For each piece of information, literally agents need to imagine (or forget) sufficiently many observation of cases to reach an assumed artificial common level of precision. This ensures that no considerations and compromising regarding different precisions is required and allows an easy averaging based only on relative relevances of the concatenating databases. However, thereby an agent also needs to know a priori that she evaluates all information in an imagined precision and the beliefs contains only (imagined) equally precise and cautious estimations. Consequently, a version of a Concatenation Axiom that cares for precision and also applies to arbitrary non-disjoint concatenations accomplishes the averaging of differently precise information by explicitly assuming (consciously) away these differences. This is somehow problematic if one wants to take precision into account seriously.

Nevertheless, this discussion delivers an explanation and intuition for the (unexplained) statement in EG: " ... we modify the Concatenation Axiom of BGSS by restricting it to databases of equal length, i.e. thus controlling for the ambiguity resulting from insufficient amount of data ". Their restriction to equal lengths is ad hoc. However, technically one could argue for the equal length assumption by referring to the discussion above. An aggregation of differently precise information is only feasible if estimations are based on a common precision level, which is a consequence of the restrictions in their axiom. More detailed, their axiom demands that for a set of n databases of the same length T, that can be concatenated to a n-times replication of a database, a belief induced by this database (not the n-th replication) is a average of the beliefs induced by each of the n databases separately. Obviously, this implies for a appropriate set of concatenating databases (consisting only of a single case)[24] that the belief induced by the database - which underlie the n-th replication- is formed by an average of the beliefs induced by T-times observed cases, i.e. for some appropriate $(\lambda_c)_{c\in D} \in (0,1)$

$$P(D) = \sum_{c\in D} \lambda_c P(c^T).$$

Thus, the restriction to equal lengths implies directly that all contained estimation are based on this common level as well. In this way, EG is indirectly adopting the procedure discussed above.

However, as already discussed, from our perspective and motivation, equal lengths of database are not sufficient to control for precision of its contained information. Moreover, in the spirit of the above discussion, EG's restriction to equal lengths cannot be meaningful interpreted as controlling for imprecision, but more as an (implicit) proposal to employ the length of the entire databases as the common (imagined) precision level according to which all estimations are made.

2.7.3 Minimal anchored axiomatization

Instead of focussing on the most precise case in a database to determine the precision of its induced belief, we can also take the least precise case as the key determinant for the precision of a belief. This modification results in a minimal version of an anchored Concatenation and Constant Similarity Axiom and a corresponding extremely cautious belief formation.

Definition 2.4
Let $F \in C^T$ be an anchored chain of $(D_i^j(k,T_j))_{j\neq i\leq m}$.
*A length $M \in \mathbb{N}$ is called the **adjusted (minimal) length**, denoted $M(k,(T_j)_{j\neq i\leq m})$, if it is such that the number of observations of the least frequent case in an anchored chain $F \in C^T$ is identical to the number of observations of the least frequent case in the anchored databases $D_i^j(k,L)$ (for all $j \neq i$) (i.e. $\min_{c\in C} f_F(c)T = \min_{c\in C} f_{D_i^j(k,M)}(c)M$).*

Minimal Anchored Concatenation Axiom:
(i) Let $F \in C^T$ be an anchored chain of $(D_i^j(k,T_j))_{j\neq i\leq m}$, i.e. $F = \circ_{j\neq i}^m D_i^j(k,T_j)$ and let $M = M(k,(T_i^j)_{j\neq i}) \in \mathbb{N}$ be the corresponding adjusted (minimal) length. Then

[24] This is always possible, for example consider $D = (c_1^2, c_2, c_3^3) \in C^6$, then $D^6 = (c_1^6) \circ (c_1^6) \circ (c_2^6) \circ (c_3^6) \circ (c_3^6) \circ (c_3^6)$, which implies $P(D) = \lambda_1 P(c_1^6) + \lambda_2 P(c_2^6) + (1-\lambda_1-\lambda_2) P(c_3^6)$

there exists $\lambda \in \Delta^m$ (where $\lambda_j = 0$ for all $j \leq m$ s. th. $T_j = 0$), such that

$$P_T(F) = \sum_{j \neq i \leq m} \lambda_j P_M(D_i^j(k)).$$

(ii) Let for three distinct $i, j, l \leq m$ and any $V, W \in \mathbb{N}$ $F = D_i^j(1, V) \circ D_j^l(1/2, 2W)$ then there exist $\lambda \in int(\Delta^2)$:

$$P_{V+2W}(F) = \lambda P_{\min\{V,W\}}(D_i^j(1)) + (1-\lambda) P_{\min\{2V,2W\}}(D_j^l(1/2))$$

For an analogously adjusted Constant Similarity Axiom the resulting theorem reads:

Theorem 2.2
Let there be given a function $P : C^* \to \Delta(R)$. Let P_T be the restriction of P to C^T for $T \in \mathbb{N}_+$. Let P satisfies the Learning and the Diversity Axiom.
Then the following are equivalent:

(i) The function P satisfies the Invariance, the minimal anchored Concatenation and the (minimal) Constant Similarity Axiom.
(ii) There exists for each $(T, c) \in \mathbb{N} \times C$ a unique $P_T^c \in \Delta(R)$, and a unique -up to multiplication by a strictly positive number- strictly positive function $s : C \to \mathbb{R}_+$, such that for all T and any $D \in C^T$:

$$P_T(D) = \frac{\sum_{c \in D} s(c) f_D(c) P^c_{T_*^D}}{\sum_{c \in D} s(c) f_D(c)},$$

where $T_*^D \in \mathbb{N}_+$ is defined by $T_*^D := T \cdot \min_{c \in D} f_D(c)$.

Interpretational, this means that all estimations are based on the least precise information contained in the database and no information needs to be imagined. However, the focus on the least precise information results in neglecting and discarding many more precise pieces of information, by processing only until the level of least precision. A detailed interpretation in terms of perception of precision and an adoption of an implied attitude of extreme cautiousness can be found in the discussion after Theorem 2.1.

2.8 Proof of Theorem 2.1, necessity part

We need to show that the representation (2.4) satisfies the axioms. The Invariance Axiom is obviously met.

Maximal Anchored Concatenation Axiom, part (i):

Let $D \in C^T$ be a chain of $D_i^j(k, T_j) = (c_j^{(1-k)T_j}, c_i^{kT_j})$ for all $j \neq i \leq m$ and $T := \sum_{j\neq i} T_j$, i.e.
$D = \circ_{j\neq i} D_i = (c_1^{(1-k)T_1}, c_2^{(1-k)T_2}, ..., c_{i-1}^{(1-k)T_{i-1}}, c_i^{kT}, c_{i+1}^{(1-k)T_{i+1}}, ..., c_m^{(1-k)T_m})$.
Let $L = L(k, (T_j)_{j\neq i})$ be the corresponding adjusted length. Hence, we have
$f_D = (\frac{(1-k)T_1}{T}, \frac{(1-k)T_2}{T}, ..., \frac{(1-k)T_{i-1}}{T}, k, \frac{(1-k)T_{i+1}}{T}, ..., \frac{(1-k)T_m}{T})^t$ and
$f_{D_i^j(k,ZT_j)} = (0, ..., 0, (1-k), 0, .., 0, k, 0, ..., 0)^t$ for any $Z \in \mathbb{N}$.

Observe that $f_{D_i^j(k,T_j)} = f_{D_i^j(k,ZT_j)}$and hence we will abbreviate $f_{D_i^j(k,ZT_j)}$ by $f_{D_i^j(k)}$.
Let $T_D^* := T \max_{c\in D} f_D(c)$, then:

$$
\begin{aligned}
P_T(D) &= \frac{\sum_{c\in C} s(c) f_D(c) P^c_{T_D^*}}{\sum_{c\in C} s(c) f_D(c)} \\
&= \frac{1}{\sum_{c\in C} s(c) f_D(c)} \cdot \Big(\sum_{j\neq i} s(c_j) \frac{(1-k)T_j}{T} P^{c_j}_{T_D^*} + s(c_i) \frac{\sum_{j\neq i} kT_j}{T} P^{c_i}_{T_D^*} \Big) \\
&= \frac{1}{\sum_{c\in C} s(c) f_D(c)} \cdot \Big(\sum_{j\neq i} \Big[s(c_j) \frac{(1-k)T_j}{T} P^{c_j}_{T_D^*} + s(c_i) \frac{kT_j}{T} P^{c_i}_{T_D^*} \Big] \Big) \\
&= \frac{1}{\sum_{c\in C} s(c) f_D(c)} \cdot \Big(\sum_{j\neq i} \frac{T_j}{T} \sum_{c\in C} s(c) f_{D_i^j(k)}(c) P^c_{T_D^*} \Big[\frac{\sum_{c\in C} s(c) f_{D_i^j(k)}(c)}{\sum_{c\in C} s(c) f_{D_i^j(k)}(c)} \Big] \Big)
\end{aligned}
$$

To proceed, we need to specify $T_D^* = \max_{c\in C} f_D(c) \cdot T$, which is by definition of the adjusted length L exactly equal to $\max_{c\in C} f_{D_i^j}(c) L$. Observe that $(*) \sum_{c\in C} s(c) f_D(c) = \sum_{j\neq i} \frac{T_j}{T} \sum_{c\in C} s(c) f_{D_i^j(k)}(c)$, hence

$$
\begin{aligned}
P_T(D) &= \frac{1}{\sum_{c\in C} s(c) f_D(c)} \cdot \Big(\sum_{j\neq i} \frac{T_j}{T} \sum_{c\in C} s(c) f_{D_i^j(k)}(c) \frac{\sum_{c\in C} s(c) f_{D_i^j(k)}(c) P^c_{\max_{c\in C} f_{D_i^j}(c) L}}{\sum_{c\in C} s(c) f_{D_i^j(k)}} \Big) \\
&\overset{(*)}{=} \frac{1}{\sum_{j\neq i} \frac{T_j}{T} \sum_{c\in C} s(c) f_{D_i^j(k)}(c)} \cdot \Big(\sum_{j\neq i} \frac{T_j}{T} \sum_{c\in C} s(c) f_{D_i^j(k)}(c) P_L(D_i^j(k, L)) \Big) \\
&= \sum_{j\neq i} \lambda_j P_L(D_i^j(k, L))
\end{aligned}
$$

From the last equation we get for all $j \neq i \leq m$

$$
\lambda_j = \frac{\frac{T_j}{T} \sum_{c\in C} s(c) f_{D_i^j(k)}(c)}{\sum_{j\neq i} \frac{T_j}{T} \sum_{c\in C} s(c) f_{D_i^j(k)}(c)} \tag{2.6}
$$

Hence the first part of the anchored Concatenation Axiom is satisfied.

For the **Maximal Anchored Concatenation Axiom, part (ii)**:
let w.l.o.g. $D_i^j(1,T) = D_1^2(1,T) = (c_1^T) \in C^T$ and $D_j^l(1/2,2W) = D_2^3(1/2,2W) = (c_2^W, c_3^W) \in C^{2W}$, then we need to show:

$$\begin{aligned} P_{T+2W}(D_1^2(1,T) \circ D_2^3(1/2,2W)) = \lambda P_{\max\{T,W\}}(D_1^2(1,\max\{T,W\})) \\ + (1-\lambda)P_{\max\{2T,2W\}}(D_2^3(1/2,2\max\{T,L\})) \end{aligned}$$

We have $P_{\max\{T,W\}}(D_1^2(1,\max\{T,W\}) = P^{c_1}_{\max\{T,W\}}$
Since P satisfies the maximal anchored Concatenation Axiom part (i) we have for

$$\begin{aligned} D_2^3(1/2,2\max\{T,W\}) &= D_1^2(0,\max\{T,W\}) \circ D_1^3(0,\max\{T,W\}) \\ &= (c_2)^{\max\{T,W\}} \circ (c_3)^{\max\{T,W\}} \end{aligned}$$

with the adjusted length L such that $\frac{1}{2}2\max\{T,W\} = L$, i.e.

$$L(0,\max\{T,W\},\max\{T,W\}) = \max\{T,W\}$$

the existence of some $\lambda \in (0,1)$ such that

$$P_{\max\{2T,2W\}}(D_2^3(1/2,2\max\{T,W\})) = \lambda P^{c_2}_{\max\{T,W\}} + (1-\lambda)P^{c_3}_{\max\{T,W\}}$$

Using this, we get for the maximal anchored Concatenation Axiom (ii) the following representation for some $\lambda \in \Delta^3$:

$$P_{T+2W}(D_1^2(1,T) \circ D_2^3(1/2,2W)) = \sum_{i=1}^{3} \lambda_i P^{c_i}_{\max\{T,W\}}$$

But this is obviously satisfied by the representation (2.4) in Theorem 2.1, since for the frequency vector $f_{D_1^2(1,T)\circ D_2^3(1/2,2W)} = (\frac{T}{T+2W}, \frac{W}{T+2W}, \frac{W}{T+2W}, 0, ..., 0)^t$, we have $\max_{c\in C} f_{D_1^2(1,T)\circ D_2^3(1/2,2L)}(c)(T+2W) = \max\{T,W\}$, and hence $P_{T+2W}(D_1^2(1,T) \circ D_2^3(1/2,2W)) = \frac{\sum_{i=1}^3 s(c_i) f_{D_1^2(1,T)\circ D_2^3(1/2,2W)}(i) P^{c_i}_{\max\{T,W\}}}{\sum_{i=1}^3 s(c_i) f_{D_1^2(1,T)\circ D_2^3(1/2,2W)}(c_i)}$, i.e. for $i = 1,2,3$

$$\lambda_i = \frac{s(c_1) f_{D_1^2(1,T)\circ D_2^3(1/2,2W)}(c_i) P^{c_1}_{\max\{T,W\}}}{\sum_{i=1}^3 s(c_i) f_{D_1^2(1,T)\circ D_2^3(1/2,2W)}(c_i)} \tag{2.7}$$

Hence the (ii)-part of the maximal anchored Concatenation Axiom is also satisfied.

The **Constant Similarity Axiom** is also satisfied, which can be shown by adopting the above proof for the anchored Concatenation Axiom.
Replacing $D = \circ_{j\neq i}^{m} D_i^j(k, T_j)$ by $D^Z = \circ_{j\neq i}^{m} D_i^j(k, ZT_j)$ (where $\sum_{j\neq i}^{m} T_j = T$) in the proof of the maximal anchored Concatenation Axiom part (i) and transforming equation (2.6) delivers the existence of some $\lambda_j(Z) \in (0,1)$ such that

$$\lambda_j(Z) = \frac{\frac{ZT_j}{ZT} \sum_{c\in C} s(c) f_{D_i^j(k,ZT_j)}(c)}{\sum_{j\neq i} \frac{ZT_j}{ZT} \sum_{c\in C} s(c) f_{D_i^j(k,ZT)}(c)} = \lambda_j,$$

where in the last equation $f_{D_i^j(k,ZT)}(c) = f_{D_i^j(k)}(c)$ is used.
For part (ii), analogous reasoning using equation (2.7) yields the desired result.

Therefore the Constant Similarity Axiom is satisfied, which completes the necessity part of the proof of Theorem 2.1.

□.

2.9 Proof of Theorem 2.1, sufficiency part

An essential step in the proof will be to identify databases with their frequency vectors of contained cases (and its length), which allows to exploit the more tractable structure of the space of frequencies on C (rather than the space of databases) and to adopt the approach taken in BGSS. By the Invariance Axiom each database $D \in \mathcal{C}^T$ can be identified (with respect to the induced belief formation) by a pair (f_D, T), where $f_D \in \Delta(C)$ represents a frequency vector of appearances of cases in the database D and T is the length of the database. Based on this, we will translate the database structure to a frequency framework.

2.9.1 General definitions for a frequency framework

The set of all frequency vectors on C is given by
$\Delta(C) := \{f = (f_1, ..., f_m)$ s. th. $f_i \in \mathbb{Q} \cap [0,1]$ for all $i \leq m$ and $\sum_{i\leq m} f_i = 1\}$
Without knowing the exact database D that a frequency vector $f \in \Delta(C)$ represents, the frequency vector can be linked to infinitely many databases D^Z for all $Z \in \mathbb{N}_+$. Hence one needs to link frequency and the length of the database for an "unique" (up to reordering) representation of a database.

The following set represents frequency vectors corresponding to databases $D \in C^T$:

$$\begin{aligned}\Delta_T(C) : &= \{f \in \Delta(C) \cap \mathbb{Q}^m,\ f(i) = \frac{l_i}{T},\ l_i \in \mathbb{N}_+,\ \sum_{i=1}^{m} l_i = T \text{ and} \\ &\quad \exists\, D \in C^T \text{ such that } f_D(i) = f(i) = l_i/T\}\end{aligned}$$

Observe that if $f \in \Delta_T(C)$, then $f \in \Delta_{TZ}(C)$ for all $Z \in \mathbb{N}_+$.
Since the set of cases C is fixed, we reduce the notational effort and will abbreviate $\Delta_T(C)$ by Δ_T, i.e. Δ_T denotes the set of all frequency vectors representing databases of length T and the set of all rational frequency vectors on C is denoted by Δ.
Hence by the **Invariance Axiom** each $D \in C^T$ can be represented by a $f \in \Delta_T$, where again $f(i) := f_D(c_i)$ denotes the frequency of case c_i for all $i \leq m$.

Definition 2.5
(i) For all $j \in \{1, 2, ..., m\}$ denote by f^j the j-th unit vector in $\mathbb{R}^m$, i.e. the frequency vector representing a database containing only cases $c_j \in C$, hence an extremal point in Δ, i.e. $f^j = (0, ..., 0, \underbrace{1}_{j-th}, 0, ..., 0)^t$

(ii) The frequency vector corresponding to the anchored database
$D_i^j(k,T) = (c_j^{(1-k)T}, c_i^{kT})$ *is given by*

$$f_{D_i^j(k,T)} = (0, ...0, \underbrace{(1-k)}_{j-th}, 0, .., 0, \underbrace{k}_{i-th}, 0, ..., 0)^t$$

Since $f_{D_i^j(k,W)} = f_{D_i^j(k,T)}$ for all T and W, the length is totally immaterial for the frequency vector and hence neglected from now on, i.e. the frequency vector corresponding to the anchored databases $D_i^j(k,T)$ for all $j \neq i \leq m$ is denoted for all T such that $kT \in \mathbb{N}$ by

$$f_i^j(k) := f_{D_i^j(k,T)}$$

Note that $f_i^j(k)$ is still the whole frequency vector, i.e. $f_i^j(k) \in \Delta$, whereas $f_i^j(k)(l)$ represents the l-th component of the vector and refers to the frequency of case c_l, i.e. $f_i^j(k)(l) \in [0,1] \cap \mathbb{Q}$.

2.9.2 Beliefs induced by frequency vectors

From now on we consider only beliefs P that satisfy the Invariance Axiom. Consequently, as mentioned above, we can transform beliefs defined on databases to beliefs defined on frequency vectors in the following way (remember that we fixed a problem $x \in X$ and skip it).

Definition 2.6
The beliefs $P : C^* \to \Delta(R)$ and its restriction $P_T : C^T \to \Delta(R)$ for all $T \in \mathbb{N}$ based on databases translates to corresponding beliefs based on frequency vectors in the following way:
(i) $P : \Delta \to \Delta(R)$ such that $P(f) := P(D)$ for $f \in \Delta$ and $D \in C$ related by $f = f_D$.
(ii) $P_T : X \times \Delta_T \to \Delta(R)$ such that $P_T(f) := P_T(D)$ for $f \in \Delta_T$ and $D \in C^T$ related by $f = f_D$.

As long as no length is fixed, $f \in \Delta$ is universal and the length T of the database that f represents becomes "visible" only through the restriction of $P(f)$ to the specific $P_T(f)$, i.e. P_T "pins" down the unique database the frequency vector is able to represent, namely the database with length T. Of course, under the condition that the frequency vector allows the existence of such a database with this specific length.

2.9.3 Axioms in the frequency framework

Maximal anchored Concatenation Axiom:
(i) Let there be $f \in \Delta_T$, for all $j \neq i \leq m$ $f_i^j(f(i)) \in \Delta_{T_j}$ and $\sum_{j \neq i}^m T_j = T$ such that for $(\alpha_j)_{j \neq i \leq m} \in [0,1]$ and $\sum_{j \neq i}^m \alpha_j = 1$, i.e. $f = \sum_{j \neq i}^m \alpha_j f_i^j(f(i))$.
Let $L = L(f(i), (T_j)_{j \neq i \leq m}) \in \mathbb{N}$ be the corresponding adjusted length, i.e.
$\max_{i \leq m} f(i)T = \max_{l \leq m} f_i^j(f(i))(l)L = \max\{f(i), 1 - f(i)\}L$.
Then there exist $\lambda \in \Delta^{m-1}$ (where $\lambda_j = 0$ for all $j \neq i \leq m$ such that $\alpha_j = 0$), such that

$$P_T(f) = \sum_{j \neq i \leq m} \lambda_j P_L(f_i^j(f(i)).$$

(ii) Let $f_i^j(1) \in \Delta_T$ and $f_j^l(1/2) = (0, ...0, 1/2, 0, ..., 0, 1/2, 0, .., 0)^t \in \Delta_{2W}$ for distinct $i, j, l \leq m$, then for all $\alpha \in (0,1)$ there exist $\lambda \in \Delta^2$ such that

$$P_{T+2W}(\alpha f_i^j(1) + (1-\alpha) f_j^l(1/2)) = \lambda P_{\max\{T,W\}}(f_i^j(1)) + (1-\lambda) P_{\max\{2T,2W\}}(f_j^l(1/2)).$$

Constant Similarity Axiom:
(i) Let there be $f \in \Delta_T$, for all $j \neq i \leq m$ $f_i^j(f(i)) \in \Delta_{T_j}$ such that for $(\alpha_j)_{j \neq i \leq m} \in [0,1]$ and $\sum_{j \neq i}^m \alpha_j = 1$, i.e. $f = \sum_{j \neq i}^m \alpha_j f_i^j(f(i))$.
Let $L = L(f(i), (T_j)_{j \neq i \leq m}) \in \mathbb{N}$ be the corresponding adjusted length.

If there exist $\lambda \in \Delta^{m-1}$ (where $\lambda_j = 0$ for all $j \neq i \leq m$ such that $\alpha_j = 0$), such that for some $Z \in \mathbb{N}_+$

$$P_{ZT}(f) = \sum_{j \neq i \leq m} \lambda_j P_{ZL}(f_i^j(k)),$$

then the equation holds for all $Z \in \mathbb{N}_+$.

(ii) Let $f_i^j(1) \in \Delta_T$ and $f_j^l(1/2) = (0, ...0, 1/2, 0, ..., 0, 1/2, 0, .., 0)^t \in \Delta_{2W}$ for distinct $i, j, l \leq m$. Then (for all $\alpha \in (0,1)$) if there exist $\lambda \in \Delta^2$ such that for some $Z \in \mathbb{N}_+$

$$P_{Z(T+2W)}(\alpha f_i^j(1)+(1-\alpha)f_j^l(1/2)) = \lambda P_{Z\max\{T,W\}}(f_i^j(1))+(1-\lambda)P_{Z\max\{2T,2W\}}(f_j^l(1/2)),$$

then the equation holds for all $Z \in \mathbb{N}_+$.

Learning Axiom:
For all $i \in \{1, 2, ..., C\}$: $(P_T(f^i))_{T \in \mathbb{N}_+}$ converges to $P_\infty(f^i) = P_\infty^i$.

Diversity Axiom:
There exist some $T^* \in \mathbb{N}_+$, such that for all $T \geq T^*$, no three elements of $\{(P_T(f^j))_{j \leq m}\}$ are collinear.

Before stating the sufficiency part of Theorem 2.1 in the frequency version we will present some helpful observations.

2.9.4 Useful observations

Smallest anchored chain of a frequency vector

Remark 2.1
For all anchor cases $c_i \in C$, $i \leq m$, there exists for each $f \in \Delta$ an (frequency based) anchored chain $f = \sum_{j \neq i} \alpha_i^j f_i^j(f(i))$, where $\alpha_i^j \in [0,1]$ are given by $f(j) = \alpha_i^j(1 - f(i))$ for all $j \neq i \leq m$.
Note that α_i^j corresponds to the (relative) lengths of the databases $D_i^j(f(i), \cdot)$ (corresponding to the particular frequency vectors $f_i^j(f(i))$) to the length of the specific database D (which is represented by the frequency vector f).

For instance, assume that $f_i^j(f(i)) \in \Delta_{V_j}$ and $\sum_{j \neq i} V_j = V$, then $f \in \Delta_V$ and $\alpha_i^j = \frac{V_j}{V}$.

In general for all $j \neq i \leq m$, $f_i^j(f(i))$ can represent a database with length $t \cdot \widetilde{T_i^j}$, where $t \in \mathbb{N}$ and $\widetilde{T_i^j} \in \mathbb{N}$ is the smallest length W such that $f(i)W$ is a natural number (and hence also $(1 - f(i))W \in \mathbb{N}$).

To specify the (smallest) length $Z_i \in \mathbb{N}$ of the database D corresponding to the anchored chain of f via anchor case c_i, i.e. $f = \sum_{j \neq i} \alpha_i^j f_i^j(f(i))$, we extend all $\widetilde{T_i^j} \in \mathbb{N}$ with the smallest $z_i^j \in \mathbb{N}$ such that all α_i^j are the fractions with the smallest common denominator Z_i, i.e. $\alpha_i^j = \frac{z_i^j \widetilde{T_i^j}}{Z_i}$ (for $j \neq i \leq m$). In this way, the smallest lengths of the databases represented by $(f_i^j(f(i)))_{j \neq i}$ that can be used for the decomposition of f via anchor c_i are exactly given by

$$T_i^j := z_i^j \widetilde{T_i^j} = \alpha_i^j Z_i \in \mathbb{N} \quad \text{and} \quad Z_i = \sum_{j \neq i \leq m} T_i^j \tag{2.8}$$

Hence $f \in \Delta_{Z_i}$ and $f_i^j(f(i)) \in \Delta_{T_i^j}$ for all $j \neq i \leq m$.
Obviously, choosing a different anchor case $c_l \in C$ for the decomposition of f will lead to a different smallest denominator Z_l (and induced length of database which f represents) and different lengths of the databases $(D_l^j(f(l), T_l^j))_{j \neq l \leq m}$ that are represented by $(f_l^j(f(l)) \in \Delta_{T_l^j})_{j \neq l \leq m}$.

Definition 2.7
For all $f \in \Delta$ and $i \leq m$, we call the anchored chain (wrt. to anchor case c_i) $f = \sum_{j \neq i} \alpha_i^j f_i^j(f(i))$ defined in (2.8) the smallest anchored chain of f representing a database of length $Z_i = \sum_{j \neq i \leq m} T_i^j$ and denote it by $(f_i^j(f(i)), T_i^j)_{j \neq i}$.

The following Lemma shows consistency of the axiomatization with respect to the possible smallest decompositions based on different anchor cases.

Lemma 2.1
Let $P : \Delta \to \Delta(R)$ and its restriction P_T to Δ satisfies the maximal anchored Concatenation and the Constant Similarity Axiom. Then, P is consistent with respect to the different possible smallest decomposition and for all $T \geq 2$ and any $f \in \Delta_T$ there exist $\lambda \in \Delta^m$

$$P_T(f) = \sum_{j \leq m} \lambda_j P_{max_{i \leq m} f(i) T}(f^j) \tag{2.9}$$

Proof:
For all $f \in \Delta$, there exists the smallest anchored chain via anchor case $c_i \in C$ (as in Definition 2.7) $(f_i^l(f(i)), T_i^l)_{l \neq i}$, where $f(l) = \frac{T_i^l}{Z_i}(1 - f(i))$ (which implies

$(1 - f(i))T_i^l = f(l)Z_i)$.
We have to show that independent of the choice of the anchor case $c_i \in C$, the induced beliefs $P_T(f)$ coincide for all $T \in \mathbb{N}$ such that $f \in \Delta_T$.
Observe that the adjusted length defined in Definition 2.3 can be stated explicitly:

$$L(k, T_1, T_2, ..., T_m) = \begin{cases} \max_j\{T_j\} & \text{if } k \leq \frac{\max_j\{T_j\}}{\max_j\{T_j\}+T} =: k^* \in (\frac{1}{m+1}, \frac{1}{2}) \\ \frac{k}{1-k}T & \text{if } k \in (k^*, \frac{1}{2}) \\ T & \text{if } k \geq \frac{1}{2} \end{cases}$$

Thus, we can differentiate the three situations of adjusted lengths depending on the different frequencies of the chosen anchor case $c_i \in C$, i.e. (i) $f(i) \leq k^*$ or (ii) $f(i) \in (k^*, \frac{1}{2})$ or (iii) $f(i) \geq \frac{1}{2}$.

(i) Let $f(i) \leq k^*$, assume w.l.o.g. that $\max_{l \leq m} f(l) = f(j)$, hence $\max_{l \leq m} T_i^l = T_i^j$: Applying the maximal anchored Concatenation Axiom in a first step for $k = f(i) \leq \frac{\max_{j \neq i} T_i^j}{\max_{j \neq i} T_i^j + Z_i}$ with adjusted length $L(f(i), (T_i^l)_{l \neq i}) = T_i^j$, and in the second line for $k = 0$ with adjusted length $L(0, f(i)T_i^j, (1-f(i))T_i^j) = (1-f(i))T_i^j$. Then, for some $a \leq m$ such that $a \neq i, l$, we get for some $\lambda, \gamma, \beta \in \Delta^m$:

$$\begin{aligned} P_{Z_i}(f) &= \sum_{l \neq i} \lambda_l P_{L(f(i),(T_i^l)_{l \neq i})}(f_i^l(f(i))) = \sum_{l \neq i} \lambda_l P_{T_i^j}(f_i^l(f(i))) \\ &= \sum_{l \neq i} \lambda_l(\gamma_l P_{L(0,f(i)T_i^j,(1-f(i))T_i^j)}(f_a^i(0)) + (1-\gamma_l) P_{L(0,f(i)T_i^j,(1-f(i))T_i^j)}(f_a^l(0))) \\ &= \sum_{l \neq i} \lambda_l(\gamma_l P_{(1-f(i))T_i^j}(f^i) + (1-\gamma_l) P_{(1-f(i))T_i^j}(f^l)) \\ &= \sum_l \beta_l P_{(1-f(i))T_i^j}(f^l) = \sum_l \beta_l P_{\max_{l \leq m} f(l) Z_i}(f^l) \end{aligned}$$

By the Constant Similarity Axiom we get that $P_T(f) = \sum_l \beta_l P_{\max_{l \leq m} f(l) T}(f^l)$ for all T such that $f \in \Delta_T$.

Analogous reasoning and application of the Constant Similarity Axiom would yield the same result for (ii) and (iii) directly. However, we just show that for all $f \in \Delta$, there exist $i \neq j \leq m$, such that $f(i) \leq k^*$. Assume that would not be true, then for all $l \leq m$ $f(l) > \frac{\max_{j \neq i} T_i^j}{\max_{j \neq i} T_i^j + Z_i}$ and hence $\sum_{l \leq m} f(l) > m \frac{\max_{j \neq i} T_i^j}{\max_{j \neq i} T_i^j + Z_i} \geq 1$, since $Z_i \leq (m-1)(\max_{j \neq i} T_i^j)$. □

The following Lemma mirrors Lemma A.4 in EG.

Lemma 2.2

Let P satisfy the maximal anchored Concatenation, Constant Similarity and Diversity Axiom and let $(s_j)_{j\leq m}$ be a collection of positive numbers (similarity weights). Define the function $P^s : \Delta(C) \to \Delta(R)$ and for any $T \in \mathbb{N}$, $T \geq 2$ and any $f \in \Delta_T$ the restriction P^s_T to Δ_T by

$$P^s_T(f) = \frac{\sum_{j\leq m} s_j f(j) P_{\max_{j\leq m} f(j)T}(f^j)}{\sum_{j\leq m} s_j f(j)}.$$

Suppose that for some $T \geq T^*$ (given by the Diversity Axiom) and $f \in \Delta_T$ it holds $P_T(f) = P^s_T(f)$.
Then, $P_W(f) = P^s_W(f)$ for all $W \in \mathbb{Z}_+$ such that $f \in \Delta_W$.

Proof:
Let $T(f)$ be the smallest T such that $f \in \Delta_{T(f)}$, then for all $l \in \mathbb{N}$ $f \in \Delta_{lT(f)}$.
By Lemma 2.1 we know that P can be represented as in representation (2.9).
Hence, if there exist some $\lambda \in \Delta^m$ (with $\lambda_i = 0$ if and only if $f(i) = 0$) such that it satisfies for some $l \in \mathbb{N}_+$,

$$P_{lT(f)}(f) = \sum_{j=1}^{m} \lambda_j P_{max_{i\leq m} f(i)T(f)}(f^j),$$

then by the Constant Similarity Axiom it also holds for all $l \in \mathbb{N}_+$. In particular, for l such that $lT(f) = T$ the following holds.

$$P_T(f) = \sum_{j=1}^{m} \lambda_j P_{max_{i\leq m} f(i)T(f)}(f^j) \overset{by\,ass.}{=} \frac{\sum_{j=1}^{m} s_j f(j) P_{max_{i\leq m} f(i)T}(f^j)}{\sum_{j=1}^{m} s_j f(j)} = P^s_T(f)$$

By the Diversity Axiom we get $\lambda_j = \frac{s_j f(j)}{\sum_{j=1}^{m} s_j f(j)}$.
Since $P^s_{lT(f)} = \frac{\sum_{j=1}^{m} s_j f(j) P_{max_{i\leq m} f(i)lT(f)}(f^j)}{\sum_{j=1}^{m} s_j f(j)} = \sum_{j=1}^{m} \lambda_j P_{max_{i\leq m} f(i)lT(f)}(f^j) = P_{lT(f)}(f)$ for all l, the proof is completed. $\square$.

Remark 2.2

Let $f \in \Delta$ be expressed as convex combination of the set $\{f_1, f_2, f_3\}$ for some $f_i \in \Delta_T$ for all $i = 1, 2, 3$, i.e. $f = \beta_1 f_1 + \beta_2 f_2 + (1 - \beta_1 - \beta_2) f_3$.
As in Remark 2.1 we apply the relative length interpretation of the weights $\beta_i \in (0, 1)$ for all $i = 1, 2, 3$, to get the (potentially) smallest induced length H of the database represented by f via the convex combination of databases $D_i \in C^T$, which are represented by $f_i \in \Delta_T$. That is, H is again the smallest possible denominator of all β_i such that for all $i = 1, 2, 3$ $\beta_i = \frac{z_i T}{H}$ for some $z_i \in \mathbb{N}$ and hence we have that $f \in \Delta_H$

can be combined by the decomposition $(f_i)_{i\leq 3}$*, where* $f_i \in \Delta_{\beta_i H = z_i T}$ *for* $i = 1, 2, 3$.

2.9.5 Theorem 2.1, sufficiency part in frequency version

Theorem 2.3

Let there be given a function $P : X \times \Delta \to \Delta(R)$*. Let* P_T *be the restriction of* P *to* $X \times \Delta_T$ *and let for* $T \geq 2$ $\quad P_T : \Delta_T \to \Delta(R)$ *satisfy the Learning, Diversity, maximal Anchored Concatenation and Constant Similarity Axiom.*
Then there exist unique probability vectors $(P_T^j)_{j\leq C} \in \Delta(R)$ *for all* $T \geq 2$ *and unique -up to multiplication by a strictly positive number- strictly positive numbers* $(s_j)_{j\leq m} \in \mathbb{R}^+$ *such that for every* $f \in \Delta_T$

$$P_T(f) = \frac{\sum_{j\leq qm} s_j f(j) P^j_{\max_j f(j)\cdot T}}{\sum_{j\leq m} s_j f(j)}. \tag{2.10}$$

Proof
Step 0:
Obviously, we have to define $P_T^j = P_T(f^j)$ for all $T \geq 2$ and $j \leq m$.

Thus it remains to show that there exist positive numbers $(s_j)_{j\leq m}$ such that the representation holds for all $T \geq 2$ and for every $f \in \Delta_T$.

Rough outline of the proof
As already mentioned in the Sketch of the Proof, we follow an inductive proof on the number of cases in the set of basic cases, i.e. on $m = |C|$.
In Step 1, which serves in Step 2 as the base case for the induction, we proof the theorem for a set of basic cases consisting only of three different basic cases, i.e. $C = \{c_1, c_2, c_2\}$.

Step 1.1: Determination of the similarity values s_1, s_2, s_3
The representation (2.9) in Lemma 2.1 and the representation (2.10) in Theorem 2.3 applied to $\overline{f} := \frac{1}{3}(f^1 + f^2 + f^3)$ yields (with the Diversity Axiom) the similarity values, which allows the definition of $P_T^s(f) := \frac{\sum_{j\leq 3} s_j f(j) P^j_{\max_j f(j) T}}{\sum_{j\leq 3} s_j f(j)}$ for all $f \in \Delta_T$ and $T \in \mathbb{N}$. Of course $f \in \{f^1, f^2, f^3, \overline{f}\}$ satisfy $P_T(f) = P_T^s(f)$.

Step 1.2: Show that $P_T(f) = P_T^s(f)$ for all simplicial points (Figure 2.1 illustrates simplicial partitions and points)

The main tool to show this claim is the observation that for four specifically structured frequency vectors (anchored frequency vectors) that satisfies already the desired equation, also the intersection of the lines between two of these (specific) vectors satisfies the above equation (Lemma 2.4). The crucial step in the proof is to apply this fact in an appropriate recursive way. In this step again the maximal anchored Concatenation Axiom and the Constant Similarity Axiom (in form of Lemma 2.2) are necessary.

Step 1.3: Show that $P_T(f) = P^s_T(f)$ for all frequency vectors $f \in \Delta(C)$
The proof is similar to a (rewritten/revised) proof of Lemma A.6 in EG, which is based on the existence of the limit of P^c_T for all $c \in C$ (Learning Axiom). Since all frequency vectors $f \in \Delta$ can be approximated by a series of simplicial triangles/points, we can show the claim (by using Lemma 2.1 and Lemma 2.2). In particular, one can show that the beliefs P and P^s induced by the sequence of simplicial points, which approximates f, converges to the belief of P and P^s induced by the limit f. Using the equivalence of $P^s_T(g)$ and $P_T(g)$ for the sequence of simplicial points $g \in \Delta$ by Step 1.2 delivers the claim.

In Step 2, the result from Step 1 is used inductively for a general set of basic cases $C = \{c_1, c_2, ..., c_m\}$ with $m > 3$.

Step: 2.1: Defining the similarity weights $s_1, ..., s_m$
Step 1 yield for any triple of cases $\{c_j, c_k, c_l\} \subseteq C$ for distinct $j, k, l \in \{1, 2, ..., m\}$ similarity weights $s_j^{(j,k,l)}, s_k^{(j,k,l)}, s_l^{(j,k,l)}$. As in the proof of Proposition 3, Step 2.1 in BGSS, one can show that each similarity weight can be chosen independent of the choice of the triple, i.e. $s_j^{(j,k,l)} = s_j$. Hence, we define $P^s_T(f) := \frac{\sum_{j \leq m} s_j f(j) P^j_{\max_j f(j)T}}{\sum_{j \leq m} s_j f(j)}$ for all $f \in \Delta(C)$ and $T \in \mathbb{N}$.

Step 2.2: Show $P_T(f) = P^s_T(f)$ for all $f \in \Delta(C)$
Inductively on $|C| = m$ for $f \in conv(\{(f^j)_{j \leq m}\})$, where we use Step 1 as base case of the induction. Each $f \in \Delta$ can be written as anchored chain with different anchors (Remark 2.1). Applying the maximal anchored Concatenation Axiom to these chains yield m-many different hyperplanes, which are spanned by $(P(f^j_i(f(i))))_{j \neq i \leq m}$, for different $i \leq m$. All these hyperplanes contain $P(f)$ and include $P^s(f)$ as well, since $P_T(f^j_i(k)) = P^s_T(f^j_i(k))$ for any $i \neq j \leq m$ and $f^j_i(k) \in \Delta_T$ by induction assumption. Using the Constant Similarity Axiom (Lemma 2.2) and Lemma 2.1 to harmonize the different hyperplanes wrt. lengths, we can show that the intersection of all these

induced hyperplanes is unique, which delivers the desired result.

2.9.6 Step 1: $C = \{c_1, c_2, c_3\}$, i.e. $m = 3$

Step 1.1:

Define $\overline{f} := \sum_{j\leq 3} 1/3 f^j$, for $f^j \in \Delta_T$, and $T \geq T^*$ then $\overline{f} \in \Delta_{3T}$. The positive numbers s_1, s_2, s_3 result from equating the evaluation of $\overline{f}$ using the representation (2.9) in Lemma 2.1, i.e. $P_{3T}(\overline{f}) = \lambda_1 P_T^1 + \lambda_2 P_T^2 + (1-\lambda_1-\lambda_2)P_T^3$ with representation (2.10) in Theorem 2.3 and solving the linear system. The solution of this linear system s_1, s_2, s_3 exist uniquely up to multiplication by a positive number due to the non collinearity condition of the Diversity Axiom for $T \geq T^*$, otherwise uniqueness is not achievable.
Define for all T and $f \in \Delta_T$

$$P_T^s(f) := \frac{\sum_{j\leq 3} s_j f(j) P^j_{\max_j f(j)T}}{\sum_{j\leq 3} s_j f(j)}$$

Obviously $P_T^s(f^j) = P_T(f^j)$ (Step 0) for all $j = 1, 2, 3$ and $P_T^s(\overline{f}) = P_T(\overline{f})$.
The aim is to show for all T and for every $f \in \Delta_T$:

$$P_T^s(f) = P_T(f) \tag{2.11}$$

In the following, we will recursively partition the simplex Δ into so called simplicial triangles, as illustrated in the Figure 2.1 below.

Definition of Simplicial Triangles:

The 0-th simplicial partition consist of vertices $q_0^j \in \Delta$, which are exactly the unit vectors f^j for $j = 1, 2, 3$. The first simplicial partition of Δ is a partition to four triangles separated by the segments connecting the middle points between two of the three unit frequency vectors, i.e. $q_1^1 := (\frac{1}{2}f^1 + \frac{1}{2}f^2)$, $q_1^2 := (\frac{1}{2}f^2 + \frac{1}{2}f^3)$ and $q_1^3 := (\frac{1}{2}f^3 + \frac{1}{2}f^1)$. The second simplicial partition is obtained by similarly partitioning each of the four triangles to four smaller triangles, and the l-th simplicial partition is defined recursively. The simplicial points of the l-th simplicial partition are all the vertices of triangles of this partition. Note that for $j = 1, 2, 3$ the q_0^j are frequency vectors representing databases consisting only of one case, but of any length $T \in \mathbb{N}$, i.e. $q_0^j \in \Delta_T$ for all $T \in \mathbb{N}_+$. All vertices q_l^v of the l-th simplicial partition are in $\Delta_{2^l T}$ for all $T \in \mathbb{N}_+$ for appropriate $v \leq n_l$ (defined below in (2.12)).

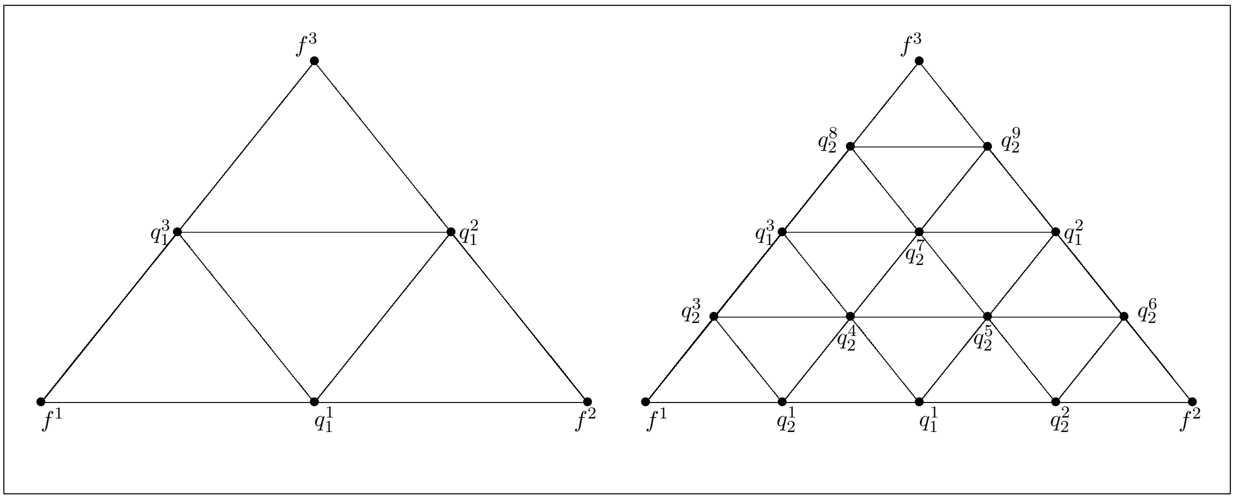

Figure 2.1: 1st and 2nd simplicial partitions

Considering the simplicial points on the line between f^1 and f^2, we get for the 0-th simplicial partition: 2 simplicial points, for 1st simplicial partition: 3 simplicial points: for 2nd simplicial partition: 5 simplicial points, for 3rd simplicial partition: 9 simplicial points and so forth, i.e. it follows the series $a_l = 2^l + 1$ for all $l \in \mathbb{N}$. Observe that for each parallel line to (f^1, f^2) between simplicial points of the l-th simplicial partition, the line which is one "step closer" to f^3, possesses one simplicial point less than the parallel line that is further away from f^3. The number of simplicial points on these parallel lines decreases until reaching the point f^3. Hence the total number n_l of simplicial points of the l-th partition is given by

$$n_l := \sum_{i=1}^{a_l} i = \sum_{i=1}^{2^l+1} i = 2^{2l-1} + 2^l + 2^{l-1} + 1 \qquad \text{where } a_l = 2^l + 1 \tag{2.12}$$

Step 1.2: $P_T(f) = P_T^s(f)$ holds for all Simplicial Points

Lemma 2.3

The vertices q_l^v with $v \leq n_l$ of the l-th simplicial partition satisfy equation (2.11) for all $l \in \mathbb{N}$.

Notation: In the following we will denote for $a, b \in \Delta$ or $a, b \in \Delta(R)$ the straight line through a and b by (a, b) (since there is no confusion with the usual interval notation).

Main tool of the proof of this Lemma is the following observation.

Lemma 2.4

Let $a, b, c, d \in \Delta$ be distinct frequency vector satisfying equation (2.11) and the lines (a, b) and (c, d) are not collinear. Then the intersection y of the line (a,b) and (c,d), i.e. $y = (a, b) \cap (c, d)$ satisfies equation (2.11) (for an appropriate T such that $y \in \Delta_T$) if either of the following conditions hold for the pairs a, b and c, d:
(i) both vectors a and b (respectively c and d) lie on a line $(f_i^j(k), f_i^h(k))$ for some $k \in [0, 1]$ and distinct $i, j, h \leq m$, which represent anchored databases with identical anchor case $c_i \in C$ or
(ii) a, b (respectively c, d) lie on a line through $(f_i^j(1), f_j^h(1/2))$ for some distinct $i, j, h \leq m$.

Proof of Lemma 2.4

We will show the situation, where both pairs a, b and c, d satisfy condition (i). Assume that $a, b \in (f_i^j(k), f_i^h(k))$, hence also $y \in (f_i^j(k), f_i^h(k))$. By Remark 2.1 we know that there exist an anchored chain of y composed of $(f_i^j(k), f_i^h(k))$, i.e. there exists some $\alpha \in (0, 1)$ and $Z^y \in \mathbb{N}$ such that $y = \alpha f_i^j(k) + (1 - \alpha) f_i^h(k) \in \Delta_{Z^y}$ with corresponding adjusted length $L^y := L(k, \alpha Z^y, (1 - \alpha) Z^y)$ such that $P_{Z^y}(y) \in \big(P_{L^y}(f_i^j(k)), P_{L^y}(f_i^h(k))\big)$.

Analogously, there exist some Z^x and L^x for all $x \in \{a, b\}$. Let $L := LCM(Z^y, Z^a, Z^b)$ (least common multiple), then for all $v \in \{y, x\}$ the following holds

$$P_{Z^v \frac{L}{L^v}}(v) \in \big(P_L(f_i^j(k)), P_L(f_i^h(k))\big)$$

In particular $P_{Z^a \frac{L}{L^a}}(a)$ and $P_{Z^b \frac{L}{L^b}}(b)$ determine already the slope of the line $\big(P_L(f_i^j(k)), P_L(f_i^h(k))\big)$ and hence $P_{Z^y \frac{L}{L^y}}(y) \in \Big(P_{Z^a \frac{L}{L^a}}(a), P_{Z^b \frac{L}{L^b}}(b)\Big)$.
The same derivation with P^s yields $P^s_{Z^y \frac{L}{L^y}}(y) \in \Big(P^s_{Z^a \frac{L}{L^a}}(a), P^s_{Z^b \frac{L}{L^b}}(b)\Big)$ and since we know that $a, b \in \Delta$ satisfy (2.11), we get:

$$P_{Z^y \frac{L}{L^y}}(y), P^s_{Z^y \frac{L}{L^y}}(y) \in \Big(P_{Z^a \frac{L}{L^a}}(a), P_{Z^b \frac{L}{L^b}}(b)\Big)$$

The same procedure applied to pair c, d instead to a, b with $G := LCM(L^y, L^c, L^d)$ yields:

$$P_{Z^y \frac{G}{L^y}}(y), P^s_{Z^y \frac{G}{L^y}}(y) \in \Big(P_{Z^c \frac{G}{L^c}}(c), P_{Z^b \frac{G}{L^c}}(c)\Big)$$

Finding a least common multiple $J = LCM(L, G)$ will deliver the desired result, since

$$P_{Z^y \frac{J}{L^y}}(y), P^s_{Z^y \frac{J}{L^y}}(y) \in \Big(P_{Z^a \frac{J}{L^a}}(a), P_{Z^b \frac{J}{L^b}}(b)\Big) \cap \Big(P_{Z^c \frac{J}{L^c}}(c), P_{Z^d \frac{J}{L^d}}(d)\Big)$$

and the intersection is unique (otherwise this would be a contradiction to the Diversity Axiom). Hence $P^s_{Z^y \frac{J}{L^y}}(y) = P_{Z^y \frac{J}{L^y}}(y)$ and by Lemma 2.2 $P_T(y) = P^s_T(y)$ for all T such that $y \in \Delta_T$.

The situation, in which one of the two pairs satisfies condition (i) and the other condition (ii) or both pairs fulfill condition (ii) can be shown analogously. $\square$

The proof of Lemma 2.3 is conducted by using the observation in Lemma 2.4 in an appropriate procedure recursively, as can be seen in the series of figures (Figures 2.2 to 2.4) below.

Proof of Lemma 2.3 by induction over l (l-th simplicial partition)

Base case for induction $l = 0$:
By Step 1.1 we know that for $(f^1 = q^1_0, f^2 = q^2_0, f^3 = q^3_0)$ the representation holds.
Induction step:
The idea of the (recursive) procedure to capture all simplicial points of the $(l+1)-$th simplicial partitions as intersections of two lines that run through simplicial points of the l-th simplicial partition can be understand by inspecting the series of Figures 2.2-2.4. However, we need some definitions first:
(i) $g^{(i,j)}_l(d)$ denotes the simplicial point in the l-th partition on the line (f^i, f^j) such that it is the d closest to f^i.
(ii) $b^i_l(d)$ denotes the simplicial point of the l-th partition that lies on the line (f^i, q^*) and also on the d- closest line to f^i that is parallel to (f^j, f^h) in the l-th partition.
We construct a procedure such that appropriately chosen lines through these points can intersect exactly in all simplicial points. Lines through these points are essential to cover all simplicial points through intersections.
More precisely, for all $l \in \mathbb{N}$ and all distinct $i, j, h \in \{1, 2, 3\}$ it holds that $b^i_{l+1}(d) := (f^i, \frac{1}{2}(f^j + f^h)) \cap (g^{(i,j)}_l(d), g^{(i,h)}_l(d))$, where all frequency vectors on the right hand side satisfy the conditions in Lemma 2.3 and given that (2.11) holds for the frequency vector of the l-th partition, also it holds for $b^i_{l+1}(d)$. Figures 2.2-2.4 show the procedure given that all simplicial points of the 2nd partition satisfy (2.11) and show how all simplicial points of the 3rd partition can be covered. In each of these figures bullets represent points that satisfy (already) equation (2.11) and any intersections of (appropriate) lines through these "bullet"-points are again simplicial points that satisfy (2.11).
For the sake of completeness we will spell out the general steps of the illustrated

procedure.

Let the claim be true for the l-th simplicial partition. For the $(l+1)$-th partition the following procedure will capture all simplicial points q_{l+1}^v for $v \leq n_{l+1}$.

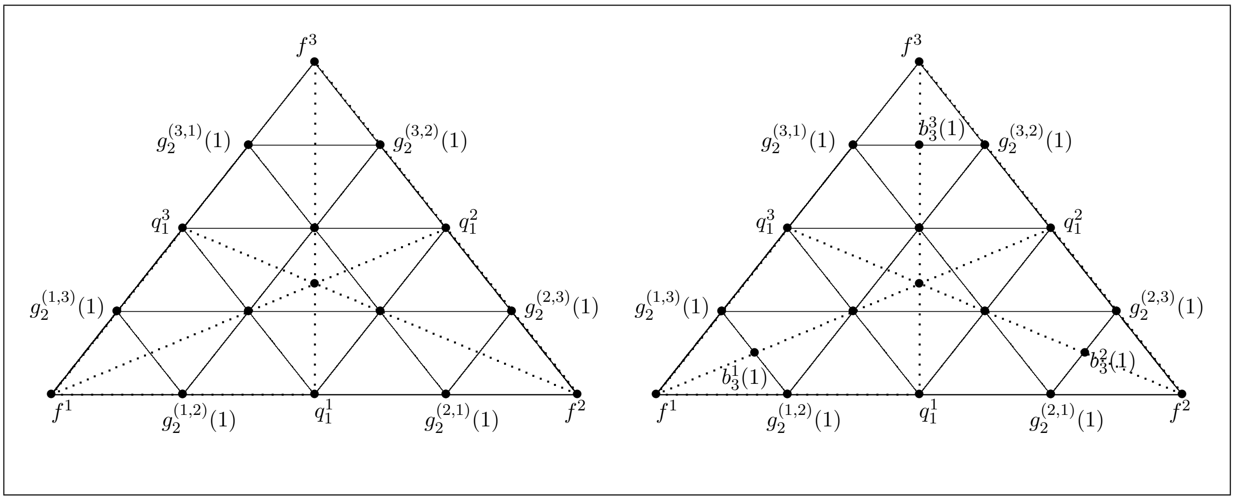

Figure 2.2: From 2nd to 3rd simplicial partition.
Assume that all simplicial points (bullets) of the 2-nd partition satisfy already equation (2.11). Here some points are named according to the notation used in the procedure.
For example take the simplicial points of the 2nd partition that is on (f^1, f^2) and closest to f^1, i.e. $g_2^{(1,2)}(1)$. Analogously, take the closest to f^1 on (f^1, f^3), i.e. $g_2^{(1,3)}(1)$. Intersecting $(g_2^{(1,2)}(1), g_2^{(1,3)}(1))$ with (f^1, q_1^2) shows that $b_3^1(1)$ satisfies equation (2.11) as well. Analogously, this can be shown for $b_3^i(1)$ $(i = 2, 3)$ using appropriate combinations of $g_2^{(i,j)}(1), q_1^u, f^h$ for $u, h \in \{1, 2, 3\}$.

Procedure:

(i) **For** $d = 1$**:**

W.l.o.g. take the perspective of $f^j = f^1$ for a $j \in \{1, 2, 3\}$. Given the l-th simplicial partition, there exist a simplicial point of the $(l+1)$-th simplicial partition $b_{l+1}^1(1)$, which is the intersection of the lines (f^1, f^2) and $(g_l^{(1,2)}(1), g_l^{(1,3)}(1))$ (Corresponds to Figure 2.2 RHS). By the induction assumption these pairs of points satisfy equation (2.11) and the conditions of the Lemma 2.4. Hence $P_{2^{l+1}}^s(b_{l+1}^1) = P_{2^{l+1}}(b_{l+1}^1)$, i.e. $b_{l+1}^1(1)$ satisfies equation (2.11). Analogously the same procedure applied to f^j for $j = 2, 3$ yields that $b_{l+1}^j(1)$ satisfies equation (2.11).

(ii) Draw the line between two elements of $\{b_{l+1}^1(1), b_{l+1}^2(1), b_{l+1}^3(1)\}$, w.l.o.g. take $b_{l+1}^1(1)$ and $b_{l+1}^3(1)$. (Corresponds to Figure 2.3 LHS) The line $(b_{l+1}^1(1), b_{l+1}^3(1))$ intersects for all $0 \leq z \leq a_l$ (defined in (2.12)) with the lines $(g_l^{(1,3)}(z), g_l^{(2,3)}(z))$ (that

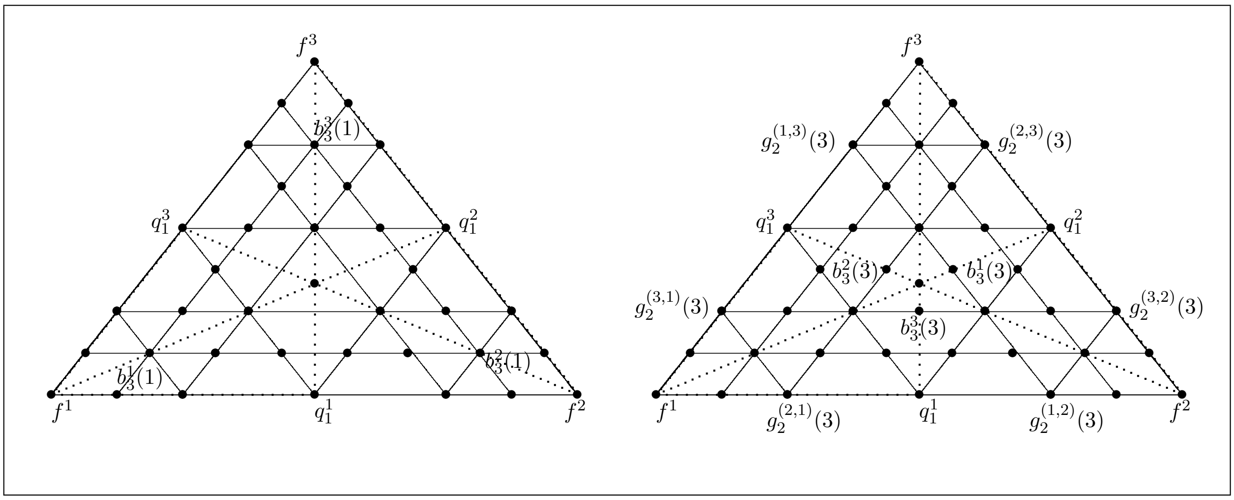

Figure 2.3: From 2nd to 3rd simplicial partition.
a) Using lines $(b_3^i(1), b_3^j(1))$ and their intersections with existing lines will show equation (2.11) for all simplicial points of the 3rd partition that are on the closest parallel lines to the rim of the simplex.
b) Now, consider the simplicial points of the 2nd partition that are on the lines (f^i, f^j) and third closest to f^i (note: the second closest are covered indirectly). For example, take $g_2^{(1,2)}(3)$ (third closest to f^1 on line (f^1, f^2)) and $g_2^{(1,3)}(3))$ (third closest to f^1 on (f^1, f^3)). Intersecting $(g_2^{(1,2)}(3), g_2^{(1,3)}(3))$ with (f^1, q_1^2) shows that $b_3^1(3)$ satisfies equation (2.11) as well. Analogously, this can be shown for $b_3^i(3)$ $(i = 2, 3)$ using appropriate combinations of $g_2^{(i,j)}(3), q_1^u, f^h$ for $u, h \in \{1, 2, 3\}$.

are parallel to the line (f^1, f^2)) and also with all lines $(g_l^{(1,2)}(z), g_l^{(1,3)}(z))$ (that are parallel to (f^2, f^3)). Lemma 2.4 yields that all simplicial points of the $(l+1)$-th partition that lie on the line $(b_{l+1}^1(1), b_{l+1}^3(1))$ satisfy equation (2.11).
Analogously, the procedure yields that all simplicial points of the $(l+1)$-th partition that lie on the lines $(b_{l+1}^i(1), b_{l+1}^j(1))$ for all combinations of $i \neq j \in \{1, 2, 3\}$ satisfy equation (2.11), i.e. all points that are on the closest parallel lines to $(f^1, f^2), (f^1, f^3), (f^2, f^3)$ and in particular, the closest $(l+1)$-simplicial points to f^1, f^2, f^3 on the boundary of $conv(\{f^1, f^2, f^3\})$.

(iii) Apply the procedure of (i) and (ii) (where $d = 1$) (Corresponds partially to Figure 2.3 RHS and 2.4) recursively for $d = 2n - 1 > 1$ for $2 \leq n \leq \frac{a_{l+1}-1}{2}$ (from f^1-view).
Derive $\{b_{l+1}^1(d), b_{l+1}^2(d), b_{l+1}^3(d)\}$ by (i) using f^h and $(g_l^{(i,j)}(d)$ for $i, j, h \in \{1, 2, 3\}$ appropriately. Using (ii), we can show that all simplicial points of the $(l+1)$-th partition that lie on the lines $(b_{l+1}^i(d), b_{l+1}^j(d))$ for all combinations of $i \neq j \in \{1, 2, 3\}$ satisfy equation (2.11).

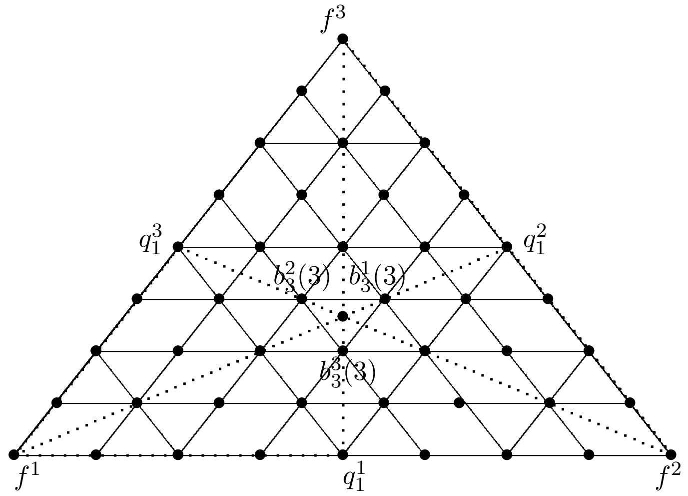

Figure 2.4: As before, the next step would be to intersect the lines between $(b_3^i(1), b_3^j(1))$ for any distinct $i, j \in \{1,2,3\}$ and the existing lines. It shows that all simplicial points of the 3rd partition that are on the third closest parallel lines to the rim of the simplex satisfy (2.11). This covers all simplicial points of the 3rd partition.

Observe that for $d = 2n$ with $1 \leq n \leq \frac{a_{l+1}-1}{2}$ the simplicial points of the $(l+1)$-th partition, which lie on the lines $(b_{l+1}^i(d), b_{l+1}^j(d))$ for all combinations of $i \neq j \in \{1,2,3\}$ are already satisfying the equation (2.11) directly, since these lines already are 'used' for the procedure in the l-th partition and the simplicial points of the $(l+1)$-th partition are just indirectly processed via the intersection steps in (ii).

Step 1.3: Completion to all $f \in \Delta$, i.e. for all $2 \leq T \in \mathbb{N}$ and $f \in \Delta_T$: $P_T(f) = P_T^s(f)$.

Some helpful considerations:

For each $f \in \Delta$ there exists a sequence of simplicial triangles of the l-th partition $(q_l^{i_l}, q_l^{j_l}, q_l^{h_l})_{l\in\mathbb{N}}$ (remember $q_l^v \in \Delta_{2^l}$ for all $v \leq n_l$) for distinct $i_l, j_l, h_l \leq n_l$ such that:

(i) $f \in conv(\{q_l^{i_l}, q_l^{j_l}, q_l^{h_l}\})$ for all $l \in \mathbb{N}$, i.e. there exist $\beta_l^v \in [0,1]$ for all $v \in \{i_l, j_l, h_l\}$ such that $f = \beta_l^{i_l} q_l^{i_l} + \beta_l^{j_l} q_l^{j_l} + \beta_l^{h_l} q_l^{h_l}$

(ii) For all $v \in \{i_l, j_l, h_l\}$ and $l \in \mathbb{N}$ it holds $q_l^v \in \Delta_{\beta_l^v H_l}$, such that H_l (as in Remark 2.2) is the smallest common denominator of all β_l^v, i.e. there exist z_l^v, such that $\beta_l^v = \frac{z_l^v 2^l}{H_l}$. Hence, if f is represented by combination of l-th simplicial points, then $f \in \Delta_{H_l}$.

(iii) $lim_{l\to\infty} q_l^v = f$ for all $v \in \{i_l, j_l, h_l\}$

Clearly this construction is possible for all $f \in \Delta$.

In order to proof that for all $f \in \Delta_T$: $P_T(f) = P^s_T(f)$ we will show the following observations.
(A) $\lim_{l\to\infty} ||P^s_{H_l}(f) - P^s_{H_l}(q^v_l)|| = 0$ and
(B) $\lim_{l\to\infty} ||P_{H_l}(f) - P_{H_l}(q^v_l)|| = 0$ for all $v \in \{1, 2, 3\}$.

Proof of (A):
By the Learning Axiom and since $P^s_T(f^i) = P_T(f^i)$ for all $T \in \mathbb{N}$, we know that for all $i \leq 3$, we have $lim_{l\to\infty} P^s_{H_l}(f^i) = P_\infty(f^i) = P^i_\infty$.

We want to show for all $v \in \{i_l, j_l, h_l\}$: $lim_{l\to\infty}||P^s_{H_l}(f) - P^s_{H_l}(q^v_l)|| = 0$.
Let for all $r \in R$, $P^i_T(r)$ be the r-th component of the probability vector.
For all l, any $v \in \{i_l, j_l, h_l\}$ and q^v_l we have that $\lim_{l\to\infty} q^v_l = f$ and hence $\lim_{l\to\infty} P^j_{\max_j q^v_l(j) H_l}(r) = P^j_{\max_j f(j) H_l}(r)$ holds. This directly implies $\lim_{l\to\infty}(P^s_{H_l}(f)(r) - P^s_{H_l}(q^v_l)(r)) = 0$ for all $r \in R$ and hence the desired result.

Proof of (B):
By Lemma 2.1 we know that $P_H(f) = \sum_{j=1}^3 \lambda_j P_{\max_{i=1,2,3} f(i) H}(f^j)$, where $\lambda \in \Delta^2$ is independent of the length of the database by the Constant Similarity Axiom .
Hence by the Learning Axiom $lim_{l\to\infty} P_{H_l}(f) = lim_{l\to\infty} \sum_{j=1}^3 \lambda_j P_{\max_{\{i=1,2,3\}} f(i) H_l}(f^j)$ exists and hence with the same reasoning as in (A) for P^s we get

$$\lim_{l\to\infty} ||P_{H_l}(q^v_l) - P_{H_l}(f)|| = 0 \quad \text{for all } v \in \{i_l, j_l, h_l\}$$

Combining Step (A) and (B) and the triangle inequality yields:

$$lim_{l\to\infty}||P^s_{H_l}(f) - P^s_{H_l}(q^v_l) - P_{H_l}(f) + P_{H_l}(q^v_l)|| \leq 0$$

By Step 1.2 we know that $P_{H_l}(q_l^v) = P^s_{H_l}(q_l^v)$ for all l, which leads to $\lim_{l\to\infty} ||P^s_{H_l}(f) - P_{H_l}(f)|| = 0$ and for all $r \in R$ we get:

$$\begin{aligned}
0 &= \lim_{l\to\infty}(P_{H_l}(f)(r) - P^s_{H_l}(f)(r)) \\
&= lim_{l\to\infty}\Big(\sum_{j=1}^{3}\lambda_j P_{\max_i f(j)H_l}(f^j)(r) - \frac{\sum_{j\leq 3} s_j f(j) P^j_{\max_j f(j)H_l}(r)}{\sum_{j\leq 3} s_j f(j)}\Big) \\
&= \lim_{l\to\infty}\sum_{j=1}^{3} P_{\max_i f(j)H_l}(f^j)(r)\Big(\lambda_j - \frac{s_j f(j)}{\sum_{j\leq 3} s_j f(j)}\Big) \\
&= \sum_{j=1}^{3} P_\infty(f^j)(r)\Big(\lambda_j - \frac{s_j f(j)}{\sum_{j\leq 3} s_j f(j)}\Big)
\end{aligned}$$

By the Diversity Axiom no three of $(P_\infty(f^j))_{j\leq 3}$ are collinear (i.e. also no three of $(P_\infty(f^j)(r))_{j\leq 3}$ are convex combinations for all r), which implies that it must hold that $\lambda_j = \frac{s_j f(j)}{\sum_{j\leq 3} s_j f(j)}$ for all $j \in \{1,2,3\}$. Therefore, $P_{H_l}(f) = P^s_{H_l}(f)$ for all l and by Lemma 2.2 $P_T(f) = P^s_T(f)$ for all T such that $f \in \Delta_T$. $\square$.
Thus, the proof for $C = \{c_1, c_2, c_3\}$ is concluded.

2.9.7 Step 2: Set of basic cases with $|C| = m > 3$ many cases

Step 2.1 Defining the similarity weights:

Consider for $T \geq T^*$ and distinct $j,k,l \leq m$ a triple $\{P^j_T, P^k_T, P^l_T\}$
Adopting the considerations of Step 1 in the previous subsection for $\{j,k,l\}$, i.e. $\overline{f_{3T}} := \sum_{i\in\{j,k,l\}} f^i$ and $f^i \in \Delta_T$ we can derive the similarity weights $(s_i^{\{j,k,l\}})_{i\in\{j,k,l\}}$ and the following representation for all $f \in conv(\{f^j, f^k, f^l\}) \cap \Delta_T$:

$$P_T^{\{j,k,l\}}(f) = \frac{\sum_{i\in\{j,k,l\}} s_i^{\{j,k,l\}} f(i) P^{\{j,k,l\}}_{max_i f(i)T}(f^i)}{\sum_{i\in\{j,k,l\}} s_i^{\{j,k,l\} f(i)}}.$$

Moreover for all $i \in \{j,k,l\}$, we have $P_T^{\{j,k,l\}}(f^i) = P_T(f^i) = P^i_T$ and $(s_i^{\{j,k,l\}})_{i\in\{j,k,l\}}$ are unique up to multiplication by a positive number.

Now we want to show that the similarity values $s_i^{\{j,k,l\}}$ are independent of the choice of j,k and l for all $i \in \{j,k,l\}$. Similar to BGSS we shown it two steps:

1. We show that $\frac{s_j^{\{j,k,l\}}}{s_k^{\{j,k,l\}}} = \frac{s_j^{\{j,k,n\}}}{s_k^{\{j,k,n\}}}$ for any $n \neq l$, i.e. the ratio between two similarity number is independent of the choice of a third case/frequency.
Consider the evaluation of a rational combinations of $f^j \in \Delta_T$ and $f^k \in \Delta_T$, i.e.

for $\alpha \in \mathbb{Q}$: $f = \alpha f^j + (1-\alpha) f^k$, where H is the smallest common denominator of $\alpha, (1-\alpha)$ and hence $f \in \Delta_H$, w.l.o.g. assume $\alpha \geq (1-\alpha)$. Then,

$P_H^{\{j,k,l\}}(f) = \frac{s_j^{\{j,k,l\}} \alpha P^j_{\alpha H} + s_k^{\{j,k,l\}}(1-\alpha) P^k_{\alpha H}}{s_j^{\{j,k,l\}} \alpha + s_k^{\{j,k,l\}}(1-\alpha)}$ and $P_H^{\{j,k,n\}}(f) = \frac{s_j^{\{j,k,n\}} \alpha P^j_{\alpha H} + s_k^{\{j,k,n\}}(1-\alpha) P^k_{\alpha H}}{s_j^{\{j,k,n\}} \alpha + s_k^{\{j,k,n\}}(1-\alpha)}$.

Equating these two expressions, we get:

$\frac{s_j^{\{j,k,l\}} \alpha}{s_j^{\{j,k,l\}} \alpha + s_k^{\{j,k,l\}}(1-\alpha)} = \frac{s_j^{\{j,k,n\}} \alpha}{s_j^{\{j,k,n\}} \alpha + s_k^{\{j,k,n\}}(1-\alpha)}$ and $\frac{s_k^{\{j,k,l\}}(1-\alpha)}{s_j^{\{j,k,l\}} \alpha + s_k^{\{j,k,l\}}(1-\alpha)} = \frac{s_k^{\{j,k,n\}}(1-\alpha)}{s_j^{\{j,k,n\}} \alpha + s_k^{\{j,k,n\}}(1-\alpha)}$,

which leads to $\frac{s_j^{\{j,k,l\}}}{s_k^{\{j,k,l\}}} = \frac{s_j^{\{j,k,n\}}}{s_k^{\{j,k,n\}}}$

Denote this ratio by $S_{j,k} := \frac{s_j^{\{j,k,l\}}}{s_k^{\{j,k,l\}}}$, which is defined for all distinct $j,k \leq m$, since the similarity numbers are strictly positive.

Further observe that the following holds:

$$S_{j,k} S_{k,l} S_{l,j} = \frac{s_j^{\{j,k,l\}}}{s_k^{\{j,k,l\}}} \frac{s_k^{\{j,k,l\}}}{s_l^{\{j,k,l\}}} \frac{s_l^{\{j,k,l\}}}{s_j^{\{j,k,l\}}} = 1 \tag{2.13}$$

2. Define $s_1 := 1$ and $s_j = S_{j,1}$ for all $j \leq m$.

The **aim** is to show that for all triple $\{j,k,l\}$ it holds that $s_i^{\{j,k,l\}} = a s_i$ for some $a \in \mathbb{R}_+$.

If we can show that $\frac{s_i^{\{j,k,l\}}}{s_m^{\{j,k,l\}}} = \frac{s_i}{s_m}$ for all $m \neq i \in \{j,k,l\}$, then $s_i^{\{j,k,l\}} = \frac{s_i}{s_m} s_m^{\{j,k,l\}} = a s_i$ for all $m \neq i \in \{j,k,l\}$, e.g. with $m = k$ we have $a = \frac{s_k^{\{j,k,l\}}}{s_k}$ and hence $s_j^{\{j,k,l\}} = a s_j$, $s_k^{\{j,k,l\}} = \frac{s_k^{\{j,k,l\}}}{s_k} s_k$, $s_l^{\{j,k,l\}} = a s_l$ and we have shown the claim.

Hence it suffices to show w.l.o.g. that $\frac{s_j^{\{j,k,l\}}}{s_k^{\{j,k,l\}}} = \frac{s_j}{s_k}$ or equivalent $S_{j,k} = \frac{s_j}{s_k}$.

The latter follows directly from (2.13), i.e. $1 = S_{1,j} S_{j,k} S_{k,1} = 1/s_j S_{j,k} s_k$, hence $S_{j,k} = \frac{s_j}{s_k}$.

The independence of the similarity values $s_i^{\{j,k,l\}}$ on $\{j,k,l\}$ allows to replace the (unique up to multiplication by a strictly positive number) $s_i^{\{j,k,l\}}$ by the just defined s_i for all $i \leq m$. Based on these $(s_i)_{i \leq m}$ we define as in the consideration in Step 1.1 for all $2 \leq T \in \mathbb{N}$ and any $f \in \Delta_T$.

$$P_T^s(f) := \frac{\sum_{i \leq m} s_i f(i) P^i_{max_i f(i) T}}{\sum_{i \leq m} s_i f(i)}$$

As in the subsection before the aim is to show that for all T and any $f \in \Delta_T$ $P_T^s(f) = P_T(f)$ holds.

Step 2.2: Completion to all $f \in \Delta$

For $M \subseteq \{1,2,....,m\}$, let $\Delta_T^M := \Delta_T \cap conv(\{f^j \,|\, j \in M\})$ denote the set of all frequency vectors $f \in \Delta_T$, which assign zero frequency to all cases $(c_i)_{i\in\{1,2,...,m\}\setminus M}$.

Lemma 2.5
For every subset $M \subseteq \{1,2,....,m\}$ with $|M| = m \geq 3$, $P_T(f) = P_T^s(f)$ holds for every $f \in \Delta_T^M$.

Proof:
For $m = 3$ the claim has been shown in Step 1 (or Step 1.3) and serves as the base case for our induction over m.
Hence we assume now that the claim holds for $m \geq 3$ and we prove it for M with $|M| = m + 1$.

(1) Let $f \in \Delta_T^M$ such that $f \in conv(\{f^j\}_{j\in M\setminus l})$ for some $l \in M$, then by induction assumption $P_T^s(f) = P_T(f)$.
(2) Now we consider $f \in int(conv(\{f^l \mid l \in M\}))$
By Remark 2.1 we know that for all $i \in M$ and $l \in M\setminus\{i\}$ there exist for all $l \neq i \leq m$ some $\alpha_i^l \in (0,1)$ and $\sum_{l\in M\setminus\{i\}} \alpha_i^l = 1$ such that $f = \sum_{l\in M\setminus\{i\}} \alpha_i^l f_i^l(f(i))$ with $f_i^l(f(i)) \in \Delta_{T_i^l}$ and then $f \in \Delta_{Z_l}$ (where $T_i^j = \alpha_i^j Z_i$). W.l.o.g. (due to Constant Similarity Axiom, Lemma 2.2) assume that for all $l \neq i \leq m$ there exist T_i^l such that $\max\{f(i),(1-f(i))\}T_i^l \geq T^*$ (to overcome potential collinearity problems). In the following we abbreviate the corresponding adjusted lengths $L(f(i),(T_i^j)_{j\neq i\in M})$ by L_i for all $i \in M$.
The maximal anchored Concatenation Axiom induces that $P_{Z_l}(f)$ lies on the following induced $(m+1)$-many hyper-planes $A_l^{m+1}(Z_l)$ for all $l \in M$, w.l.o.g. assume that $M = \{1,2,...,m+1\}$:

$$\begin{aligned}
P_{Z_1}(f) &\in int(conv(\{P_{L_1}(f_1^2(f(1))), P_{L_1}(f_1^3(f(1))), ..., P_{L_1}(f_1^{m+1}(f(1)))\})) =: A_1^{m+1}(Z_1) \\
P_{Z_2}(f) &\in int(conv(\{P_{L_2}(f_2^1(f(2))), P_{L_2}(f_2^3(f(2))), ..., P_{L_2}(f_2^{m+1}(f(2)))\})) =: A_2^{m+1}(Z_2) \\
&\in \cdots \\
P_{Z_{m+1}}(f) &\in int(conv(\{P_{L_{m+1}}(f_{m+1}^1(f(m+1))), P_{L_{m+1}}(f_{m+1}^2(f(m+1))), ... \\
&\qquad ....., P_{L_{m+1}}(f_{m+1}^m(f(m+1)))\})) =: A_{m+1}^{m+1}(Z_{m+1})
\end{aligned}$$

Since for all $l \neq j \leq m$, $P_T^s(f_l^j(f(l))) = P_T(f_l^j(f(l)))$ for all T such that $f_l^j(f_l) \in \Delta_T$, we have also $P_{Z_l}^s(f) \in A_l^{m+1}(Z_l)$ for all $l \in M$.
For $Z = LCM(Z_1, ...Z_{m+1})$ Lemma 2.2 implies that $P_Z(f), P_Z^s(f) \in A_l^{m+1}(Z)$ for

all $l \in M$, i.e. $P_Z(f), P_Z^s(f) \in \bigcap_{l \in M} A_l^{m+1}(Z)$. By Lemma 2.1 we have that for all $l \in M$ the sets $A_l^{m+1}(Z)$ consist of identical $(P^j_{max_{l\in M} f(l)Z})_{j \in M}$ (with (different) positive weights after evaluation of $P_Z(f_i^j(f(i))) = \lambda_j P^j_{max\{f(i),(1-f(i))\}Z} + (1-\lambda_j)P^i_{max\{f(i),(1-f(i))\}Z}$ for particular $\lambda_j \in (0,1)$).

This implies that determining $\bigcap_{l \in M} A_l^{m+1}(Z)$ means solving the $(m+1) \times (m+1)$ system of linear equations. We know that $|\bigcap_{l \in M} A_l^{m+1}(Z)| \geq 1$, since $P_Z^s(f)$ and $P_Z(f)$ are included in the intersection. The claim of $P_Z(f) = P_Z^s(f)$ is proofed if we can show that $\bigcap_{l \in M} A_l^{m+1}(Z)$ is a singleton.
We will proof this by contradiction:
Assume that $P_Z(f) \neq P_Z^s(f)$, then the line $g := (P_Z(f), P_Z^s(f))$ has to be contained in $A_l^{m+1}(Z)$ for all $l \in M$. Hence this line g must intersect two of the faces

$H_j := conv(\{(P^k_{\max_{i \leq m} f(i)Z})_{k \in M \setminus \{j\}}\})$ (for $j \in M$).
W.l.o.g. let these two faces be named H_u, H_v for some distinct $u, v \in M$. But then for all $l \in M$ $A_l^{m+1}(Z)$ has to intersect with these two faces H_u, H_v. We will show that this is not true. Observe that each $A_l^{m+1}(Z)$ intersects with all $(H_j)_{j \neq l \in M}$. Further, observe that applying the successive intersection, we get for $t \leq m+1$ $\cap_{j=1}^t A_j^{m+1}(Z) \cap \{H_1, ...H_t\} = \emptyset$, which implies that there is no faces H_j for any $j \in M$ that intersects with $\cap_{j=1}^{m+1} A_j^{m+1}(Z)$, i.e. there exist no faces such that all $A_l^{m+1}(Z)$ intersect them.
Hence there cannot exist a line g such that $g \in A_l^{k+1}(Z)$ for all $l \in M$, which implies that there cannot be more than one unique element in the intersection of all $(A_l^{m+1}(Z))_l$, i.e. $\cap_{l \in M} A_l^{m+1}(Z) = P_Z^s(f) = P_Z(f)$. By Lemma 2.2 we get $P_T(f) = P_T^s(f)$ for all T such that $f \in \Delta_T$, which completes the proof of the Theorem 2.3 and hence also Theorem 2.1.

□

3 Limited Attention in Case Based Belief Formation

Abstract

An agent wants to derive her belief over outcomes based on past observations collected in her database (memory). There is well establish evidence in the psychology and marketing literature that agents consistently fail (or choose not) to process all available information. An agent might be constraint to pay attention (recall) and consider only parts of her potentially available information due to unawareness, cognitive or psychological limitations or intentionally for effort-efficiency. Based on this insight, we axiomatize a two-stage belief formation process in which in a first step agents filter ((un)intentionally) the available information. In a second step individuals employ the remaining observations to express a belief. We impose cognitively and normatively desirable properties on the filtering process. The axioms on the belief formation stage describe the relationship between databases and their induced beliefs. The axiomatized belief induced by a filtered databases is representable by a similarity weighted average of the estimations induced by each past attention-grabbing observation. An appealing application is a satisficing filter that induces a filtered belief that relies only on past experiences that are sufficiently relevant for a current problem. For the specific situation that agents (are able to) always pay attention to all available information, our axiomatization coincides with the axiomatization of belief formation in Billot et al. (Econometrica, 2005).

3.1 Introduction and motivation

In many situations agents need to evaluate uncertain consequences of their actions. In order to compare different potential consequences agents need to assign likelihoods to these outcomes. How can individuals form (probabilistic) beliefs over outcomes?

Traditionally, economic theory models uncertainties in a state space representation a la Savage (1954) and Bayes and derive a subjective prior based on observable actions of an agent. This implicitly requires that an agent already posses a subjective prior belief, which is expressed by her observable actions. However, the Savage and Bayesian approach does not help an agent to find or form a prior explicitly, for instance by incorporating pieces of information directly into a belief formation. In particular in situations in which an agent might not be able to condense her insufficient or too complex information into a consistent state space, their normatively appealing and convincing approach to endogenously derive a belief is not feasible.[1]

We will consider such an environment and axiomatize a belief formation that allows to take directly into account the available information. This is strongly related to the aim of (asymptotic) statistical inference, where from data a distribution is derived. However, in this chapter we give a behavioral foundation for a belief formation in "non-asymptotic environments" that are characterized by heterogenous and limited information gathered in a list or database.

The impact of data and experience on the formation of a probabilistic belief was examined initially by the axiomatization of Billot et al. (2005) (BGSS from now on).[2] The axiomatizations of BGSS and related ones of Eichberger and Guerdjikova (2010) (EG) (for ambiguous multiprior beliefs) and Bleile (2014a) (precision dependent cautious beliefs) yield that a belief induced by a database is a similarity weighted average of the estimations induced by **all** observed cases in the database.[3] Thereby similarity weights capture different degrees of relevance of the potentially very heterogenous information.

A common shortcoming of these approaches is that an agent is obliged to take into consideration and account **all** past observations in her database. This precludes reasonable situations in which an agent might want to neglect, does miss or just forgets some pieces of information that would be in principle available. Our work relaxes this drawback of "compulsory" paying attention to all obtainable information. For this purpose we extend the mentioned axiomatic approaches (in particular BGSS) by adding a component of limited attention or consideration regarding available information.

A traditional and widely accepted assumption in economic theory is that gaining more information is beneficial and leads to improved actions. In this way, it is usually assumed that agents incorporate and take into account all available pieces of information.[4] However, the assumption of full attention and consideration of all

[1]For more details on the difficulties of the Savage and Bayesian approach, see e.g. Gilboa et al. (2012) and Chapter 1 and Section 2.1.

[2]The framework is based on Case based Decision Theory and its application to prediction problems (Gilboa and Schmeidler (1995, 2001, 2003)).

[3]Basically, the three axiomatizations differ in the way the estimations are made.

[4]In addition, it is also a key assumption in all traditional revealed preference approaches.

available information requires that agents are aware of it, perceive it (unbiased) and eventually are able to process it without any cognitive and psychological constraints.[5]

The idea and concept of limited attention goes back to the seminal studies in psychology of Miller (1956), in which he identified limited cognitive abilities in processing information as the source for incomplete consideration, especially deficits and constraints in parallel (simultaneous) processing of information. Since then, mounting evidence in psychology and marketing shows that agents process and restrict attention to only a small fraction of the overall available information and consistently fail to consider all potentially available information due to their limited attention span (e.g. Broadbent (1958), Stigler (1961), Pessemier (1978), Hauser and Wernerfelt (1990), Chiang et al. (1998)). Often agents employ (implicitly) a multi-stage process to assign different degrees of attention to specific pieces of information (Bettman (1979)). In an initial rough filtering or screening stage agents pre-selects these elements that receive (or are worth to capture) full attention and consideration. In the literature this set of "surviving" elements is called a consideration set (Wright and Barbour (1977), Bettmann (1979), Roberts and Lattin (1991)).

A formation of a consideration set might emerge for many reasons. Cognitive constraints in parallel processing of information and unawareness of the presence of information (due to complexity, size, sequential processing or search) might cause an unintentional filtering (Miller (1956), Nedungadi (1990), Schwartz (2004)). The formation of a consideration sets as a (unintentionally) reply to avoid cognitive overload has been also studied in economic problems, e.g. recently Masatlioglu et al. (2012) axiomatized choices under (unintentional) limited attention.

In contrast, a consideration set can also be the result of a purposeful strategic elimination process. Agents often use (heuristic) filtering procedures to screen information rapidly and roughly before engaging into a costly and detailed evaluation (e.g. Wright and Barbour (1977), Gensch (1987), Nedungadi (1990), Gigerenzer et al. (1999), Hauser (2013)). Usually, these heuristics are noncompensatory cutoff or satisficing rules that allow for an uncomplicated "effort-efficient" comparison. This approach has recently gained prominence in economics (in particular in decision theory), e.g. through the works of Lleras et al. (2010) and Eliaz and Spiegler (2011a,b).

Another reason for the emergence of a consideration set relies on mounting evidence from psychology showing that often non-objective criteria like value systems,

[5] According to Simon (1959, p.272) perception and cognition intervene between subjective view and the objective real world. In this context perception is often referred to as a "filter", where filtering can not only be seen as a passive, but also as an active selection process involving exclusion of almost all that is not within the scope of attention.

subjective motives or aversions, etc. restrict the attention of agents. Recent work has modeled these subjective and psychological biases ranging from overwhelming temptation (Gul and Pesendorfer (2001), Dekel and Lipman (2012)), rationalization (Cherepanov et al. (2013)), status quo bias (Masatlioglu and Ok (2005)), routes (Apesteguia and Ballester (2013)) and reason based choice (Lombardi (2009), De Clippel and Eliaz (2012)).

In this chapter, we want to incorporate the formation of a consideration set induced by limited attention (consideration) as an intermediate stage into a belief formation process. In this way, our agent is not obliged to take into account all potentially available information, but might base her belief only on (survived) filtered information in the consideration set. In order to illustrate the basic idea and plausibility of such a two stage belief formation process we modify the doctor example of BGSS.

A doctor needs to evaluate different outcomes of a treatment. She has some working experience or access to some medical database $D = (c_1, ..., c_l)$, where she recorded in a case $c_i = (x_i, r_i)$ the vector of characteristic of a patient i, $x_i \in X$ (e.g. age, gender, weight, blood count) and the observable outcome of the treatment $r_i \in R$ (e.g. better, worse, adverse effects). A new patient characterized by x enters her office and using a medical record D, the doctor wants to derive a belief $P_x(D) \in \Delta(R)$ over potential outcomes in R. She might apply an empirical frequency and use only a part D_x of the database D, which contains only cases $c = (x, r_c)$ of patients with *"identical"* characteristic x compared to the current patient:

$$\text{"Frequentist"}: \qquad P_x(D) = \frac{\sum_{c \in D_x} \delta_{r_c}}{|D_x|}.$$

However, if the database contains not sufficiently many of these *"identical"* patients x, she might want to include also *"similar"* patients. She judges the degree of similarity between patients x and x' by $s(x, x') \in \mathbb{R}_+$. Further, she might induce from a case $c = (x_c, r_c)$ not only a point estimate δ_{r_c} on the realized outcome, but derives a more general estimate $P^c \in \Delta(R)$ on likelihoods of particular (related) outcomes and forms the belief as axiomatized in BGSS by:

$$\text{"BGSS-belief"}: \qquad P_x(D) = \frac{\sum_{c \in D} s(x, x_c) P^c}{\sum_{c \in D} s(x, x_c)}. \tag{3.1}$$

However, if the database D is long, complex or retrieved partly from her memory, the doctor might not want or is just not able to (recall) pay attention to and take into account all potential cases (patients) in the database D. She filters out some patients contained in her record D with specific features $\Gamma(D) \subseteq D$. An intuitive example is a **similarity satisficing** procedure, in which she considers only sufficiently relevant patients that surpass a threshold of similarity s^*, i.e. $\Gamma(D) = (c \in D | s(x, x_c) \geq s^*)$:

$$\text{"Filtered-belief":} \qquad (P_x \circ \Gamma)(D) = \frac{\sum_{c \in D} \mathbf{1}_{\{s(x,x_c) \geq s^*\}}(c) s(x, x_c) P^c}{\sum_{c \in D} \mathbf{1}_{\{s(x,x_c) \geq s^*\}}(c) s(x, x_c)} \tag{3.2}$$

The filtered belief formation based on the similarity satisficing principle (under additional restrictions on the threshold value s^*) represents a special case of the general result we obtain in our representation theorem in Section 3.5.
Roughly, our filtered belief formation consist of two stages, in which initially a subjective (specific) filter process "selects" the information that builds the consideration set. In the second step, the agent forms her belief based on the remaining un-eliminated information in her consideration set.

One might be tempted to interpret such a filtered belief as the belief (3.1) of BGSS, based on an already (exogenously) ex-ante and independently filtered database $\Gamma(D)$. However, such an separation of filtering and belief formation would exclude plausible and appealing filters based on similarities (as in (3.2)), since the similarity values are endogenously derived in the belief formation. Moreover, in our axiomatization both stages are merged by an axiom that focusses on the relationship between filtered databases and their induced beliefs.

The initial filtering process captures appealing and desirable psychological properties that are rooted in psychology and marketing literature. The main property is the well known and accepted consideration property. It is based on the idea and evidence that if a case is considered in a database, i.e. is attention grabbing, then it should attract attention also in all of its subdatabases, since it faces less competition for attention by fewer pieces of information.[6] Further, we make some assumptions on the cognitive ability of agents and assume that an agent is able to pay attention to at least k ($k \geq 3$)-many available different pieces of information. A slightly more demanding characteristics requires that order and frequency in which information appears in a databases does not affect the level of attention an agent attributes to it. Basically that means that pieces of information are per se attention grabbing and not due to their specific position or a sufficiently high number of appearance.

The second layer of a filtered belief formation concerns the axiomatization on the belief level. The normatively reasonable axioms follow the basic intuition of the axioms in BGSS, but are modified to capture the previous filtering stage. We

[6] This property is also implied by the appealing axiomatic rationalization theory of Cherepanov et al. (2013) (based on specific kinds of rationales).

generalize the Concatenation Axiom of BGSS in order to capture the previous filtering process on the involved databases. The original Concatenation Axiom says that a belief induced by a combination of two databases is formed as an average of the beliefs that are induced by each of these databases separately. We cannot directly translate this to filtered database, since concatenations of already filtered databases can differ enormously from the result of filtering the concatenation of the two underlying databases (before they were filtered).[7] Thus, in order to ensure a reasonable averaging of the induced beliefs in the spirit of the axiom, we require consistent relationships between the involved filtered databases. Another (implicit enforced) axiom ensures that the order in which information appears is irrelevant for the belief it induces.

As a result, our filtered belief formation can be represented as a similarity weighted average of the estimates induced by each case that the agent actually pays attention to, i.e. of those that survive the filtering. Hence the representation coincides with BGSS if the agent does not filter any information, but takes into account always all available information.

Apart from the appealing intuition of filtering according to similarities (as in (3.2)), the filtering process can be any general arbitrary process that satisfies the required properties. Various subjective and psychological motives, constraints, biases and justifications can be employed as elimination criteria.[8] In particular, many recently developed multi-criteria decision procedures include elimination procedures to (implicitly) form a consideration set. The literature varies in the kind of criteria that are employed, e.g. using (compositions of) rationale(s) (sequentially) to eliminate alternatives (Manzini and Mariotti (2007, 2012a), Apesteguia and Ballester (2013), Houy (2007, 2010), Houy and Tadenuma (2009), Horan (2013)), focusing only on subjectively justifiable alternatives (Cherepanov et al. (2013), Gerasimou (2013)), considering only alternatives belonging to un-dominated or best category(ies) (Manzini and Mariotti (2012b)) or considering only top N eye-catching elements according to some exogenously given order or ranking (Salant and Rubinstein (2008)). Most of these approaches form consideration set in a way that can be interpreted as seeking for a reasons to select (based on Shafir et al. (1993) and Tversky (1972) and more related also in Lombardi (2009) and de Clippel and Eliaz (2012)).

Another approach to form a consideration set can be seen in a satisficing procedure

[7]This is unproblematic for unfiltered databases, where the result of their combinations is directly clear.

[8]For instance salience, familiarity, "roughly" identical features, heuristics, fairness, extremeness aversion, limited memory, reference effects, etc.

(Simon (1959) and more related Tyson (2008, 2013), Papi (2012), Manzini et al. (2013a)). The resulting consideration set contains only these elements which surpass a (endogenously) given threshold level according to some criteria. This is close to our motivating similarity satisficing example in (3.2).

However, for our purpose the most interesting paper is the axiomatic rationalization theory of Cherepanov et al. (2013), since our assumed consideration property (of the filter) is a direct consequence of their normatively and descriptively appealing rationalization process. Modifying their justification procedure in order to cope with our other filter properties yields a corollary of our main representation result that allows for an interpretation in terms of a "rationalized" filtered belief.

The next section gives the general framework for the two stage filtered belief formation. In Section 3.3 we introduce and discuss the properties on the filtering process. Section 3.4 deals with the axioms on the belief formation induced by a filtered database. The main representation theorem and a sketch of the proof is presented in Section 3.5 and we derive as a corollary the similarity satisficing belief process. Section 3.6 relates the filtering process to the recently developed multi-criteria/stage decision procedures. In particular we exemplarily modify the choice model of Cherepanov et al. (2013) to a filter process in our terms. An interpretation of the resulting representation in terms of multi-similarities is given. Section 3.7 concludes. All proofs can be found in the last section.

3.2 The model

In this section, we introduce the case-based information framework and the basic building blocks of our belief formation based on filtered information. Further, we introduce some definitions and notations necessary for our approach.

3.2.1 Database framework

A basic case $c = (x, r)$ consists of a description of the environment or problem $x \in X$ and an outcome $r \in R$, where $X = X^1 \times X^2 \times .... \times X^N$ is a finite set of all characteristics of the environment, in which X^j denotes the set of possible values features j can take. R denotes a finite set of potential outcomes, $R = \{r^1, ..., r^n\}$
The **ordered** set $C \subseteq X \times R$ consists of all $m \geq 3$ basic cases, i.e. $C = \{c^1, ..., c^m\}$. A database D is a sequence or list of basic cases $c \in C$. The set of databases D consisting of L cases, i.e. $D = (c_1, ..., c_L)$ where $c_i \in C$ for all $i \leq L$, is denoted by C^L and the set of all databases by $C^* = \cup_{L \geq 1} C^L$, including the empty database $\emptyset$. The description of databases as sequence of potentially identical cases allows

multiple observation of an identical case to be taken into account and treated as an additional source of information.
For a database $D \in C^*$, $f_D(c)$ denotes the relative frequency of case $c \in C$ in databases D.
The concatenation of two databases $D = (c_1, c_2, ..., c_L) \in C^L$ and $E = (c'_1, c'_2, ..., c'_T) \in C^T$ (where $c_i, c'_j \in C$ for all $i \leq L, j \leq T$) is denoted by $D \circ E \in C^{L+T}$ and is defined by $D \circ E := (c_1, c_2, ..., c_L, c'_1, c'_2, ..., c'_T)$.
In the following we will abbreviate the concatenation or replication of L-times the identical databases D by D^L. Specifically, c^L represents a database consisting of L-times case c.
For any $D \in C^*$ the diversity of a database D is given by $div(D) := |\{D\}|$, where as usual $\{D\}$ denotes the set of different cases contained in database D. So $div(D)$ gives the number of different cases contained in database D.
We need to translate some relations from sets to the list framework.

Definition 3.1
(i) The $\in$-relation on databases is defined by $c \in D$ if $f_D(c) > 0$.
(ii) The $\subseteq$-relation on the set of databases C^* is defined by $D \subseteq E \Leftrightarrow f_D(c)|D| \leq f_E(c)|E|$ for all $c \in C$. We will call such databases to be nested.
(iii) The $\cap$-relation on databases is given by $D \cap E = ((c^{\min\{f_D(c)|D|, f_E(c)|E|\}})_{c \in C})$
(iv) Two databases D and E are disjoint if for all $c \in C$: $c \in D$ if and only if $c \notin E$.

The definitions are basically independent of the order of cases in the databases. Note however that the definition of $\cap$-relation in (iii) is very specific, since the order of C is transferred, i.e. by intersection a specific order (on C) is induced.[9]

3.2.2 Filter

In the literature so far, a filter on C^* is usually defined as a set function $\Gamma : C^* \to C^*$, such that for all $E \in C^*$ $\emptyset \neq \Gamma(E) \subseteq E$. In this way, implicitly the information E is filtered from the perspective of the available information E. However, it does not cover and specify how any subdatabase D of E is filtered, while knowing database E. In particular a case c in subdatabase D might attract attention in D if the agent does not take into account the additional information in E. But she might not pay attention to this case when having the larger database E in mind (and its relevant cases). Thus, the underlying perspective from which a filtering occurs might affect the assigned attention (see Section 3.3.1 for further discussion). Our

[9] In contrast to intersections of sets, where orderings are immaterial, intersection of databases do require some assumption on resulting orderings.

filter determines as well, how any subdatabase D of E is filtered from the perspective of E. Formally, a function $\Gamma : X \times C^* \times C^* \to C^*$ such that for all $x \in X$ and $D \subseteq E \in C^*$, $\Gamma(x, D, E) := \Gamma^E(x, D) \subseteq D$ is called a filter on (sub)database D induced by (perspective) database E and given a problem x. We use the notation $\Gamma^E(x, D)$ to highlight the roles of the two database. Basically the (sub)database D is screened from the perspective of (or having in mind) the richer information E.

3.2.3 Induced belief

For a finite set S, $\Delta(S)$ denotes the simplex of probability vectors over S and for $n \in \mathbb{N}$ Δ^n denotes the simplex over the set $\{1, 2, ..., n\}$.

In the axiomatizations of BGSS, EG and Bleile (2014a) (or Chapter 2), an agent will form a belief over outcomes $P(x, D) \in \Delta(R)$ for a certain problem characterized by $x \in X$ using her information captured in a database $D \in C^*$, i.e. $P : X \times C^* \to \Delta(R)$.

In the current approach a filtered belief is formed based only on parts of the information captured in D and that is filtered from the perspective of richer information $E \in C^*$, i.e. filtered belief $(P \circ \Gamma) : X \times C^* \times C^* \to \Delta(R)$ and $(P \circ \Gamma)(x, D, E) = (P \circ \Gamma^E)(x, D)$ for $D \subseteq E$ (with slight abuse of notation). In this sense the filtered belief $(P \circ \Gamma^E)(x, D)$ is induced by a nested pair of databases D and E and can be interpreted as the belief over outcomes induced by database $D \in C^*$ seen through a filter that rests on perspective E (given problem $x \in X$). Hence, a filtered belief is a two stage process of filtering followed by a belief formation.

Technically, the filtered belief induced by the pair of nested databases D and E coincides with the BGSS belief based on an a priori already filtered database $\Gamma^E(D)$.[10] However, as discussed in the introduction, a priori filtering would separate the filter procedure and the belief formation process that would exclude desirable applications based on endogenously derived similarity values, like the similar satisficing behavior (as discussed in our motivating example (equation (3.2)). Further, the nice axiomatization of BGSS based on relationships between filtered database and their induced beliefs would need to be defined on filtered databases as well, which merges both stages again. However, for our motivation and purpose the most intuitive (and desirable) filtered belief formation is based on a single database, i.e. the filtered belief $(P \circ \Gamma)(x, D, D)$ induced by database D.

Throughout the paper the problem x is fixed, therefore x is often suppressed in the following, i.e. $(P \circ \Gamma^E)(x, D) = (P \circ \Gamma^E)(D)$.

[10] Basically meaning that $(P \circ \Gamma^E)(D) = P(\Gamma^E(D))$.

3.3 Filter definition and properties

Instead of defining explicitly a procedure how to filter information we are more general and rather impose natural, well accepted and established properties in psychology and marketing. These properties are normatively and descriptively compelling in several situations and are indeed true for many heuristics people actually use in real life to screen their "information set". In particular many of the recently developed multistage decision models contain a wide variety of (endogenous and exogenous) filter procedures that satisfy our filter properties, which supports their relevance and generality (see Section 3.6). In this sense, we can also interpret our filter as a choice correspondence in which elements surviving the pre-choice process form a consideration set of acceptable and relevant information. This (filtered) consideration set represents the underlying basis for a filtered belief formation process.

3.3.1 Basic properties of a filter

In the last section, we mentioned briefly our concept of a filter. More precisely:

Definition 3.2 ***Filter***
A set function $\Gamma : C^* \times C^* \to C^*$ *is called a* ***filter on D induced by (perspective) database*** E *if for all* $D, E \in C^*$ *such that* $D \subseteq E \in C^*$ *it holds*
(i) $\Gamma(E, E) =: \Gamma^E(E) =: \Gamma(E) \neq \emptyset$ *and*
(ii) $\Gamma(D, E) =: \Gamma^E(D) = \Gamma(E) \cap D$.

Note that by definition $\Gamma^E(D) \subseteq D$ holds for all $D \subseteq E$, but it does not imply that $\Gamma^E(D) \neq \emptyset$.

As mentioned already, the traditional definition of a filter is based only on one database, i.e. $\emptyset \neq \Gamma(D) \subseteq D$, which is covered as well by our definition for $D = E$. However, our definition also specifies how all subdatabases D of a given database E are filtered. The knowledge of a database E affects the evaluation of the relevance of cases in the subdatabase D, since the "attractiveness" of pieces of information might vary strongly with different perspectives, available information or knowledge. Basically, the aim is to capture that a case $c \in D \cap E$ may attract attention in database D, but not in database E, i.e. $\Gamma^D(c) = c$, but $\Gamma^E(c) = \emptyset$. Thus, the perspective from which filtering occurs is an important characteristic to determine whether some piece of information survives a filtering process or not. Our definition requires a consistent filtering of parts of the available information, i.e. knowing database E, an agent should pay only attention to those cases in a (smaller) subdatabase D that (already) captured attention in the richer database E.

We will discuss the three properties the definition proposes, namely the filtering, its specific content and induced ordering.

There are many reasons, why an agent demonstrates partial attention to available information. Cognitive limitations and constraints in parallel processing of complex and large information sets is often associated with an unconsciously allocation of attentional resources (Anderson (2005) pp.72-105). This kind of cognitive unawareness can lead to a formation of a consideration sets as well as an intentional heuristic elimination procedure in order to explicitly simplify the information set in a cost or effort efficient way, while still keeping the most relevant information. In addition, consideration sets are often constructed unrelated to the objective features of the elements in their information sets, e.g. subjective constraints, like motives, mood, rationales, value systems, biases, attitudes might affect acceptability, attractiveness or moral justification of an object and affects its assigned attention level. In sum, filtering a set of available information is a natural and plausible behavior under many circumstances and for many reasons.

The nested perspective-structure in our definition of the induced filter, i.e. $\Gamma^E(D) = \Gamma(E) \cap D$ for $D \subseteq E$, states that a subdatabase D is filtered from the perspective of a database E. In this specific form an agent already processed and filtered out the information in E (resulting in $\Gamma(E)$) that she deems relevant, wants to consider or is able to pay attention to. Since an agent actually needs to process and screen only information in a subdatabase D, she will pay attention to all elements in D that already grabbed her attention and are still in her mind, i.e. $\Gamma(E) \cap D$. Basically D is not filtered independently but compared with the information that is attention grabbing in E. Put differently, D is filtered through E- colored glasses and not by inspecting elements in D in detail. However, it is important to stress that this definition applies only to situation in which both databases are processed "simultaneously". If databases are processed independently, such an interwoven filtering (from the perspective of one) is not feasible or desirable. In particular, in the following we will introduce a (consideration) property which specifies how independently filtered nested databases are related.

Finally, we are concerned with the ordering of the resulting filtered database, which stems from the definition of the intersection and the ordering on C. Of course, it is restrictive to assume ad hoc a specific arbitrary ordering of a resulting filtered database, since amongst other things this may depend on orders of D or/and E. However, in course of the axiomatization of a belief formation, we would adopt the Invariance Axiom of BGSS, which states that only the content and not the ordering of cases is important for an induced belief. From that perspective any reordering of

a resulting filtered database would lead to the same induced belief, which leaves the definition and its induced ordering rather harmless.
Moreover, the implied $\Gamma^E(D) = \Gamma^E(\pi(D))$ for all $D \subseteq E$ and any reordered database $\pi(D)$ will be further generalized by the following property.

Definition 3.3 ***Filter order invariance***
For all $L \in \mathbb{N}$ and $D \in C^L$ and any permutation π on $\{1, .., L\}$, let $D = (c_1, .., c_L)$ and $\pi(D) = (c_{\pi(1)}, ..., c_{\pi(L)})$. A filter Γ is ***order invariant*** *if it holds that $c \in \Gamma(D)$ if and only if $c \in \Gamma(\pi(D))$.*[11]

Basically it states (in close relationship to BGSS Invariance Axiom for beliefs) that the order of the cases is immaterial for the resulting filtering of the cases, i.e. only the content of a database matters. Information should be attention grabbing per se and not due to its specific position in the database. From a first sight the property seems to be rather restrictive since agents appear to be able to consider all cases in the database "simultaneously" without any order biases, like first impressions and recency effect (see Rubinstein and Salant (2006)). However, if some of these effects are important, then they can be captured by a more elaborated description of the cases. For example, its description can include time or the order in which it was observed.
The filter order Invariance implies that in combination with the definition of intersections of databases

$$\Gamma^E(D) = \Gamma^{\pi'(E)}(\pi(D)) \text{ for all } D \subseteq E \text{ and any reordered databases } \pi(D) \subseteq \pi'(E) \quad (3.3)$$

It will imply the identity of beliefs induced by all possible combinations of reordered databases $\pi(D)$ and $\pi'(E)$ that enter the filtering process.

A closely related property states and ensures quite naturally that if a case catches attention then all other cases of this type are attention grabbing as well.

Definition 3.4 ***Equal treatment of information***
A filter Γ on $C^* \times C^*$ ***treats information equally*** *if for all $c \in D$*

$$\Gamma^D(c) = c \textit{ if and only if } \Gamma^D(c^T) = c^T \textit{ for all } T \textit{ such that } c^T \in D.$$

Basically, this means that all pieces of the same type are treated equally and either are attention grabbing or not.

[11]Note that independent of the filter definition, the property would just enforce that the content is identical. Only in combination with our definition of a filter the ordering coincide, i.e. $\Gamma(D) = \Gamma(\pi(D))$ is implied.

Another content dependent property is similarly based on the view that attention or disregard is structural in a way that per se either a piece of information is eye-catching or attention grabbing or not. Namely, the filtering process is not affected by an (sufficiently large) amount of occurrence of a piece of information.

Definition 3.5 ***Filter ignorance of repeated information***
A filter Γ satisfies the ***ignorance property*** *if for all $D \in C^*$ it holds:*
$\Gamma(D \circ c) \cap D = \Gamma(D)$ *for all $c \in D$.*[12]

A repeated appearance of a case does not influence its perception in the sense that a case attracts attention if it is relevant or outstanding in itself and not because itself or another case appears in a specific amount. The assignment of attention to a case might be altered only if another new case appears or all appearances of a specific case are removed from the database (as induced by the consideration filter property defined below). In this sense, additional observation of already known and evaluated cases do not alter (already attached) attention levels.

The property finds support in Gul et al. (2012), which examines the probabilities of choices when alternatives are duplicated. They propose, that duplicate alternatives should be identified as observational identically and should be (in a specific sense) irrelevant for the likelihoods with which an observational identical alternative is chosen. This can be related to a "pay attention"- choice in which we restrict the probabilities to pay attention to zero or one. However, duplication of evidence might affect the composition of a filtered consideration set in the list-framework of Rubinstein and Salant (2006) and hence violate our property (and their partition independence property). Of course, the number of times an element appears might have an influence on attention, for instance in a procedure that pays attention only to the most frequent element.

3.3.2 Main structural properties of a filter

The two following structural properties characterize the impact of nested databases on their induced filtered consideration set and the cognitive ability of an agent to process and pay attention to at least a minimum amount of available information.

[12] This is equivalent to: for all $D \subseteq E$ and all $c \in E$, $\Gamma^{E \circ c}(D) = \Gamma^{E}(D)$.

Consideration property

The consideration property specifies the relationship between the induced consideration sets of nested databases in a quiet naturally way. It states that if an agent pays attention to a case in a database, then also her attention is drawn to this case in all of its subdatabases. This follows the idea and evidence that elements in an information set compete for attention and need to outperform other pieces of information to be considered. If a case manages to attract attention or is salient enough in a database, then it also gains attention in a subdatabase, in which some of its rivals for attention are not anymore present. The classical example in marketing deals with the attention an agent assigns to products in a supermarket with a huge variety compared to a small neighborhood store. A specific jam catching your eye standing in front of a large supermarket shelf with fifty different jams will catch your attention also in the convenience store selling only five different sorts of jam.

Definition 3.6 ***Consideration property***
A filter Γ *on* $C^* \times C^*$ *satisfies the* ***consideration property*** *if for all* $D \subseteq E \in C^*$ *and* $c \in D$: $\quad c \in \Gamma(E)$ *implies* $c \in \Gamma(D)$.

From another point of view, if there is no reason (e.g. an (outstanding) piece of information) in the larger sample E that shadows case c, i.e. $c \in \Gamma(E)$, then it still cannot happen that case c is shadowed by any reason (e.g. any information) in the smaller subdatabase.

Apart from the "competition for attention"-explanation the consideration property can be motivated by the finding that with increasing complexity and size of a set agents reduce the amount of alternatives they consider and also lower its intensity and seriousness.[13] If the complexity of a database is caused by a difficult detailed evaluation of alternatives involving compromising and tradeoffs, an agent might find it much harder to find reasons why to "choose" to consider a case. As a consequence agents might stick to only superficial analysis, where only the very important, salient or extraordinary alternatives receive attention. With decreasing complexity of a set agents might return to a more detailed analysis that facilitates the selection of more alternatives to be worth or justified to be included into the consideration set.

[13] See e.g. Hauser and Wernerfelt (1990) and Shugan (1980).

Minimal attention span

The following property describes the cognitive ability of an agent to handle and process information. Our minimal attention property requires that an agent considers in full detail at least k-many ($k \geq 3$) available different pieces of information. More precisely, for all databases containing less than k many different information an agent takes into account all pieces of information. For more complex databases an agent is free to apply filtering techniques to process her information and neglects some pieces of information, as long as she pays attention to at least k-many different pieces of information.

Definition 3.7 ***Minimal attention span***
A filter is ***minimal attentive*** *with $k \geq 3$ if for all $D \in C^*$ such that*
(i) for $div(D) = l \leq k$ it holds $div(\Gamma(D)) = l$ and
(ii) for $div(D) > k$ it holds $div(\Gamma(D)) \geq k$.

From this perspective, the property induces that agents are cognitively sophisticated enough to handle at least k-many different pieces of information completely and with full attention. Also, agents might want to gather a certain minimal amount of information in order to evaluate and act in an informed and confident way. Thus they take into account all available information, when only few (less than k different) information is around.[14]

The property combines basically two components -minimum required amount of information and filtering in case of potential information overload - which are well supported by empirical findings. We state it in terms of a general minimal attention span k, but for our purpose $k = 3$ is sufficient and also meets empirical evidence. For example Gensch (1987) found that screening and filtering rules may be invoked for as few as four alternatives, but agents consider and rely on all information for less diverse information sets. In the marketing literature Jarvis and Wilcox (1977) examined the usual size of consideration sets and discovered that the average size of consideration sets is three to eight products, independent of the size of the initial information.[15]

So far a minimal attention property is usually not assumed in the (choice) literature in which usually no lower limit is given for the filtering stage unless a non-emptiness condition. In a special case of Lleras et al. (2010) agents have no limited

[14]One can also think about it as in Simon (1959, p.263) based on an aspiration or satisfaction level such that at least k cases are attracting interest or attention. If not, the level was too high and a reduction of the level leads to some search behavior of the agent to pay attention to more available alternatives.

[15]Miller' s insight (1956) that agents can process or remember at least seven case is also covered.

attention problem for situations of binary alternative set, i.e. their agents are always able to pay attention to both alternatives.[16] However, their version does not cover any restrictions for more general situations, as specified in our property.

The requirement of filtering a minimum amount of information constitutes a strong constraint to identify filters as choice correspondences. We will discuss this problematic issue in Section 3.6, where we interpret filtering in a choice theoretic perspective.

Definition 3.8 ***Admissible Filter***
A filter Γ on $C^* \times C^*$ is called an admissible filter if it satisfies the invariance, equal treatment, ignorance, consideration and minimal attention span property.

3.4 Axioms on level of belief formation

In this section we introduce the axioms on the stage of belief formation.
As mentioned in the section before, a natural axiom for a belief process that is already implied by the definition of intersections and the filter properties, is a version of an Invariance Axiom as in BGSS for filtered databases that reads:

Filtered Invariance Axiom (already implied)
For every $T \geq 1$, every database $D, E \in C^*$ with $D \subseteq E$. Let $\Gamma^E(D) = (c_1, .., c_T) \in C^T$ and for every permutation $\pi : \{1, ..., T\} \to \{1, ..., T\}$ and any filter $\Gamma^E(D)$, where $\pi(\Gamma^E(D)) = (c_{\pi(1)}, ..., c_{\pi(T)})$ the following holds:

$$(P \circ \Gamma^E)(D) = (P \circ \pi(\Gamma^E))(D)$$

Basically it says that an induced belief over outcomes depends only on the content of a filtered database and is insensitive to the sequence and order in which data arrives. However, by our definitions of a filter any filtered databases containing the same content have exactly the same specific ordering (according to the order on C), which makes the invariance property superfluous, since re-orderings do not occur after filtering (see eq. (3.3)). In this sense the filtered Invariance Axiom is substituted by the definition of a filter and the filter ignorance property.[17]

[16] See also Masatlioglu et al. (2012), where $\Gamma(D) \geq 2$ allows the full revelation of preferences

[17] We would have required the axiom directly if our filtering process would have allowed for different orderings of databases consisting of the same content. Technically speaking, there is no difference between restricting the filtering process to specific orders or allowing for different orderings and requiring an Invariance Axiom for beliefs.

Per se the invariance property does not allow for different impacts whether a case appears earlier or later. However, the order in which information is provided or obtained can influence the judgment strongly and may carry information by itself. One way to cope with these order effects is to describe the cases informative enough. E.g. if one wants to capture the position or time of occurrence of a case in the filtered database, one could implement this information into the description of the cases itself. Put differently, if one challenges the consequences of an invariance property, then there must be some criteria which distinguish the cases. Taking into account these differences explicitly in the description of the cases may lead the agent to reconcile with such an invariance.

Filtered Concatenation Axiom
Let Γ be a filter. For all database $D, E, F \in C^*$ such that $D \circ E \subseteq F$ there exists $\lambda \in [0,1]$ such that:

$$(P \circ \Gamma^F)(D \circ E) = \lambda (P \circ \Gamma^F)(D) + (1-\lambda)(P \circ \Gamma^F)(E)$$

where $\lambda = 0$ if and only if $\Gamma^F(D) = \emptyset$.

In the following we will call the database which emerges from concatenation of other databases as the **combined or concatenated** database, whereas the databases used for the concatenation will be called **combining or concatenating** databases.

The filtered Concatenation Axiom says that a filtered belief induced by a concatenated database is a weighted average of the filtered beliefs induced by their respective combining databases. The axiom captures the idea that the belief based on the combination of two databases cannot lie outside the interval spanned by the beliefs induced by each combining database separately. Intuitively it can be interpreted in the following way (stated from an exclusion point of view): if the information in any database induces an agent's belief not to exclude an outcome r, then the outcome r cannot be excluded by the belief induced by the combination of all these databases.[18] Alternatively, if a certain conclusion is reached given two filtered databases, the same conclusion should be reached given their filtered union.

However, in order to sustain the normative appealing interpretation of averaging (filtered) beliefs, the filtered concatenation of two databases must coincide with

[18]Of course the axiom is stronger in the sense that it not only requires that the probability of such an r is positive, but it should lie between the minimal and maximal assigned probabilities induced by the combining (filtered) databases.

the concatenation of these two filtered databases. This is achieved by employing the common perspective $F \supseteq D \circ E$ according to which all involved databases are filtered, i.e. Γ^F. Otherwise it would not be in general reasonable to require the existence of such an average of beliefs, since the elements surviving the filtering process for each single database might differ from the elements surviving the elimination of the database generated by the combination of the two.[19] In this situation, it would be implausible and unreasonable to determine a relationship between the induced filtered beliefs. The required structure ensures that a filtered belief induced by the concatenated database relies on information that is also employed in the filtered beliefs induced by the single concatenating databases and thus allows for an interwoven filtering and belief formation.

Moreover, another important reason for requiring this specific common perspective is based on the motivation that an agent should filter from the perspective of the richest information available, which is at least the concatenated database $D \circ E$, i.e. $F \supseteq D \circ E$. This is very reasonable and natural, since the agent processed already at least the information in the concatenated database due to the fact that she actually wants to form a (filtered) belief based on this filtered database. In this way she cannot remove (intentionally) some (gained or experienced) information for filtering the concatenating database from a less informative perspective.[20] In this way, the concatenated databases represent a quite natural choice as a "smallest" perspective from which the filtering process is initiated.

Collinearity Axiom
No three elements of $\{((P \circ \Gamma^c)(c))_{c \in C}\}$ are collinear.

Technically speaking this axiom allows to derive a unique similarity function (in combination with the other axioms), but it has also some reasonable intuition. Roughly it states that the estimation based on a case is never equivalent to the combined estimations based on two other cases. Hence, a case is always informative in the sense that no combination of two other cases can deliver the same estimate and would make this case "redundant".

[19] In an unfiltered concatenated database any information appear in either of their (unfiltered) concatenating databases and find weight in their induced beliefs. However, for filtered databases $(P \circ \Gamma^{D \circ E})(D \circ E) = \lambda(P \circ \Gamma^D)(D) + (1-\lambda)(P \circ \Gamma^E)(E)$ is meaningless, since the relationships between $\Gamma(D \circ E)$ and the parts $\Gamma(D)$ and $\Gamma(E)$ are unclear.

[20] This is even impossible if the filtering occurs unconsciously, since the information entered already her mind by the nature of the task.

3.5 Representation Theorem

Theorem 3.1
Let there be given a function $(P \circ \Gamma) : C^* \times C^* \to \Delta(R)$*, where* $P : C^* \to \Delta(R)$ *and* Γ *be an admissible filter on* $C^* \times C^*$*. Let* $(P \circ \Gamma) : C^* \times C^* \to \Delta(R)$ *satisfy the Collinearity Axiom.*
Then the following are equivalent:
(i) The function $(P \circ \Gamma)$ *satisfies the filtered Concatenation Axiom*
(ii) There exists for each $c \in C$ *a unique* $P^c \in \Delta(R)$*, and a unique -up to multiplication by a strictly positive number- strictly positive function* $s : C \to \mathbb{R}_+$*, such that for all* $D \subseteq E \in C^*$ *such that* $\Gamma^E(D) \neq \emptyset$

$$(P \circ \Gamma^E)(D) = \frac{\sum_{c \in D} s(c) 1_{\Gamma(E)}(c) P^c}{\sum_{c \in D} s(c) 1_{\Gamma(E)}(c)}. \tag{3.4}$$

Rough sketch of the proof
The necessity part is straightforward calculation. The sufficiency part follows the rough structure of the proof of BGSS and Bleile (2014a) (or Chapter 2), but differs in the crucial arguments. The idea is to transform the framework from the space of databases to the space of frequency vectors that is structural more tractable, i.e. the filtered belief based on databases $(P \circ \Gamma^E)(D) = \frac{\sum_{c \in D} s(c) 1_{\Gamma(E)}(c) P^c}{\sum_{c \in D} s(c) 1_{\Gamma(E)}(c)}$ for $D \subseteq E \in C^*$ translates to frequency vectors $f_D \subseteq f_E$ by $(P \circ \Gamma^{f_E})(f_D) = \frac{\sum_{j \leq m} s_j \Gamma_j^{f_E}(f_D) P^j}{\sum_{j \leq m} s_j \Gamma_j^{f_E}(f_D)}$. In order to show that this is viable we exploit some properties of the filter and the Concatenation Axiom.
The essential part of the proof is to derive the similarity weights $(s_i)_{i \leq m}$. This will be shown inductively over $|C| = m$ and $div(f_E) \leq m$.

Step 1: Base case for the induction, i.e. $|C| = m = 3$ and $div(f_E) \leq 3$, w.l.o.g. $C = \{c_1, c_2, c_3\}$, i.e. aim to find s_1, s_2, s_3.
For pairs $(f_D, f_E) \in \Delta \times \Delta$ (i.e. such that $div(f_E) \leq 3$) the properties of the filter (in particular the minimal attention span) induce that the filtering stage disappears and the axioms coincide with BGSS. Thus, the same steps (using simplicial partitions) as in BGSS (or Chapter 2) will show the above representation for all pairs (f_D, f_E) such that $f_D \subseteq f_E$ and $div(f_E) \leq 3$.

Step 2: $|C| = m > 3$ and $div(f_E) \leq m$.
As in BGSS or Chapter 2, we can show (again using the minimal attention property) that the similarity weights derived in Step 1 for any set of basic cases $C = \{c^i, c^j, c^k\}$

are independent of the triplet $\{i, j, k\}$ and thus we can define for all $f_D \subseteq f_E \in \Delta$

$(P \circ \Gamma^{f_E})_s(f_D) = \frac{\sum_{j \le m} s_j \Gamma_j^{f_E}(f_D) P^j}{\sum_{j \le m} s_j \Gamma_j^{f_E}(f_D)}$ using the derived $s = (s_1, .., s_m)$

The aim is to show $(P \circ \Gamma^{f_E})(f_D) = (P \circ \Gamma^{f_E})_s(f_D)$ for all $f_D \subseteq f_E \in \Delta$ via induction over m and using Step 1 ($m = 3$) as base case.
a) For all pairs (f_D, f_E) such that $div(f_E) = m$ and $div(f_D) < m$ the "idempotence" of the filter (i.e. $\Gamma(\Gamma^{f_E}(f_D)) = \Gamma^{f_E}(f_D)$) is used to exploit the induction assumption.
b) For $div(f_D) = div(f_E) = m$ we adopt (with different reasoning) the construction used in BGSS, i.e. $f_D = \alpha q^j + (1 - \alpha) f(j)$ (for some $\alpha \in (0, 1)$, where $f(j)$ denotes the point on $conv\{(q^l)_{l \in \{1,...,m\} \setminus j}\}$ on the line through f_D and j-th unit vector q^j. The result of a) allows to apply the filtered Concatenation Axiom, i.e. $(P \circ \Gamma^{f_D})(f_D) = \lambda (P \circ \Gamma^{f_D})(q^j) + (1 - \lambda)(P \circ \Gamma^{f_D})(f(j))$. The filter properties (minimal attention and consideration property) ensure that there exist at least three q^j such that $\Gamma^{f_D}(q^j) \neq 0^m$ and therefore there exist three P^j that do not lie on one line by the Collinearity Axiom. Thus, there are at least three different lines on which $(P \circ \Gamma^{f_D})(f_D)$ and $(P \circ \Gamma^{f_D})_s(f_D)$ lie and since their intersection is unique the beliefs $(P \circ \Gamma)$ and $(P \circ \Gamma)_s$ induced by the pair (f_D, f_D) have to coincide. The filter ignorance property concludes the proof for all (f_D, f_E).

Interpretation of Theorem

The main difference to the axiomatized representation of BGSS (i.e. (3.1)) lies obviously in the inclusion of the filtering process that captured by the indicator function in the representation (3.4). Thereby the agent only employs and needs to take into account the information that she really filtered to be most important, "relevant" or acceptable (according to some criteria, such that the admissible filter properties are met).[21] Thus, the belief formation follows a two-stage procedure of filtering and subsequent belief formation. Another deviation from BGSS concerns the dependence of the axiomatization on pairs of databases and no single database as in BGSS. Such a structure is necessary and reasonable for our axiomatization. However, for interpretational purpose and the motivation behind this work, the situation $E = D$ is most interesting. In this context, an agent is not concerned with additional information or priming of her mind by E, but evaluates the given or evoked database D. By interpreting a database D as her potentially available memory, an agent might recall or retrieve only some of her memorized experiences

[21]It allows also a conditional belief formation, i.e. $s(x, c) = 1_{\{x = x_c\}}(c)$, mentioned already in BGSS, but unfeasible in their setup.

$\Gamma(D)$. Basically $\Gamma(D)$ can be seen as the result of a brainstorming or coming to mind process of those past experiences that she deems most appropriate, valuable, salient or wants to take into account.

3.5.1 Example: Similarity satisficing

Initially Simon (1955, 1957) introduced satisficing behavior as an alternative approach to the classical rational choice theory. According to him, in most global models of rational choice all alternatives are evaluated before a choice is made, but " in actual human decision-making alternatives are often examined sequentially. We may, or may not, know the mechanism that determines the order of procedure. When alternatives are examined sequentially, we may regard the first satisfactory alternative that is evaluated as such as the one actually selected" (Simon (1955), p. 110). In general, satisficing behavior is a relevant and often observed heuristic in reality (e.g. for experimental evidence see Caplin et al. (2011), Reutskaja et al. (2011)).
In our motivating example of a similarity satisficing procedure (3.2), the filtered belief was

$$(P \circ \Gamma^E)(D) = \frac{\sum_{c \in D} s(c) 1_{\{s(c) \geq s^*\}}(c) P^c}{\sum_{c \in D} s(c) 1_{\{s(c) \geq s^*\}}(c)}.$$

Its interpretation is especially appealing if a database is identified with recalled memory and assuming that those experiences are retrieved earlier that are most similar to the current problem. In this situation an agent will stop to contemplate after some time and starts to process the till-then recalled (most relevant) information.

Obviously, setting the threshold level to zero, i.e. $s^* = 0$, would result in the BGSS representation (3.1). This directly shows that s^* needs to be restricted in order to be meaningful embedded in our filtered belief formation framework. In particular, our filtering process must satisfy the minimal attention and consideration property. The former property requires that we need to take into consideration the k ($k \geq 3$) most similar cases. This determines the threshold values s^*. Obviously, such a threshold needs to be database-dependent, i.e. $s^*(E) =: s^E$ for all $E \in C^*$. More precisely, define for all $E \in C^*$ $S_E := \{(s(c))_{c \in E}\}$ and denote by s^{E_j} the j-largest number $s(c)$ according to $\geq$ in S_E. Then we get directly that for all $E \in C^*$ the database-dependent threshold level s^E is given by $s^E = s^{E_k} 1_{\{div(E) \geq k\}}(E)$ for any cognitive ability $k \geq 3$. Such a definition of the threshold s^E implies the minimal attention property. In order to also satisfy the consideration property - i.e. $c \in \Gamma(E)$, then $c \in \Gamma(D)$ for all $D \subseteq E$ - we need to enforce $s^E \geq s^D$ for $D \subseteq E$. The resulting filter $\Gamma^E(D) = (c \in D \mid s(c) \geq s^E)$ satisfies the remaining properties directly.

Summarized, this yields the following Corollary.

Corollary 3.1
Let P be as in Theorem 3.1 and Γ a similarity satisficing filter $\Gamma^E(D) := (c \in D | s(c) \geq s^E)$ for $D \subseteq E \in C^*$ with database dependent similarity thresholds s^E as defined above. Then the equivalence in Theorem 3.1 holds with the specific representation

$$(P \circ \Gamma^E)(D) = \frac{\sum_{c \in D} s(c) 1_{\{s(c) \geq s^E\}}(c) P^c}{\sum_{c \in D} s(c) 1_{\{s(c) \geq s^E\}}(c)}.$$

A database dependent threshold is even more in the spirit of Simon's satisficing behavior. Simons hypothesis is that most subjects search sequentially and stop search when an environmentally determined level of reservation utility (similarity in this context) has been surpassed. Hence, for the specification of their reservation or satisficing level individuals take into account the environment, i.e. in our setup the perspective or information set E. In addition, Simon proposed that the levels of reservation utility (similarity here) increase with set size and object complexity, i.e. for larger databases in our setup. Thus, both conditions on the threshold - database dependence and increasing in database complexity - are well-grounded in the satisficing literature.

A recent related paper that is concerned with the axiomatization of a two-stage threshold representation (Manzini et al. (2013a)) obtains various structures for the threshold values. Comparing it to Lleras et al. (2010) (or our admissible filter) they get as well $s^E > s^D$ for $D \subset E$. However, the attention filter model of Masatlioglu et al. (2012) results in $s^E = s^D$ for any nested D and E, such that $\Gamma(E) \subseteq D$ and for the two stage salience model of Tyson (2013) even the converse inequality holds.

3.6 Related literature with consideration or elimination stage

3.6.1 Relationship to multi-criteria/stage Decision Theory

As mentioned in the introduction the (implicit) formation of a consideration set is part of many recently developed multistage decision procedures in which a consideration set is constructed by eliminating several alternatives according to some criteria. The literature varies in the process of filtering. Some employ (sequences of) rational(es) to eliminate alternatives (Manzini and Mariotti (2007, 2012a), Apesteguia and Ballester (2013), Houy (2007, 2010), Houy and Tadenuma (2009)) or accept alternatives that can be justified by some of multiple criteria (Cherepanov et al. (2013) (CFS from now on), Gerasimou (2013)). Other procedures are based on undominated category(ies) (Manzini and Mariotti (2012b)) or specific frames (orders, lists, moods, fairness) which are unrelated to preferences (Salant and Rubinstein (2008)). Also mental constrains might induce consideration sets (Masatlioglu and Ok (2005)).

The aim of this section is to interpret and identify our filter in terms of a multi-criteria choice correspondence by adopting the above mentioned multistage elimination procedures. A problematic issue in merging both concepts lies in the fact that choice models usually are intended to identify a single chosen alternative by choosing in the final step the "best" alternative within the remaining consideration set. However, for a filtering process and implied corresponding consideration sets singletons are not desirable. For this reason we do not identify and compare the entire choice process with a filtering process, but we are mainly interested in the filtering and elimination stages and not on the final choice stage. On the other hand we can stick to these choice models if we replace the criteria applied in the final choice step (often binary asymmetric relations) by appropriate satisficing criteria (as discussed above) such that we end up with a set of acceptable alternatives. Roughly speaking, we discuss and relate the models in a more approximative and intuitive style, being aware of the difficulties and basic differences.

3.6.2 Filter as choice correspondence in multistage procedures

In order to identify a filter Γ as a choice correspondence we need to discuss our filter properties in a multistage decision theoretic framework.

The **filter definition** $\emptyset \neq \Gamma(D) \subseteq D$ is plausible for a choice correspondence since active, non-empty choices need to be made from D. The second part of the definition -$\Gamma^E(D) = \Gamma(E) \cap D$- can be interpreted as a usual consistency condition or as choice from E given a (budget) constraint D.

Since usual decision theoretic frameworks deal with sets of alternatives in which orderings and repetitions are immaterial, the **properties of invariance, equal treatment** and **ignorance of additional identical information** are directly satisfied for a choice correspondence.

Our **minimal attention span** property can be interpreted as a restriction to multiple choices (i.e. correspondences) such that a minimum of k available cases need to be chosen. Of course this requirement differs from common decision theoretic frameworks in which no restrictions on the quantity of chosen elements is enforced (unless non-emptiness). Thus, it will be crucial to implement this property into an adopted choice correspondence. A possible approach will be proposed and discussed later in this section. In particular, we exemplarily adopt CFS's rationalization model, but the taken approach can be applied to other models as well (see below). Another approach to guarantee for a minimum amount of choices can be imposed by enforcing an appropriate satisficing strategy at some stage of the choice process.

However, for the moment we want to focus on the **consideration property**. In the choice literature it is known as Sen's property α or akin as Contraction Axiom. Many modified versions of the weak Axiom of revealed preferences (WARP) satisfy the consideration property. Manzini and Mariotti (2007) introduce a **Weak WARP** that is also satisfied by CFS (2013), Lleras et al. (2010) and Lombardi (2009) that states that for all $c, c' \subset D \subset E$ and $c = \Gamma((c, c')) = \Gamma(E)$, then $c' \notin \Gamma(D)$. This coincides with **ReWARP** of Gerasimou (2013). Lleras et al. (2010) introduce a **Limited Consideration WARP** that states that for any $c \in D \cap E$ it holds that $c \in \Gamma(E)$ if (i) $\Gamma(E) \in D$ and (ii) $c \in \Gamma(A)$ for some $A \supseteq E$.
However, there are also some refinements of WARP in our context that do not satisfy the consideration property, e.g. Gerasimou (2013)'s **DeWARP** is neither satisfied in generality nor **Moody WARP** of Manzini, Mariotti (2013b).

For the remaining section our interest lies on the link between the elimination procedures suggested in the multistage decision models and the consideration property. We translate the choice theoretic approaches directly into our database framework and will not state the original versions of their models.

Models without an explicit procedure to form consideration sets

In Lleras et al. (2010) the formation of the consideration set is directly characterized by the consideration property, as in our approach.

Masatlioglu et al. (2012) axiomatizes choice based on a consideration set generated by an attention filter, which is characterized by

$$\Gamma(D) = \Gamma(D \backslash c) \text{ for all } c \in D \backslash \Gamma(D).$$

Basically, such a filter selects only those cases in a database D that she is aware of and effectively pays attention to. Hence, if an agent is not aware of $c \in D \backslash \Gamma(D)$, then removing c from the database D should not affect the set of cases an agent would pay attention to. In contrast, our consideration property does not rely on such an unawareness component and in general both properties (attention and consideration) are independent (see Lleras et al. (2010) for an example demonstrating the differences).

Models with explicit formation procedures

Given a binary relation R on C, we denote by $U(D,R) := (c \in D \mid \not\exists \tilde{c} \text{ with } \tilde{c}Rc)$ the cases in D that are undominated[22] in D and by $Dom(D,R) := (\tilde{c} \in D \mid \exists c \in D \backslash \tilde{c}$ such that $cR\tilde{c})$ the set of cases in D that are dominated by a case in D.

Sequential elimination Procedures

The following approaches adopt the same rough idea of "short-listing" in which multi-criteria are checked sequentially and only those alternatives survive until the final consideration stage that meet some or all criteria. The elimination is based mainly on un-dominance inspections regarding the specific criterion.

Manzini and Mariotti (2007) axiomatize choice following a rational shortlisting behavior. The final choice is made according to criterion R_2 within the alternatives in the consideration set that are undominated and survived the elimination according to R_1 (asymmetric and transitive binary relation)

$$\Gamma(D) = (c \in D \,|c \in U(R_2, U(R_1, D))).$$

Houy (2007, 2010) introduces as well another mechanism to form a consideration set that is based on a composition of some binary relations R_i, in which agents sequentially check for a certain pattern of (un)dominance according to ordered criteria

$$\Gamma(D) = (c \in D \,| \text{ for all } \tilde{c},\ \ cR_1\tilde{c} \ \text{ or } \ (\neg(\tilde{c}R_1c) \text{ and } cR_2\tilde{c}) \text{ or } (\neg(\tilde{c}R_2c) \ \vee \ cR_3\tilde{c}...)).$$

In a similar spirit Horan (2013) summarizes many of these two stage models in which a consideration stage is formed according to undominance based on a asymmetric relation and then a second (asymmetric) relation is used for the choice. Obviously, all such short-listing procedures (and extended to more criteria $(R_i)_i$) do satisfy the consideration property.

A sequential elimination procedure of a different kind is discussed in Manzini and

[22]For our purpose it might make more sense to think rather in terms of satisficing than maximal elements, e.g. in the sense of $U(R,D) = (c \in D|cRc^*)$ for some threshold case c^*.

Mariotti (2012b). Their consideration set is formed according to a (asymmetric, possibly incomplete) relation P on subsets that are interpreted as categories. The un-dominated categories survive the elimination phase.

$$\Gamma(D) = (S \subset D \mid \nexists \tilde{S} \subset D \text{ such that } \tilde{S}PS).$$

In order to capture the consideration property, specific requirements on the categorization structure are necessary. Bleile (2014c) (or Chapter 4) implements two potential versions of categorizations on databases into a belief formation process.

Satisficing procedures

The following branch of literature adopts the satisficing idea of Simon (1955, 1957).

Tyson (2008) axiomatizes a satisficing procedure based on a considerations set that contains only those elements that exceed some database dependent threshold level Θ according to some numerical representation of a criterion f

$$\Gamma(D) = (c \in D \mid f(c) \geq \Theta(D)).$$

Such a filter satisfies the consideration property for an appropriate definition of the database dependent threshold value Θ (as in our similarity satisficing example in Section 3.5.1).

Tyson (2013) and Manzini et al. (2013a) modify and generalize this procedure to salience measures and general relations.

Papi (2012) proposes an axiomatic characterization of the satisficing heuristic under various informational structures in which the order of inspecting alternatives are either full, partially or not observable. Especially the case of unobserved sequences can be interpreted within a framework of choice correspondences by assuming that for all possible orders the satisficing elements enter the consideration set.

Frame related elimination procedures

Salant and Rubinstein (2008) model a general approach in which the in principle available choice set is restricted by subjective and psychological constraints, rationales or biases. They call such additional characteristics that are not directly covered by the objective description of the alternatives as frames

$$\Gamma(D) = (c \in D \mid \Gamma(D, f) = c \text{ for some frame } f).$$

Obviously, for an unspecified frame our filter properties are not directly satisfied.

Reason Based Choice procedures

In general, the above mentioned procedures to form a consideration set can be interpreted under the premise of having and/or seeking a reason to accept or eliminate alternatives. This general idea follows the stream of literature on reason based choice initiated by Shafir et al. (1993) or even Tversky (1972). That is, elements in the consideration set are those that can be (internally) justified most easily (according to some reasons). Thereby, as above, an (un)dominance structure (according to one, some or all criteria) serves as convincing reason for choosing the specific element. A link between reason based choice and the consideration property can be established by the insight that in a smaller set it might be easier to find a reason to choose some alternative, whereas in a larger set it also might be easier to find a reason to reject.

Lombardi (2009) relies on the concept of reason based choice in the sense of finding the best and most easily justifiable alternatives by possessing the "most convincing" dominance structure. It constructs a consideration filter by employing the same criterion for the screening as well as for the final evaluation, but in different ways.[23] The rational (acyclical binary relation) used as a criterion in the first stage is used to construct another criterion for the second stage.

$$\Gamma(D) = (c \in U(D) \mid \nexists\, \tilde{c} \in U(D) \text{ such that } Dom(\tilde{c}, D, R) \supset Dom(c, D, R)),$$

where $Dom(c, D, R) := (\tilde{c} \in D \text{ such that } cR\tilde{c})$. Such a procedure does not satisfies the consideration property in general.

Gerasimou (2013) also relies on a procedure based on a single (acyclical or asymmetric) relation R. The consideration set contains elements that are justified by the fact that they are un-dominated, but at least one alternative is worse[24]

$$\Gamma(D, R) = (c \in D \mid \nexists\, \tilde{c} \in D \text{ such that } \tilde{c}Rc \text{ and } \exists c' \in D \text{ s. th. } cRc').$$

Such a procedure does not satisfy in general the consideration property.

Similarly, in the vein of seeking reasons to justify the selection, De Clippel and

[23] In general, the psychological literature (e.g. Andrews and Srinivasan (1995), Roberts and Lattin (1997)) states that the criteria influencing consideration and the final evaluation stage may differ as well as (partially) overlap. Overlapping criteria, however, play different roles at both stages.

[24] Note, there is a difference to undomination

Eliaz (2012) employ a pro-cons bargaining procedure based on linear orders P_1, P_2. The agent forms the consideration set via (internal) compromising between P_1 and P_2 by trying to receive as many as possible dominated alternatives for both rationales.

$$\Gamma(D) = (c \in D \,|\, argmax_c min_i |Dom(D, P_i)|)$$

The consideration property needs not to hold in general for this procedure.

However, the most interesting and elegantly fitting model for our approach is CFS's (2013) rationalization theory that we want to merge with our approach exemplarily.

Rationalization Theory and related psychological filter

CFS model a consideration filter explicitly as a rationalization procedure. For a set of binary relations $R = \{R_1, ..., R_n\}$ on C, a case in database D is rationalized if $cR_i\tilde{c}$ for all $\tilde{c} \in D$ for some $i \leq n$. This psychological filter contains those alternatives that are justifiable by at least one criterium, rational, reason, story, etc.

$$\Gamma^R(D) = (c \in D \,|\, \exists i \leq n \text{ such that } cR_i\tilde{c} \text{ for all } \tilde{c} \in D).$$

This procedure is very interesting, since such a psychological filter satisfies the consideration property directly and hence the rationalization procedure can be seen as a generator of any filter satisfying the consideration property. However, for our purpose we need to take care of our additional minimal attention property. Roughly speaking, we want to find a reasonable procedure that delivers an admissible filter via a rationalization similar to CFS.

For a binary relations S and for $D \in C^*$, we define the following recursive (maximal) domination sets for an attention level $k \geq 3$

$$\begin{aligned} &Max(D,S) =: Max^1(D,S) := (c \in D \,|\, cS\tilde{c} \text{ for all } \tilde{c} \in D) \\ &\text{for } n > 1: \quad Max^n(D,S) := Max(D \setminus \cup_{i \leq n-1} Max^i(D,S), S) \quad \text{and} \\ &\qquad\qquad\quad\; Max^*(D,S) := Max^d(D,R) \text{ for } d := argmin_n\{div(Max^n(D,S)) \geq k\}. \end{aligned}$$

Note that $Max^*(D,S)$ can be empty, e.g. for $div(D) < k$.

Definition 3.9

a) Let $S = \{S_1, ..., S_N\}$ be a be a set of binary relations on C and $D \in C^*$. Then we call a case c minimal k-attentive rationalizable (MAR) by S in D if and only if $c \in Max^*(D, S_1) \;\vee\; c \in \circ_{i=2}^{N} Max(D, S_i)$.

The set of all MAR cases (by rationales S) in database D is denoted by $MAR(D, S)$.

b) A filter Γ is called a minimal k-attentive psychological filter based on a set of binary relations $S = \{S_1, .., S_N\}$, i.e. $\Gamma = MAR(D, S)$, if it holds
(i) for any $D \in C^*$ such that $div(D) \leq k$: $\Gamma(D) = D$ and
(ii) for any $D \in C^*$ such that $div(D) \geq k$ and $Max^*(D, S_1) \neq \emptyset$: $\Gamma(D) = MAR(D, S)$.

Basically the modification of a CFS filter serves the reason to capture the required minimal attention property by enforcing ad hoc that always the k-best cases according to a most important, seminal, distinguishing, leading rational (criterium, reason, story) S_1 are consider for sure. In addition, if they differ from the "best" cases according to the other criteria, also all cases which are rationalizable by these other rationales survive the elimination procedure. A plausible way to justify such a formation process would emphasize the extraordinary role of criterion S_1. An agent is focussing on the (at least) k-best un-dominated or most salient alternatives for the most important criterion S_1 and only the best alternatives according to the minor, rather marginal or negligible criteria are worth to consider. For instance, an agent buying a car would choose according to different criteria, like speed, mileage, gas consumption, etc. Her major criteria might be gas consumption and hence includes the k best cars regarding economy into her consideration set, whereas she only takes the fastest car and that with lowest mileage into account, since they are outstanding or salient within the minor criteria.

For the definition and the underlying recursive domination we had in mind $k = 3$. For larger minimal attentions k one can generalize this approach to any specific structure of ranking criteria. For instance for $k = 4$, one might assume to consider for sure the two best alternatives according to rational S_1, and additional the two best remaining alternatives regarding to story S_2. In this sense a minimal attentive psychological filter can be generalized in arbitrary ways for specific attention levels.

The non-emptiness requirement of a MAR-filter, i.e. $Max^*(D, S_1) \neq \emptyset$, is for example satisfied if the binary relation S_1 is complete. But also for an incomplete "benchmark" or satisficing relation S_1[25] the non-emptiness is satisfied if c^* is chosen such that there exist at least k cases $c \in D$ such that cS_1c^*.
Consequently, we can state the following corollary.

Corollary 3.2
Let Γ be an (extended) MAR-filter on $C^* \times C^*$ based on a set of binary relations

[25] I.e. for a benchmark c^* cS_1c' if and only if $cS_1c^*S_1c'$ or (cS_1c^* and $c'S_1c^*$) for $c, c', c^* \in D$, see also Definition 6.2 later.

$S = \{S_1, ..., S_N\}$. Then Γ is an admissible filter and the equivalence in the Theorem 3.1 holds for all $D \subseteq E \in C^*$, such that $\Gamma^E(D) \neq \emptyset$ with the specific representation

$$(P \circ \Gamma^E)(D) \quad = \quad \frac{\sum_{c \in D} s(c) 1_{MAR^*(E,S)}(c) P^c}{\sum_{c \in D} s(c) 1_{MAR^*(E,S)}(c)}.$$

In general, most of the above mentioned multistage procedures can be adopted to satisfy the minimal attention property by replacing an usual dominance structure by our defined Max^* structure for some rational at some stage of their elimination procedures. In this way, these modified multistage choice processes (that satisfy the consideration property) can be interpreted as an admissible filter and incorporated into our filtered belief formation.

An appealing and intuitive example for rationales S_i in S is to interpret them as rationales that are related to componentwise similarities on the characteristics space $X = X^1 \times ... \times X^N$, i.e. $s_i(c) \in \mathbb{R}_+$ for all $c \in C$ $(i \leq N)$. One can understand the endogenously derived similarity value s as an (complex) aggregation f (f: $\mathbb{R}^N \to \mathbb{R}$) of these componentwise similarities- i.e. $s(c) = f(s_1(c), ..., s_N(c))$. An underlying motive for choosing such a form relies on the fact that agents tend to evaluate the dimensions lexicographically instead of aggregating multi-evaluation criteria (Tversky et al. (1988), Dulleck et al. (2011)), i.e. $(s_i)_{i \leq N}$ versus $s = f(s_1, .., s_N)$. In addition, research shows that in the filtering stage agents use noncompensatory heuristics for a rough screening based on simple criteria, e.g. a satisficing behavior like comparing s_i with thresholds s_i^*. But for the final evaluation stage a compensatory, more detailed multi-component and compromise-based procedure is taken, as in our aggregated similarity measure $s = f(s_1, .., s_N)$.[26]

Two candidates for a reasonable and plausible, but ad hoc, definition of binary relations S_i are based on comparisons of the componentwise similarities s_i in the following way.

Definition 3.10
Let there be functions $s_i : C \to \mathbb{R}$ for all $i \leq N$.
(i) "Componentwise similarity": For all $i \leq N$, a transitive and complete binary relation $\tilde{S}_i$ is defined on $C \times C$ by $c\tilde{S}_i c'$ if and only if $s_i(c) \geq s_i(c')$
(ii) "Benchmark exceeding componentwise similarity": For all $i \leq N$ the asymmetric, transitive and possibly incomplete binary relation S_i^* is defined by $cS_i^*\tilde{c}$ if and only if $(s_i(c) \geq s_i^* > s_i(\tilde{c}) \quad \vee \quad s_i(c) \geq s_i(\tilde{c}) \geq s_i^*)$ for componentwise threshold

[26] See e.g. Bettman (1979), Gensch (1987), Hauser and Wernerfelt (1990), Roberts and Lattin (1991, 1997), Andrews and Srinivasan (1995) and for a model Lombardi (2009).

values $(s_i^*)_i$.

Obviously, the binary relation defined in (i) can be used to define a MAR-filter. The relations in (ii) can be applicable for a MAR-filter if for any $D \in C^*$ s_1^* is chosen according to $s_1^* \leq s_1^{D_k} 1_{\{div(D) \geq k\}}(D)$ (as defined in Section 3.5.1).

3.7 Conclusion

Chapter 3 examines how beliefs are formed by agents that are constraint or not willing to pay attention to all potentially available pieces of information. It is well known in the psychology and marketing literature that humans do not take into account all available information due to many different reasons. Based on this insight we axiomatize a two stage belief formation procedure in which agents employ only these pieces of information that "survived" a first step of (un(intentional)) filtering (or screening). The filter is required to satisfy natural, reasonable and well known properties. The axioms on the belief level are closely related to the axioms introduced in BGSS and modified in a way to capture the link to filtering information and their consequences for induced beliefs. The resulting filtered belief is a weighted sum of estimates induced by past observed information that are attention grabbing. Thus, only pieces of information that attracted the attention and consideration of an agent are taken into account. The weights are determined by the similarities of the observed cases with the problem under consideration.

The axiomatized filtered belief formation generalizes the axiomatizations of BGSS, EG and Bleile (2014a) in which all available pieces of information are necessarily taken into account which prevents unintentional forgetting or unawareness as well as intentional application of a heuristic screening techniques that often drive human judgment. Hence, a filtered belief formation offers a cognitively less demanding and more realistic behavioral procedure to form beliefs based on data.

An intuitive and natural application of a filtered belief formation are models of satisficing behavior regarding the relevance or appropriateness of information for the current problem. Moreover, it captures also a conditional belief formation process that only takes into account identical problems in the past and neglects all not perfectly similar observations - which cannot be covered by BGSS, EG and Bleile (2014a).

In particular interesting is that filtering (and elimination) of information (or alternatives) emerged very recently as a research topic in the decision theory literature. These multi-stage/criteria models incorporate as well a first step of filtering before engaging into the final choice step. Many of these models can be easily translated and embedded into the filtering stage of our belief formation process such that our filter stage can be interpreted as a choice correspondence in terms of decision theory.

3.8 Proof of Theorem 3.1, necessity part

Let $(P \circ \Gamma^E)(D) = \frac{\sum_{c \in D} s(c) 1_{\Gamma(E)}(c) P^c}{\sum_{c \in D} s(c) 1_{\Gamma(E)}(c)}$ for all $D \subseteq E \in C^*$, such that $\Gamma^E(D) \neq \emptyset$.
Let $D = D_1 \circ D_2$ and $D \subseteq E \in C^*$.

$$\begin{aligned}
(P \circ \Gamma^E)(D) &= \frac{\sum_{c \in D} s(c) 1_{\Gamma(E)}(c) P^c}{\sum_{c \in D} s(c) 1_{\Gamma(E)}(c)} \\
&= \frac{1}{\sum_{c \in D} s(c) 1_{\Gamma(E)}(c)} \Big(\sum_{c \in D_1} s(c) 1_{\Gamma(E)}(c) P^c + \sum_{c \in D_2} s(c) 1_{\Gamma(E)}(c) P^c \Big) \\
&= \sum_{i=1,2} \frac{\sum_{c \in D_i} s(c) 1_{\Gamma(E)}(c)}{\sum_{c \in D} s(c) 1_{\Gamma(E)}(c)} \Big(\frac{\sum_{c \in D_i} s(c) 1_{\Gamma(E)}(c) P^c}{\sum_{c \in D_i} s(c) 1_{\Gamma(E)}(c)} \Big) \\
&= \lambda (P \circ \Gamma^E)(D_1) + (1-\lambda)(P \circ \Gamma^E(D_2)
\end{aligned}$$

Thus, the filtered Concatenation Axiom is satisfied. $\square$.

3.9 Proof of Theorem 3.1, sufficiency part

3.9.1 Important observations

In the proof, we treat the situation in which the level of minimal attention k is set equal to three, i.e. $k = 3$. This simplifies notational effort and is sufficient to follow the main steps of the proof that analogously work for any $k \geq 3$.

The following Lemma states useful and crucial properties for the proof.

Lemma 3.1
Let Γ be an admissible filter, then the following holds:
(i) For all $c \in D \in C^*$, $\Gamma^D(c^T) \in \{\emptyset, c^T\}$ for all T such that $c^T \in D$.
(ii) For all $D \subseteq E \in C^*$ such that $div(E) \leq 3$: $D \subseteq \Gamma^E(D) \subseteq D = ((c^{f_D(C)|D|})_{c \in C})$.
(iii) For all $D \subseteq E \in C^*$ we have some kind of idempotence, i.e.

$$\Gamma(\Gamma^E(D)) = \Gamma^E(D).$$

Proof:
(i) By definition $\Gamma^D(c) \subseteq c$, hence the equal treatment property delivers directly the desired result.

(ii) By definition and the equal treatment property, we have for all $D \subseteq E \in C^*$ that $\Gamma^E(D) = \circ_{c\in C}(\Gamma^E(c))^{f_D(c)|D|}$
If $div(E) = k \leq 3$, then by minimal attention property $div(\Gamma(E)) = k$ and thus by the equal treatment property $\Gamma(E) = \circ_{c\in C} c^{f_E(c)|E|}$. Since $D \subseteq E$, we get directly $\Gamma^E(D) = \circ_{c\in C} c^{f_D(c)|D|}$.

(iii) By definition of a filter, we have $\Gamma(\Gamma^E(D)) \subseteq \Gamma^E(D)$. The consideration property implies
$$\Gamma(\Gamma^E(D)) = \Gamma^{\Gamma^E(D)}(\Gamma^E(D)) \supseteq \Gamma^E(\Gamma^E(D)) = \Gamma^E(E) \cap \Gamma^E(D) \supseteq \Gamma^E(D)$$
and hence the claim holds true. □

Basically (ii) just says that no filtering of D takes place in this situation and only a reordering takes place (which is "immaterial" for the induced belief).

3.9.2 Translating the database framework to frequencies

An essential step in the proof is to identify databases with their frequency vectors. The space of frequency vectors is more tractable and enables us to adopt the structure of BGSS's proof (and use the procedure of Bleile (2014a) (or Chapter 2)). However, the proof here requires some additional features, since in addition filters are involved, which alters the crucial steps in the inductive proof.

General Definitions for a Frequency Framework

We need to introduce some definitions regarding the frequency framework.

The set of all frequency vectors on the ordered set of basic cases $C = \{c^1, .., c^m\}$ is given by (since C is fixed we skip it in the following)
$\Delta(C) = \Delta := \{f = (f_1, ..., f_m) \text{ s. th. } f_i \in \mathbb{Q} \cap [0,1] \text{ for all } i \leq m \text{ and } \sum_{i\leq m} f_i = 1\}$
The following set represents all frequency vectors related to databases $D \in C^T$:

$$\Delta_T := \{f \in \Delta \; f_i = \frac{l_i}{T},\; l_i \in \mathbb{N}_+,\; \sum_{i=1}^{m} l_i = T \;\text{ and } \exists\, D \in C^T \text{ such that } f_D(c_i) = f_i = l_i/T\}$$

Observe that if $f \in \Delta_T(C)$, then $f \in \Delta_{TZ}(C)$ for all $Z \in \mathbb{N}_+$, i.e. the frequency vector f_D represents all databases D^Z for some $Z \in \mathbb{N}$ and we cannot relate it to any specific database D^k for a specific $k \in \mathbb{N}$.

Definition 3.11
(i) O^m *denotes the null-vector on* $\mathbb{R}^m$*.*
(ii) For all $j \in \{1, 2, ..., m\}$ *denote by* q^j *the j-th unit vector in* $\mathbb{R}^m$*, i.e. the frequency vector representing a database containing only case* $c_j \in C$*.*
(iii) For all $d \in \Delta$ *its diversity is given by* $div(d) := |\{i \leq m \,|\, d_i > 0\}|$

Definition 3.12
(i) The $\subseteq$*-relation on frequencies* $\Delta \times \Delta$ *is defined as follows for* $d, e \in \Delta$*:*

$$d \subseteq e \text{ if and only if } d_i \geq 0 \text{ only if } e_i > 0 \text{ for all } i \leq m.$$

(ii) Let $d \in \Delta_T$ *and* $e \in \Delta_L$*, then the* $\cap$*-relation on* $\Delta \times \Delta$ *is defined by*

$$d \cap e := \left(\left(\frac{\min\{d_i T, e_i L\}}{\sum_{i \leq m} \min\{d_i T, e_i L\}}\right)_{i \leq m}\right)^{27}.$$

For definition (i) we have in mind that there exist T and L such that d represents a database of length T, i.e. $D \in C^T$, and e represents an $E \in C^L$ such that $\min_{j \leq m} d_j T \leq \min_{j \leq m} e_j L$.

Why is a transformation viable?

Roughly, we want to show that for a filtered belief formation we can identify databases $D \subseteq E \in C^*$ in the filtering process by frequencies $f_D \subseteq f_E$ such that $\Gamma^{f_E}(f_D)$ corresponds to $\Gamma^E(D)$. For this purpose, we exploit the properties of an admissible filter and the axioms on filtered belief formation in the following way.

(i) The filter ignorance property for Γ implies directly $\Gamma^{E^L}(D) = \Gamma^E(D)$ for $D \subseteq E$ and for all $L \in \mathbb{N}$, i.e. $(P \circ \Gamma^E)(D) = (P \circ \Gamma^{E^L})(D)$.
(ii) The filtered Concatenation Axiom implies (by $D^Z = D \circ ... \circ D$) $(P \circ \Gamma^F)(D) = (P \circ \Gamma^F)(D^Z)$ for an appropriate F such that $D^Z \subseteq F$ holds for $Z \in \mathbb{N}$.
However, since for all $Z \in \mathbb{N}$ there exists a $L \in \mathbb{N}$ such that $D^Z \subseteq F^L$, observations (i) yields that for $D \subseteq E$

$$(P \circ \Gamma^E)(D) = (P \circ \Gamma^{E^L})(D^Z) \text{ for all } Z \in \mathbb{N} \text{ and sufficiently large } L \in \mathbb{N}. \quad (3.5)$$

[27] **(Consistency) Remark:** If $d \subseteq e$, then obviously $d \cap e = \left(\left(\frac{\min\{d_i T, e_i L\}}{\sum_{i \leq m} \min\{d_i T, e_i L\}}\right)_{i \leq m}\right) = \left(\left(\frac{d_i T}{\sum_{i \leq m} d_i T}\right)_{i \leq m}\right) = d$

Further, by the definition of a filter and its order invariance property (and hence its implied belief invariance property) the order of the cases do not matter, which enables us to represent all involved databases by their frequency vector on C (which are independent of lengths) and their corresponding lengths.
However, by the above observations, within the filtered belief formation, lengths of databases become irrelevant in the sense of equation (3.5). In particular, since each $D^Z \in C^*$ (for any $Z \in \mathbb{N}$) is represented by the same frequency vector f_D, the "sufficiently large"-condition looses its bite (see below). Thus, we can identify the filtered belief formation process on databases by frequency vectors.

Filtered belief induced by frequencies

Definition 3.13
The filtered beliefs $(P \circ \Gamma) : C^* \times C^* \to \Delta(R)$ *based on databases* $D \subseteq E$ *translates to corresponding beliefs based on frequency vectors* $f_D \subseteq f_E$ *in the following way:* $(P \circ \Gamma) : \Delta \times \Delta \to \Delta(R)$ *such that* $\Gamma(f_D, f_E) := \Gamma(D, E)$ *and* $(P \circ \Gamma)(f_D, f_E) := (P \circ \Gamma)(D, E)$.

Basically, the weakening of the condition $D \subseteq E$ to $f_D \subseteq f_E$ runs through the implicit or intuitive interpretation of $\Gamma^{f_E}(f_D)$ in a way such that there exists an appropriate replication Z of database E (i.e. $f_E = f_{E^Z}$) such that $E^Z \supseteq D$ is matched and since by the ignorance property $\Gamma^{E^Z}(D) = \Gamma^E(D)$ the nestedness condition can be relaxed.
Thus an application of filter properties and belief axioms show the viability of the transformation from databases to the frequency framework.

Filter definition in frequency terms

Definition 3.14
For $d \subseteq e \in \Delta$, *a function* Γ: $\Delta \times \Delta \to \Delta$ *is called an e-induced filter on d if (i)* $\Gamma(e, e) := \Gamma^e(e) := \Gamma(e) \in \Delta$ *and (ii)* $\Gamma^e(d) := \Gamma(e) \cap d$ *hold.*

Definition 3.15
(i) ***Consideration property***
A filter Γ *on* $\Delta \times \Delta$ *satisfies the consideration property if for* $d \subseteq e \in \Delta$: $\Gamma(e) \cap d \subseteq \Gamma(d)$

(ii) ***Minimal attention span***
A filter Γ *satisfies the minimal attention of* $k \geq 3$ *if for all* $d \in \Delta$:

(a) If $div(d) = l \leq k$*, then* $div(\Gamma(d)) = l$
(b) If $div(d) > k$*, then* $div(\Gamma(d)) \geq k$*.*

(iii) ***Filter ignorance of repeated information***
Let $e = (\frac{p_1}{p}, \frac{p_2}{p}, ..., \frac{p_m}{p})$*,* $f = (\frac{p_1}{p+1}, \frac{p_2}{p+1}, .., \frac{p_i+1}{p+1}, ..., \frac{p_m}{p+1}) \in \Delta$*, where* $p := \sum_{j \leq m} p_j \in \mathbb{N}$ *and* $p_i > 0$*. Let* $d \subseteq e \subseteq f$*. A filter* Γ *satisfies the ignorance property if* $\Gamma^e(d) = \Gamma^f(d)$*.*

The **equal treatment** property is directly satisfied by the definition of a filter in frequency terms.

Definition 3.16
A filter Γ *on* $\Delta \times \Delta \to \Delta$ *satisfying the consideration, equal treatment, minimal attention span and ignorance property is called admissible.*

Analogously to Lemma 3.1, we get in frequency terms:

Lemma 3.2 ***(Lemma 3.1 in frequency terms)***
Let Γ *be an admissible filter on* $\Delta \times \Delta$*.*
(i) For all $q^j \subseteq d \in \Delta$*, we have* $\Gamma^d(q^j) \in \{\emptyset, q^j\}$*.*
(ii) For all $d \subseteq e \in \Delta$ *such that* $div(e) \leq 3$ $\Gamma^e(d) = d$ *holds.*
(iii) For all $d \subseteq e \in \Delta$*:* $\Gamma(\Gamma^e(d)) = \Gamma^e(d)$*.*

Axioms in frequency terms

Filtered Concatenation Axiom
Let Γ be a filter on $\Delta \times \Delta$. For all $d \in \Delta_T$ and $e \in \Delta_L$ for any $T, L \in \mathbb{N}$, there exists $\lambda \in [0, 1]$, such that for $g \supseteq f := \frac{T}{T+L}d + \frac{L}{T+L}e$

$$(P \circ \Gamma^g)(f) = \lambda(P \circ \Gamma^g)(d) + (1 - \lambda)(P \circ \Gamma^g)(e),$$

where $\lambda = 0$ if and only if $\Gamma^g(d) = 0^m$.

Collinearity Axiom
No three of $\{((P \circ \Gamma^{q^j})(q^j))_{j \leq m}\}$ are collinear.

Sufficiency part of Theorem 3.1 in frequency terms

Proposition 3.1
Let there be given a function $(P \circ \Gamma) : \Delta \to \Delta(R)$*, where* $P : \Delta \to \Delta(R)$ *and* Γ *an*

admissible filter on $\Delta \times \Delta$. Let a filtered belief $(P \circ \Gamma) : \Delta \times \Delta \to \Delta(R)$ satisfies the filtered Concatenation and Collinearity Axiom.
Then, there exist unique probability vectors $(P^j)_{j \leq m} \in \Delta(R)$, and unique -up to multiplication by a strictly positive number- strictly positive numbers $(s_j)_{j \leq m} \in \mathbb{R}$ such that for all $q \subseteq f \in \Delta$ such that $\Gamma^f(q) \neq 0^m$

$$(P \circ \Gamma^f)(q) = \frac{\sum_{j \leq m} s_j \Gamma_j^f(q) P^j}{\sum_{j \leq m} s_j \Gamma_j^f(q)},$$

where $\Gamma_j^f(q)$ denotes the frequency of case c_j in $\Gamma^f(q)$.

3.9.3 Proof of sufficiency part of Theorem 3.1 in frequency terms

We have by Lemma 3.2 (ii) directly that $\Gamma^f(q^j) = q^j$ for all $f \supseteq q^j$ such that $div(f) \leq 3$ and hence we need to choose

$$P^j = (P \circ \Gamma^f)(q^j) \tag{3.6}$$

The aim of the inductive proof over m with $|C| = m$ and $div(f) \leq m$ is to find the similarity values $s_1, ..., s_m$.
Step 1: $|C| = m = 3$, w.l.o.g. $C = \{c_1, c_2, c_3\}$, thus $\Delta = \Delta(q^1, q^2, q^3)$
Step 1.1: Defining similarity weights
We define $q^* := \frac{1}{3}(q^1 + q^2 + q^3)$ and for $f \supseteq q^*$ Lemma 3.2 (ii) yields

$$(P \circ \Gamma^f)(q^*) = \frac{\sum_{j \leq 3} s_j \Gamma_j^f(q^*) P^j}{\sum_{j \leq 3} s_j \Gamma_j^f(q^*)} = \frac{\sum_{j \leq 3} s_j \frac{1}{3} P^j}{\sum_{j \leq 3} s_j \frac{1}{3}} = \frac{\sum_{j \leq 3} s_j P^j}{\sum_{j \leq 3} s_j}$$

According to the filtered Concatenation Axiom there exist $\lambda \in int(\Delta^3)$ (by minimal attention, i.e. $\Gamma^f(q^j) = q^j$) such that

$$(P \circ \Gamma^f)(q^*) = \sum_{j \leq 3} \lambda_j (P \circ \Gamma^f)(q^j) = \sum_{j \leq 3} \lambda_j P^j,$$

where the last equality follows from (3.6).
By equating both representations we can derive the corresponding similarity weights s_1, s_2, s_3 uniquely up to multiplication by a strictly positive number and define for all $q \subseteq f \in \Delta$

$$(P \circ \Gamma^f)_s(q) := \frac{\sum_{j \leq m} s_j \Gamma_j^f(q) P^j}{\sum_{j \leq m} s_j \Gamma_j^f(q)}$$

The aim is now to show that for all $(q, f) \in \Delta \times \Delta$ such that $q \subseteq f$

$$(P \circ \Gamma^f)_s(q) = (P \circ \Gamma^f)(q). \tag{3.7}$$

All such (q,f) are collected in $E := \{(q, f) \in \Delta \times \Delta \,|\, (P \circ \Gamma^f)_s(q) = (P \circ \Gamma^f)(q)\}$, where obviously by definition

$$\begin{aligned} (q^j, f) \in E \qquad & \text{for all } j \leq 3 \text{ and } f \in \Delta \text{ such that } q^j \subseteq f \\ \text{and } \; (q^*, f) \in E \qquad & \text{for all } f \in \Delta \text{ such that } q^* \subseteq f. \end{aligned} \tag{3.8}$$

Step 1.2: All simplicial points (with appropriate perspective) satisfy equation (3.7)

Notation: In the following we will denote for $a, b \in \Delta$ or $a, b \in \Delta(R)$ the straight line through a and b by (a, b) (since there won't be a confusion to the usual interval notation).

The main tool of the proof is the following observation, which will be recursively applied in an appropriate manner in the proof.

Lemma 3.3

Let $a, b, c, d, e \in \Delta$, where $e = (a, b) \cap (c, d)$ and for all $f \in \{a, b, c, d\}$ let $div(f) \leq 3$ and $(f, f) \in E$. If $((P \circ \Gamma^f)(f))_{f \in \{a,b,c,d\}}$ are not collinear, then $(e, g) \in E$ for $g \in \Delta$ such that $div(g) \leq 3$ and $e, f \subseteq g$ for all $f \in \{a, b, c, d\}$.

Proof:

W.l.o.g. let e be between a and b on the line through a and b. Since $(P \circ \Gamma)_s$ and $(P \circ \Gamma)$ satisfy filtered Concatenation Axiom we get

$$\begin{aligned} (P \circ \Gamma^e)(e) &\in ((P \circ \Gamma^e)(a), (P \circ \Gamma^e)(b)) \;\text{ and} \\ (P \circ \Gamma^e)_s(e) &\in ((P \circ \Gamma^e)_s(a), (P \circ \Gamma^e)_s(b)). \end{aligned}$$

For $f \in \{a, b\}$ such that $div(f) \leq 3$, we get with Lemma 3.2 (ii) for all $g \in \Delta$ such that $div(g) \leq 3$ and $e, f \subseteq g$

$$\begin{aligned} (P \circ \Gamma^g)(e) &\in ((P \circ \Gamma^g)(a), (P \circ \Gamma^g)(b)) \;\text{ and} \\ (P \circ \Gamma^g)_s(e) &\in ((P \circ \Gamma^g)_s(a), (P \circ \Gamma^g)_s(b)) \end{aligned}$$

Analogously, we get a similar result for the segment (c, d).

By Lemma 3.2 we know that $\Gamma^g(f) = \Gamma^f(f)$ for all $f \in \{a, b, c, d\}$ and since by assumption $(f,f) \in E$, i.e. $(P \circ \Gamma^f)_s(f) = (P \circ \Gamma^f)(f)$, we directly get $(P \circ \Gamma^g)_s(f) = (P \circ \Gamma^g)(f)$ and

$$(P \circ \Gamma^g)(e), (P \circ \Gamma^g)_s(e) \in ((P \circ \Gamma^g)(a), (P \circ \Gamma^g)(b)) \cap ((P \circ \Gamma^g)(c), (P \circ \Gamma^g)(d)).$$

Since the intersection of the two line is unique due to the Collinearity Axiom, we get the desired result, i.e. $(P \circ \Gamma^g)(e) = (P \circ \Gamma^g)_s(e)$ and $(e,g) \in E$. $\square$.

By Lemma 3.2 we know that $\Gamma^f(q) = q$ for any $q \subseteq f$ such that $div(f) \leq 3$, hence we need to show equality (3.7) not for all appropriate pairs (q, f), but only for any q and some appropriate f such that $q \subseteq f$. Then it will hold for all f such that $q \subseteq f$.

In the following, we will partition the simplex Δ into so called simplicial triangles recursively, as illustrated in the Figure 3.1 below.

Definition of Simplicial Triangles:

The 0-th simplicial partition consist of vertices $q_0^j \in \Delta$, which are exactly the unit vectors q^j for $j = 1, 2, 3$. The first simplicial partition of Δ is a partition to four triangles separated by the segments connecting the middle points between the two of the three unit frequency vectors, i.e. $q_1^1 := (\frac{1}{2}q^1 + \frac{1}{2}q^2)$, $q_1^2 := (\frac{1}{2}q^2 + \frac{1}{2}q^3)$ and $q_1^3 := (\frac{1}{2}q^3 + \frac{1}{2}q^1)$. The second simplicial partition is obtained by similarly partitioning each of the four triangles to four smaller triangles, and the l-th simplicial partition is defined recursively. The simplicial points of the l-th simplicial partition are all the vertices of triangles of this partition.

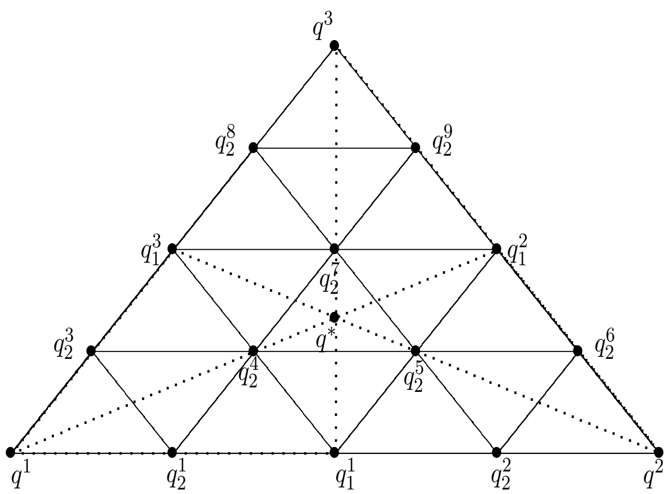

Figure 3.1: 1st and 2nd Simplicial partitions

We want to show that all simplicial points (with appropriate perspective) satisfy equation (3.7), i.e. are in E, by induction over the l-th simplicial partitions. Step 1.1. showed the claim for $l = 0$ (equation (3.8)). We proceed to the points in the **First simplicial partition**:

Since

$$q_1^1 = (q^1, q^2) \cap (q^3, q^*),$$

and we already know that for all $f \in \{q^1, q^2, q^3, q^*\}$ $(f, f) \in E$ we can apply Lemma 3.3 if the collinearity condition holds. However, since $(P \circ \Gamma^{q^i})(q^i) = P^i$ for $i = 1, 2, 3$ and $(P \circ \Gamma^{q^*})(q^*) \in int(conv(\{P^1, P^2, P^3\}))$ the Collinearity Axiom directly induces the non-collinearity condition. Hence, by Lemma 3.3 we get that $(q_1^1, f) \in E$ for all f such that $q_1^1 \subseteq f$.

With the same reasoning, we get $(q_1^2, f), (q_1^3, f) \in E$ where $f \supseteq q_1^2$ (respectively q_1^3). Thus all pairs (q_1^i, f) consisting of a simplicial points of the first simplicial partition and all f such that $q_1^j \subseteq f$ are included in E.

For the **second simplicial partition** we distinguish between inner simplicial points and points on the boundary of the simplex Δ, i.e. between two of the corners q^j. Figure 3.2 demonstrates the intuition.

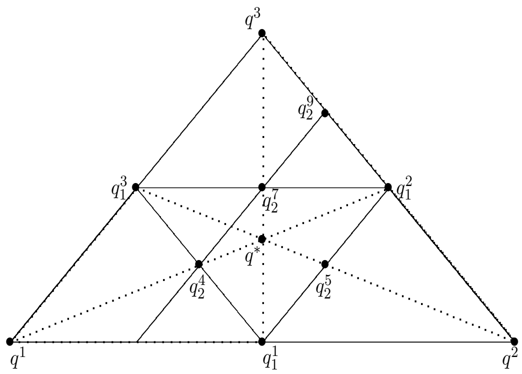

Figure 3.2: Step from 1st to 2nd Simplicial partition

(a) The **first step** involves the inner simplicial points $q_2^4, q_2^5, q_2^7 \in int(conv(\{q^1, q^2, q^3\}))$.

Since

$$q_2^4 \in (q_1^1, q_1^3) \cap (q^1, q_1^2)$$

and for all $f \in \{q_1^1, q_1^3, q^1, q_1^2\}$ $(f, f) \in E$ by Step 1.1 and Step 1.2 for the first simplicial partition, we need to check the collinearity condition to apply Lemma 3.3.

However, the condition is met since the induced beliefs for (q_1^j, q_1^j) with $j = 1, 2, 3$ are in $int(conv(\{P^i, P^j\}))$ (for appropriate $i \neq j$) that cannot lie on one line since $(P^j)_{j\leq 3}$ are not collinear by the Collinearity Axiom. Consequently by Lemma 3.3 we get that $(q_2^4, f) \in E$ for all $f \supseteq q_2^4$.
Analogously, we can get that all simplicial points (combined with appropriate perspective f) of the 2nd partition in the interior of Δ, i.e. q_2^5, q_2^7 with appropriate super-frequencies f are in E.

(b) In the **second step** we will deal with and focus on the simplicial points on the boundary of Δ (e.g. representative q_2^9, see Figure 3.2).

We have that $q_2^9 \in (q^3, q^2) \cap (q_2^4, q_2^7)$. All frequencies f involved in the intersection are shown (Step 1.1. and Step 1.2 (a) for second partition) to be contained in E, in the sense of $(f, f) \in E$. Again, the non-collinearity is fulfilled since (P^2 and P^3) and induced beliefs in $int(conv(\{P^1, P^2, P^3\}))$ are involved and $(P^2, P^3) \notin int(conv(\{P^1, P^2, P^3\}))$ since $(P^j)_{j\leq 3}$ are not collinear. Thus, Lemma 3.3 delivers $(q_2^9, f) \in E$ for all $f \supseteq q_2^3$.
The same procedure with analogous and adjusted arguments yield that all simplicial points on the boundary of the 2nd simplicial partition combined with appropriate super-frequencies (perspectives) are also included in E.

The same kind of algorithm works for all simplicial points of any l-th simplicial partitions, i.e. obviously each $q \in rim(conv(\{q^1, q^2, q^3\}))$ is for some l captured. For $q \in int(conv(\{q^1, q^2, q^3\}))$ one can approximate q via a series of simplicial points (q_l^1, q_l^2, q_l^3) such that $q \in int(conv(\{q_l^1, q_l^2, q_l^3\}_l))$ for all l. In detail, the completion for all permissible $(q, f) \in \Delta \times \Delta$ can be shown almost similarly as in Step 1.3 Bleile (2014a) (or Chapter 2) (or differently in BGSS Step 1.2 in their proof) and hence we refer to these papers for the entire procedure.
This concludes the proof for the case $|C| = 3$ and (q, f) with $q \subseteq f$ such that $div(f) \leq 3$.

Now we need to show the claim for $|C| = m > 3$ and $div(f) \leq m$.
Step 2: $|C| = m > 3$
Step 2.1: Defining the similarity weights
Using the considerations from Step 1 above for $\{j, k, l\}$ (i.e. $q^*_{\{j,k,l\}} := \frac{1}{3}\sum_{i\in\{j,k,l\}} q^i$, $q^i \in \Delta$ and $q \subseteq f \in \Delta(q^j, q^k, q^l)$) we can derive the similarity weights $(s_i^{\{j,k,l\}})_{i\in\{j,k,l\}}$.
Further, for all (q,f) such that $q \subseteq f \in conv(\{q^j, q^k, q^l\})$ the following representation

holds

$$(P \circ \Gamma^f)^{\{j,k,l\}}(q) = \frac{\sum_{i=j,k,l} s_i^{\{j,k,l\}} \Gamma_i^f(q) P^{\{j,k,l\}}(q^i)}{\sum_{i=j,k,l} s_i^{\{j,k,l\} \Gamma_i^f(q)}}.$$

Moreover for all $i \in \{j,k,l\}$, we have $(P \circ \Gamma^f)^{\{j,k,l\}}(q^i)) = P^i$ and $(s_i^{\{j,k,l\}})_{i \in \{j,k,l\}}$ are unique up to multiplication by a positive number.
Similar to BGSS or Bleile (2014a) (or Chapter 2), we can show that the similarity values $s_i^{\{j,k,l\}}$ are independent of the choice of j, k and l for all $i \in \{j,k,l\}$, since filtering is not present in the arguments. Thus we can define for all $q \subseteq f \in \Delta$

$$(P \circ \Gamma^f)_s(q) := \frac{\sum_{i \leq m} s_i \Gamma_i^f(q) P^i}{\sum_{i \leq m} s_i \Gamma_i^f(q)}.$$

The aim is to show that for all $(q,f) \in \Delta \times \Delta$ such that $q \subseteq f$ $(P \circ \Gamma^f)(q) = (P \circ \Gamma^f)(q)$.

Step 2.2: Completion to all $(q,f) \in \Delta \times \Delta$

By Step 1 we know that the claim $(P \circ \Gamma^f)_s(q) = (P \circ \Gamma^f)(q)$ is true for all (q,f) such that $div(f) \leq 3$. We take this as the base case of our induction.
For the induction assumption, we have that $(P \circ \Gamma^f)_s(q) = (P \circ \Gamma^f)(q)$ for all $(q,f) \in \Delta \times \Delta$ with $q \subseteq f$ and $div(f) \leq k-1$.
The induction step considers $q, f \in \Delta$ with $q \subseteq f$ and $div(f) \leq k$:
We can restrict the analysis to f such that $div(f) = k$, since for all other $f \in \Delta$ the claim is true by the induction assumption.
We split the proof into two parts. First for which $div(q) \leq k-1$ and then for $div(q) = k$.

First Situation: Consider $q \subset f$**, i.e.** $div(q) \leq k-1$**.**

By Lemma 3.2 (iii), we have $\Gamma(\Gamma^f(q)) = \Gamma^{\Gamma^f(q)}(\Gamma^f(q)) = \Gamma^f(q)$ and hence directly

$$(P \circ \Gamma^f)(q) = (P \circ \Gamma^{\Gamma^f(q)})(\Gamma^f(q)), \tag{3.9}$$

since $\Gamma^f(q) \subseteq q$ by definition of a filter and hence $div(\Gamma^f(q)) \leq k-1$ the induction assumption applies to the RHS of equation (3.9), i.e. $(P \circ \Gamma^{\Gamma^f(q)})(\Gamma^f(q)) = (P \circ \Gamma^{\Gamma^f(q)})_s(\Gamma^f(q))$ which is again identical to $(P \circ \Gamma^f)_s(q)$ and hence the desired result $(P \circ \Gamma^f)_s(q) = (P \circ \Gamma^f)(q)$ is implied directly.

Second Situation: Consider $q \subseteq f$ with $div(q) = k$

A similar construction as in BGSS, but with different reasoning, yields the result. Let $q = \sum_{l \in K} \alpha_l q^l$ with $\alpha_l > 0$ and $K \subseteq \{1, ..., m\}$ such that $|K| = k$.
Define the frequency vector $q(l)$ to be the vector in $conv(\{(q^j)_{j \in K \setminus l}\})$ such that q lies on the line $(q(l), q^l)$.
By the minimal attention span property $div(\Gamma^q(q)) = div(\circ_{\{l: q_l > 0\}} \Gamma^q(q^l)) \geq 3$, i.e. there exist at least three l's (e.g. $l = i, j, k$) such that $\Gamma^q(q^l) = q^l \neq 0^m$ and hence $(P \circ \Gamma^q)(q^l) = P^l = (P \circ \Gamma^q)_s(q^l)$.
Further, for these $l \in \{i, j, k\}$ we get $(P \circ \Gamma^q)_s(q(l)) = (P \circ \Gamma^q)(q(l))$ by the result of the first situation, since $div(q(l)) \leq k - 1$.
Hence we have that $(P \circ \Gamma^q)_s(q), (P \circ \Gamma^q)(q) \in (P^l, (P \circ \Gamma^q)(q(l))) =: L(l)$, for those three $l = i, j, k$. Since no three P^j are collinear, there are at least two distinct lines L(l), i.e. $L(l) \neq L(n)$ for at least two distinct $l, n \in \{i, j, k\}$. Since $(P \circ \Gamma^q)_s(q), (P \circ \Gamma^q)(q)$ are both on these distinct lines and these lines need to intersect uniquely, we have $(P \circ \Gamma^q)_s(q) = (P \circ \Gamma^q)(q)$. By the ignorance property $(P \circ \Gamma^q)(q) = (P \circ \Gamma^f)(q)$ for all f with $div(f) = k$, which completes the proof. $\square$.

4 Belief Formation Based on Categorization

Abstract

An agent needs to determine a belief over potential outcomes for a new problem based on past observations gathered in her database (memory). There is a rich literature in cognitive science showing that human minds process information in categories, rather than piece by piece. We assume that agents are naturally equipped (by evolution) with a efficient heuristic intuition how to categorize information in general.. Depending on how available categorized information is activated and processed, we axiomatize two different versions of belief formation relying on categorizations. In one approach an agent relies only on the estimates induced by the single pieces of information contained in so called target categories that are activated by the problem for which a belief is asked for. Another approach forms a prototype based belief by averaging over all category-based estimates (so called prototypical estimates) corresponding to each category in the database. In both belief formations the involved estimates are weighted according to their similarity or relevance to the new problem. We impose normatively desirable and natural properties on the categorization of databases. On the stage of belief formation our axioms specify the relationship between different categorized databases and their corresponding induced (category or prototype based) beliefs. The axiomatization of a belief formation in Billot et al. (Econometrica, 2005) is covered for the situation of a (trivial) categorization of a database that consists only of singleton categories and agents basically do not process information categorical.

4.1 Introduction and motivation

Often agents need to evaluate and judge the likelihood of future uncertain events. On which basis can individuals derive and assign likelihoods and form probabilistic beliefs over random incidents?

Traditionally, economic theory models uncertainties in a state space representation a la Savage (1954) and Bayes and derive a subjective prior based on observable actions of the agent. However, this procedure implicitly assumes that agents already know or are endowed with a subjective prior belief, which they express through their observable actions. In this way, the Savage and Bayesian approach does not advice agents how to find or form a prior explicitly. Basically, the belief is purely subjective and offers no mechanism to incorporate information directly into a belief formation. Consequently, their normatively appealing and convincing approach to endogenously derive a belief is not feasible in situation in which an agent might not be able to condense her insufficient or too complex information into a consistent state space.

We consider an axiomatization of belief formation that allows and requires to take directly into account the available information (gathered in form of a list or database of past observations or cases). The influence of data and experience on the formation of a probabilistic belief was examined initially by the axiomatization of Billot et al. (2005) (BGSS from now on). The axiomatizations of BGSS and related ones of Eichberger and Guerdjikova (2010) (EG) (for ambiguous multiprior beliefs) and Bleile (2014a) (precision dependent cautious beliefs) yield that a belief induced by a database is a similarity weighted average of the estimations induced by all observed cases in the database. Thereby similarity weights capture different degrees of relevance of the potentially very heterogenous information.

A common shortcoming of these approaches to belief formation is that an agent processes each distinct single piece of information separately and forms its induced estimate. Interpreting a database as memory an agent is assumed to store (memorize) all single pieces of information and needs to retrieve any single piece of information from her memory.[1]

However, numerous studies in (social) psychology and cognitive science show that humans do not store and treat single pieces of information in such a one by one procedure, but classify information in different categories. The prominent social psychologist Allport (1954) memorably noted "the human mind must think with the aid of categories. We cannot possibly avoid this process. Orderly living depends upon it ". There is a wealth of research demonstrating that humans' cognition processes information by employing categorical thinking, reasoning and stereotyping.[2]

[1]An axiomatization that does not take into account all potentially available information in this vein is Bleile (2014b)(or Chapter 3). It deals with a two stage belief formation that consists of a initial filtering process that "screens and selects" the information that finally flows into the belief formation process.

[2]The psychology literature on categorization is vast, e.g. see Rosch and Lloyd (1976), Murphy and Medin (1985), Goldstone (1994), Rips (1989), Smith et al. (1998), Medin and Aguilar (1999), Murphy (2002). Real life examples discuss that consumers categorize products (Smith

In particular, one can interpret categorization as model of similarity-based reasoning (Tversky (1977), Gilboa and Schmeidler (1995)) in which information needs not to be understood in its particularity, but as member of a larger classified category that allows to generalize properties from category members to new members through analogies and similarities. This makes categorical thinking especially helpful for predictions (Osherson et al. (1990), Anderson (1991)).

In order to capture the impact of categorical thinking and reasoning in agent's belief, we modify and extend the mentioned axiomatic approaches (in particular BGSS) by adding a categorization procedure that affects the processing, storing, retrieving and employing of potentially available information.

In complex and poorly understood environments, categorizations emerge naturally to simplify actions by gathering many distinct experiences together and ignoring the details of each single piece of information. Limited learning and memorizing opportunities drive agents into relying on abstractions and (categorical) summarizations rather than on single past cases. Processing and storing of all past cases in full detail bears costs in storing and retrieving the information, since the finer information is stored the more effort is required to activate it. The classification of information in different categories offers a less demanding way of storing and retrieving information, since only the assignment to suitable categories and their characteristics needs to be memorized. In particular, the literature on "optimal" categorization focuses on the issue how fine or coarse categories ought to be formed in order to process information in a way to gain a maximum amount of information with the least cognitive effort. In particular, it should be more efficient than some other form of case-based reasoning, as for instance kernel-based estimation.

Another important function of categorical reasoning concerns its role for facilitating and improving inductive inference and prediction. The underlying idea is that an assignment to categories does allow an agent not only to use the information contained in the current problem, but exploit as well the additional information provided by the categories to which this problem belongs (or which it activates). Of course, this is only helpful if the previous experiences contained in the specific categories provide some information for the actual problem such that the agent can infer or generalize some information and properties from past observations in the categories. From this perspective, a categorization of information enables and implicitly provides an agent with additional (more detailed) information than mentioned in the initial description of the problem. Ideally categories are formed like sufficient statis-

1965), investors engage in "style investing" (Sharpe (1992), Bernstein (1995)), rating agencies categorize firms wrt. default risk (Coval et al. (2009)), etc.

tics for its assigned members and thus would make prediction particularly simple and reliable.[3]

In this chapter, we are not concerned with the formation of categories, but assume that a set of (optimal) categories is already naturally or evolutionary determined.[4] In particular, we are solely interested in an axiomatic description on how categorized information is incorporated into a belief formation by agents.

The categorization literature identified several procedures in using categorical thinking for belief formation. The approaches differ in the way how many categories are taken into account. Either all categories are considered or only some specific target category(ies) are taken into account. Another difference concerns (a still ongoing discussion about) the issue how categories are represented themselves. Either categories are represented by an aggregated summarizing representative that captures the essence or central tendency of the category - a so called prototype - or all members of the category are used for its representation.[5]

There is experimental evidence in psychology that individuals tend to rely on (a single) most likely target category(ies), whereas the other categories (and their content) are immaterial for the belief formation (e.g. Murphy and Ross (1994), Krueger and Clement (1994), Malt et al. (1995)). When faced with a new problem, an agent's mind activates automatically some already generated category(ies) that are best fitting according to some metric for the current problem.[6] Depending on how an agent treats categories she will form her belief either based on all single pieces of information contained in the target category(ies) or use the estimates induced by a prototypical representative associated with the (target) category(ies). Our first axiomatization of a category based belief formation will adopt this approach based on activated target categories, which simplifies (cognitively) the belief formation, since an agent only needs to process the information that is directly evoked for the current problem.

The second stream of literature -which is covered in our second axiomatization of

[3]Peski's (2011) categorization model can be interpreted as such an optimal statistical procedure.

[4]Traditionally, categories are formed based on (attribute-wise, overall, functional or casual) similarity considerations. Roughly speaking, in general categories are often formed as to maximize the similarity of objects within a category and the dissimilarity of objects from different categories. However, there is an ongoing discussion and debated whether categorization presupposes a notion of similarity or not (see Goldstone (1994) and Gärdenfors (2000), Pothos (2005)). For instance some literature argue that categorization is theory or rule-based (according to various criteria).

[5]There are also approaches in between (Vanpaemel and Storms (2008)), but we stick to the benchmark procedures.

[6]For example by comparing the actual problem to the prototypical problem of different categories until a closest match is found. The automaticity in categorical thinking is discussed e.g. in Allport (1954), Bargh (1994, 1997, 1999).

a prototype based belief- is based on the prototype of all categories in a database. Such a prototype based belief adopts the approach taken in Anderson (1991) and models the situation, in which an agent might not be able to figure out best fitting target category(ies). The simplifying power of categorization in this approach results from taking into account all (categorized) information through prototypical summaries and not by memorizing and retrieving all single pieces of information. This reduces the cognitive and memorization effort. It averages the particular prototypical estimates induced by the categories.

Our axiomatizations of both kinds of belief formation -category and prototype based- are based on modified and extended versions of the axioms of BGSS and partly Bleile (2014b) (or Chapter 3). Categorization based belief formations can be seen as two stage procedures. First, agents are endowed with a natural categorization structure on the in principle available information (e.g. like through a natural or evolutionary developed optimal heuristic algorithm). On the basis of this categorization structure and the current problem, the available information (provided by the database/memory of past experiences) will be categorized. We will assume some reasonable, natural and well known -but rather weak- properties on this induced categorization of databases. The (structural) properties on the (induced) categorization of databases differ for our two versions of categorization based belief formation. However, a common feature concerns our main requirement that an agent will not categorize any database, but that databases must be sufficiently complex or diverse to initiate a categorization process. We assume that a database need to contain a minimum amount of distinct cases such that an agent really (wants to) thinks in categories.[7] However, this is a quite natural requirement, since one reason for categorizing information is to overcome limitations in processing cognitively challenging information or environments. Another common property says that the order of cases in a database is immaterial for the categorization, i.e. a categorization procedure depends only on content and not on the sequences of pieces of information.

For a prototype based belief we require that a "real" category ought to contain a minimum number of distinct cases. In particular, we require for a prototype based belief formation that a database is categorized (in some accordance with the natural categorization) in such a way that at most one singleton category exits and all other non empty categories consist of at least two members. Of course a degenerate categorization in which each category contains only one member is meaningless for

[7]Complexity is certainly related to the number of options to be considered, but also few options characterized by difficult interwoven features might be challenging to evaluate.

our purpose (and is covered directly by BGSS).

The second stage of a belief formation based on categorized information deals with the behavioral impacts on the actual belief level. As in BGSS, we require that the belief is independent of the order of the (categorized) information and that some form of Concatenation Axiom holds. Explicitly, a belief induced by the combination of two databases should lie between the beliefs induced by the two databases separately. However, to keep the normatively appealing spirit of the axiom, we need to ensure that the categorized information of the combination of two databases coincides with the combination of the two separately categorized underlying databases.[8]

The particular properties of the categorization procedures and the axioms on the belief formation level guarantee that the beliefs based on categorized information can be represented only based on information in the target categories evoked by the new problem (i.e. as the category based belief) or based only on all prototypes of the categorized information (i.e. as the prototype based belief).

As already mentioned, there exist research discussing predictions based on categorical thinking, but none of them is of an axiomatic nature. However, we are solely interested in the behavioral foundation of a belief formation based on categorized information. Our approach is closest to BGSS and Bleile (2014b) with regard to the proposed axioms. In particular, a category based belief can be seen as a special case of a belief formation under limited attention as axiomatized in Bleile (2014b). Such a filtered belief relies only on pieces of information that are contained in a so called consideration set that gather all information that survive a filtering stage. Basically, the evoked target categories can be interpreted as such a consideration set and thus a category based belief is a specifically filtered belief.

Concerning the two axiomatized procedures of employing categorized information, the most relevant works are Anderson (1991) for the prototype based belief and Murphy and Ross (1994) for the category based belief. The existing literature deals either with applying prediction procedures based on categorized information and discussing its consequences for specific situation (e.g. Mullainathan (2002)) or is concerned with how and why "optimal " formations of categories emerge (Fryer and Jackson (2008), Mohlin (2014), Peski (2011)). The mentioned papers all employ a belief formation relying on prototypes (of the main target categories or the category the current problem belongs to). However, they all differ with regard to their notion of an "optimal" categorization. In Fryer and Jackson (2008) the optimal categoriza-

[8]For the axiomatization of a prototype based belief we deal with this issue in a similar spirit as in Bleile (2014b) by introducing another simultaneously available (super)-database, which serves as the common reference which "dictates" the categorization in a consistent manner for any subdatabase.

tion minimizes the sum (across categories) of within category variations between objects that have already been encountered (for an exogenously fixed number of categories). Mohlin (2014) aims to find the optimal number of categories in order to minimize the prediction error, which amounts to tackling the tradeoff between small and fine or more coarse but larger categories to avoid overfitting problems (which is also the principle rational of Peski (2011) and Al Najjar and Pai (2014) for decision making). Peski (2011) shows when categorical learning is optimal for prediction (in the sense of an (asymptotic) statistical tool equivalent to learning by Bayesian updating). He compares a categorization algorithm with Bayesian updating. The underlying assumption is that the environment is symmetric meaning that the Bayesian prior is symmetric. The categorization algorithm is such that categories are formed in order to minimize the inner entropy of the categories, i.e. to maximize the informational content in the categories. His categorization procedure combines deductive reasoning (i.e. learn to form categories) and applying the deduced categorization inductively to belief formation. This is structural impossible in our axiomatization.

The following section will introduce the database related framework. Section 4.3 illustrates by means of an example the two beliefs based on categorizations. In Section 4.4 our natural categorization structure is discussed. Section 4.5 and 4.6 cover the categorization of databases, the axioms on the belief level and the resulting belief formations for both categorization based belief formations separately. We conclude in Section 4.7. All proofs can be found in the last two sections.

4.2 The model

In this section, we introduce the case-based information framework and the basic building blocks of our belief formation based on filtered information. Further, we introduce some definitions and notations necessary for our approach.

4.2.1 Database framework

A basic case $c = (x, r)$ consists of a description of the environment or problem $x \in X$ and an outcome $r \in R$, where $X = X^1 \times X^2 \times .... \times X^N$ is a finite set of all characteristics of the environment, in which X^j denotes the set of possible values features j can take. R denotes a finite set of potential outcomes, $R = \{r^1, ..., r^n\}$
The **ordered** set $C \subseteq X \times R$ consists of all $m \geq 3$ basic cases, i.e. $C = \{c^1, ..., c^m\}$. A database D is a sequence or list of basic cases $c \in C$. The set of databases D consisting of L cases, i.e. $D = (c_1, ..., c_L)$ where $c_i \in C$ for all $i \leq L$, is denoted

by C^L and the set of all databases by $C^* = \cup_{L\geq 1} C^L$, including the empty database $\emptyset$. The description of databases as sequence of potentially identical cases allows multiple observation of an identical case to be taken into account and treated as an additional source of information.
For a database $D \in C^*$, $f_D(c)$ denotes the relative frequency of case $c \in C$ in databases D.
The concatenation of two databases $D = (c_1, c_2, ..., c_L) \in C^L$ and $E = (c'_1, c'_2, ..., c'_T) \in C^T$ (where $c_i, c'_j \in C$ for all $i \leq L, j \leq T$) is denoted by $D \circ E \in C^{L+T}$ and is defined by $D \circ E := (c_1, c_2, ..., c_L, c'_1, c'_2, ..., c'_T)$.
In the following we will abbreviate the concatenation or replication of L-times the identical databases D by D^L. Specifically, c^L represents a database consisting of L-times case c.
For any $D \in C^*$ the diversity of a database D is given by $div(D) := |\{D\}|$, where as usual $\{D\}$ denotes the set of different cases contained in database D. So $div(D)$ gives the number of different cases contained in database D.
We need to translate some relations from sets to the list framework.

Definition 4.1
(i) The $\in$-relation on databases is defined by $c \in D$ if $f_D(c) > 0$.
(ii) The $\subseteq$-relation on the set of databases C^* is defined by $D \subseteq E \Leftrightarrow f_D(c)|D| \leq f_E(c)|E|$ for all $c \in C$. We will call such databases to be nested.
(iii) The $\cap$-relation on databases is given by $D \cap E = ((c^{\min\{f_D(c)|D|, f_E(c)|E|\}})_{c \in C})$
(iv) Two databases D and E are disjoint if for all $c \in C$: $c \in D$ if and only if $c \notin E$.

The definitions are basically independent of the order of cases in the databases. Note however that the definition of $\cap$-relation in (iii) is very specific, since the order of C is transferred, i.e. by intersection a specific order (on C) is induced.[9]

4.2.2 Induced belief

For a finite set S, $\Delta(S)$ denotes the simplex of probability vectors over S and for $n \in \mathbb{N}$ Δ^n denotes the simplex over the set $\{1, 2, ..., n\}$.
As in BGSS, EG and Bleile (2014a) an agent forms a belief over outcomes $P(x, D) \in \Delta(R)$ in a certain problem characterized by $x \in X$ using her information captured in a database $D \in C^*$, i.e. $P : X \times C^* \to \Delta(R)$.

[9] In contrast to intersections of sets, where orderings are immaterial, intersection of databases do require some assumption on resulting orderings.

4.3 Motivating example

In order to illustrate the basic idea and plausibility of categorization based belief formation, we incorporate the two categorization procedures into the doctor example of BGSS.

A doctor needs to evaluate different outcomes of a treatment. She has some working experience or access to some medical database related to the treatment $D = (c_1, ..., c_l)$, where she recorded in a case $c_i = (x_i, r_i)$ the vector of characteristics of a patient i, $x_i \in X$, (e.g. age, gender, weight, blood count, specific illness) and the observable outcome of the treatment $r_i \in R$ (e.g. better, worse, adverse effects). A new patient characterized by x enters her office and using a medical record D, the doctor wants to derive a belief $P_x(D) \in \Delta(R)$ over potential outcomes in R. She might apply an empirical frequency and use only a part D_x of the database D, which contains only cases $c = (x_c, r_c)$ of patients with *"identical"* characteristics $x_c = x$ compared to the current patient,

$$\text{"Frequentist":} \qquad P_x(D) = \frac{\sum_{c \in D_x} \delta_{r_c}}{|D_x|}.$$

However, if the database contains not sufficiently many of these *"identical"* patients x, she might want to include also *"similar"* patients. She judges the degree of similarity between patients x and x' by $s(x, x') \in \mathbb{R}_+$. Further, she might induce from a case $c = (x_c, r_c)$ not only a point estimate δ_{r_c} on the realized outcome, but derives a more general estimate $P^c \in \Delta(R)$ on likelihoods of particular (related) outcomes and forms the belief as axiomatized in BGSS (2005) by

$$\text{"BGSS-belief":} \qquad P_x(D) = \frac{\sum_{c \in D} s(x, x_c) P^c}{\sum_{c \in D} s(x, x_c)}.$$

However, as discussed above, the cognitive science literature emphasized the role of categories in storing, retrieving and processing of information. The literature argues as well for a naturally (by evolution) given ability or heuristic feeling to categorize. Thus, assume the doctor is implicitly able to categorize the set C of all potentially possible patient-outcome pairs $(x, r) \in C$ into different categories $\tilde{C}_l \subset C$ for $l \leq L$, such that the set $\tilde{C} = \{\tilde{C}_1, .., \tilde{C}_L\}$ partitions the set of all possible cases $c = (x, r)$. For example category $\tilde{C}_1$ contains all male patients, with age below 60 years, any weight, good blood count and sore throats and category $\tilde{C}_l$ contains all male patients, overweight and heart problems and so forth. In general, categories might be exclusive or non-disjoint, for instance male patients might appear in different categories. Such an implicit preexisting natural categorization structure might be subconsciously rooted in the mind of a doctor and induces some consistent (embedded) categorization of a database D. Alternative, one might think of the preexisting natural categorization structure as the patient groups that the doctor are taught at medical school and the database D as the experience she made working in a hospital. Receiving information about a new patient, she wants to predict the outcome based on (a natural or taught) categorization of her past experience. Depending on how the doctor uses her categorized patient database D, she might form a category or prototype based belief as follows.

Suppose the current patient x is male, older than 40 years, overweight, etc. and the doctor knows for each characteristic its value. The doctor partitioned her experience D with the treatment into categories consistent with the natural (preexisting) categorization. A specific patient x might trigger, activate or evoke automatically the category(ies) to which this patient belongs, is related to or matches best. For example this might be the category $\tilde{C}_1$ (as above) "synchronized" (intersected) with the actually experienced database D, which we denote by $\tilde{C}(x, D) \subseteq D$. Thus, not only the identical patient profiles are recalled and considered, but the entire x-evoked category in D that also may contain different, but somehow similar or related patients (according to the criteria for categorization, which might be optimal for predictive tasks by evolution). A doctor might form a **category based belief** based only on the members the category(ies) that patient x evokes, i.e. $\tilde{C}(x, D)$,

$$\text{"Category based belief":} \qquad P_x(D) = \frac{\sum_{c \in \tilde{C}(x,D)} s(x, x_c) P^c}{\sum_{c \in \tilde{C}(x,D)} s(x, x_c)},$$

where as above s measures the similarity between the current patient and the already treated patients.
Of course $\tilde{C}(x, D)$ might not consist only of a single category, but several categories that are activated by the patient x.

Alternatively, the doctor might have already categorized her experience D prior a new patient arrives, i.e. $\tilde{C}^D = \{\tilde{C}_1^D, .., \tilde{C}_L^D\}$ for $l \leq L$, such that $\bigcup_{l \leq L} \tilde{C}_l^D = D$ in a consistent manner with respect to the (preexisting) natural categorization structure $\tilde{C}$. Furthermore, she might have formed some prototypical estimates $P^{\tilde{C}_l^D} \in \Delta(R)$ for each of these categories. For a given database D and a new patient x, the doctor's **prototype based belief** in such a situation might be given by a weighted average of these prototypical estimates, where the weights are determined by the relevance of the particular category for the current patient, i.e. $\tilde{s}(x, \tilde{C}_l^D)$ for all $l \leq L$,[10]

$$\text{"Prototype based belief":} \qquad P_x(D) = \frac{\sum_{l \leq L} \tilde{s}(x, \tilde{C}_l^D) P^{\tilde{C}_l^D}}{\sum_{l \leq L} \tilde{s}(x, \tilde{C}_l^D}.$$

4.4 Natural evolutionary (optimal) categorization structure

Definition 4.2

A (natural) categorization structure $\tilde{C} = \{\tilde{C}_1, .., \tilde{C}_L\}$ on the set of basic cases C partitions C into L different nonempty categories $\tilde{C}_l \subseteq C$ for $l \leq L \in \mathbb{N}$, i.e. $C = \bigcup_{l \leq L} \{\tilde{C}_l\}$.

The definition allows non-disjoint categories, since it is quite naturally that a case could be classified into multiple categories if a categorization depends on more than one criterion. For instance, the categories of young and male patients are not necessarily disjoint, when gender and age are criteria. Moreover, hierarchical categories are not mutually exclusive, e.g. the category of young patients and the one of young male patients.

We will not care about the formation of categories and assume that agents are naturally endowed with an idea, how to construct categories. In particular, through evolutionary pressure nature equipped us with an heuristic algorithm or tool that allows to form categorizations that organize our experience in an almost optimal way for prediction and that tends to minimize prediction errors (and thus increase the likelihood to survive and stay fit). This is supported by the findings that young children appear to form, acquire and use categories from very early on (Gelman and Markman (1986), Smith (1989), Murphy (2002)), showing also that many categorizations are innate.[11] However, if the categorization is based on data, the developmental literature shows that especially in the beginning of learning (i.e. with small databases (i.e. for children)) the categorization is still flexible (Hayne (1996), Quinn and Eimas (1996)). This concern is immaterial for a natural categorization, since it is based already on all potential pieces of information. However, we will take care of it, when we consider categorization induced by database (in particular in the framework of a prototype based belief formation).

Another justification for assuming such a fixed preexisting categorization structure on the set of all potential cases is by interpreting it as a result of an already developed

[10] Alternatively, $\tilde{s}$ measures likelihood of patients x to be assigned or belonging to the particular category.

[11] It is not observable that children rely on on purely empirical learning.

optimal categorization with regard to numbers and content of the categories. The literature varies in the way how they define an optimal categorization (as already discussed in the introduction, e.g. Fryer and Jackson (2008), Peski (2011), Mohlin (2014)).

A more direct reason, why we assume a preexisting natural categorization is based on the fact that we want to avoid the many difficult and interacting mechanisms involved in a categorization process[12] and want to focus solely on the belief formation issue.

Also in the literature it is not uncommon to assume a fixed preexisting categorization structure (Anderson (1991), Murphy and Ross (1994), Mullainathan (2002), Manzini and Mariotti (2012), Al-Najjar and Pai (2014) or Mohlin's (2014) ex ante optimal categorization).

4.5 Axiomatization of category based belief formation

4.5.1 Specific properties of the categorization procedure

In this section we specify the list of categories a problem $x \in X$ evokes or activates and our concept of a x-evoked categorization of a database.

Problem evoked categorization of a database

The definition of a problem evoked categorization of information captures the intuition that a new problem activates specific most appropriate or relevant categories (that are already generated in the natural categorization structure). An agent takes into account these "target" categories for the current problem.

Definition 4.3
For all $x \in X$ a categorization structure $\tilde{C}$ induces a list $\tilde{C}_x$ of categories that problem x evokes. For all x there exist a $M_x \subseteq \{1, .., L\}$, such that $\tilde{C}_x := ((\tilde{C}_l)_{l \in M_x}) \in C^*$
We call $\tilde{C}_x$ the ***categories that are activated (evoked) by problem x or short x-activated categories****.*

There is substantial experimental evidence showing that when faced with an object, humans' brains automatically activate category(ies) that (according to some metric) appears to suit the current problem best (with regard to best fitting, most likely

[12]e.g. initial encoding, abstraction of conceptual representation (if any), storage of the abstraction and/or exemplars in memory, retrieval of stored representations, decision process that produce categorization or typicality, see Murphy and Ross (1994).

or analogous category(ies)).[13] Basically, a new problem does not trigger some most relevant single pieces of information, but the activation process is based on categorical thinking.

Our definition does not specify the exact procedure of activation. However, for our purpose to form a belief given a current problem, i.e. the characteristic part of case $x \in X$, it is most reasonable to think about a categorization of past observations according to their characteristics.[14] In this way, the characteristic x can be seen as a basic sensory input such that categories are formed based on a relationships between the characteristics of the cases (e.g. like a metric on the characteristics space or feature overlaps, etc.). In this way, for instance a very specific procedure to evoke categories might solely activates categories that contain a case that coincides (with regard to the characteristic part) with the current problem, i.e. $\tilde{C}_x$ is the list of categories that contain at least one case with characteristic $x \in X$, i.e. let $A_l := \{x \in X | \exists c = (x, r) \in C \text{ s. th. } c \in \tilde{C}_l\}$ for all $l \leq L$, then

$$\tilde{C}_x := (1_{A_1}(x)\tilde{C}_1, 1_{A_2}(x)\tilde{C}_2, ..., 1_{A_L}(x)\tilde{C}_L) \in C^*.$$

Such an procedure would imply that cases with different characteristics than x are only activated if they are contained in categories that include member cases with characteristic (problem) x.[15]

However, our definition is general and thus may not necessary rely on the x characteristics specifically, but may take into account the categories that are activated by any characteristics that are close or related to x. Furthermore another widely accepted procedure depends on the closest "distance" (with respect to similar, salient, related, familiar) to the prototypical element of a categories (e.g. Rosch and Lloyd (1978)), which then trigger their corresponding category(ies) members.

Properties on problem activated categories

Now we define a consistent transfer to activations of the categories in a database. Based on a natural categorization structure $\tilde{C} = \{\tilde{C}_1, \tilde{C}_2, ..., \tilde{C}_L\}$, we will define a function $\tilde{C}: X \times C^* \to C^*$ that determines the single pieces of information in a database $D \in C^*$ that belong to categories that are activated by a specific problem $x \in X$.

[13]Some research on this issue is mentioned in the introduction. Note that under automaticity subjects are often not even aware of this process.

[14]However, an outcome dependent categorization is in principle also possible.

[15]Such a categorization is only useful if cases with different characteristics are in the same categories.

Thereby, our main assumption concerns a minimal amount of distinct cases in a database that initiates categorical processing of information in agents' minds. For a less diverse or complex database (i.e. $div(D) \leq k$) an agent's brain does not start to simplify and reduce the set of information by categorizing the information, but will just process, inspect and take into account all single pieces of information directly. However, for more diverse or complex databases agent's mind will initiate a rough (problem evoked) classification of the database in accordance with the natural categorization $\tilde{C}$ and then consider in detail only the information in her database that are also contained in categories activated by the current problem.

Definition 4.4 ***Induced minimal categorization***
Let $\tilde{C}$ be a categorization structure on C and $k \in \mathbb{N}$ with $k \geq 3$. A database $\tilde{C}(x, D)$ results from a categorization function $\tilde{C} : X \times C^* \to C^*$[16] that categorizes each database $D \in C^*$ according to the categories in $\tilde{C}_x$ that are evoked by problem $x \in X$. We call $\tilde{C}(x, D)$ the x-evoked/activated categorized database D (for cognitive ability level k) and define it for all $D \in C^T$ for all $T \in \mathbb{N}$ by

$$\tilde{C}(x, D) := \begin{cases} D \cap (c^T)_{c \in C} & \text{for } div(D) \leq k \\ D \cap (\tilde{C}_x)^T & \text{for } div(D) > k. \end{cases}$$

In the following, we will fix $k = 3$ without loss of generality. Obviously $\tilde{C}(x, D) \subseteq D$.

Example:
In order to clarify the definition of the categorization function for a database $D \in C^T$ such that $div(D) > 3$, i.e. $\tilde{C}(x, D) := (\tilde{C}_x)^T \cap D$, consider the following situation. Let $D = (c_1^2, c_2, c_3^4, c_6, c_7^3) \in C^{11}$, i.e. $div(D) = 5$, and a categorization structure induced by problem $x \in X$ $\tilde{C}_x = (\tilde{C}_1, \tilde{C}_2)$, where $\tilde{C}_1 = (c_1, c_2)$ and $\tilde{C}_2 = (c_2, c_4, c_7)$. Then $(\tilde{C}_x)^{11} \cap D = (c_1, c_2)^{11} \circ (c_2, c_4, c_7)^{11} \cap D \approx (c_1^2, c_2, c_7^3)$ (up to order), which is normatively and descriptively appealing for such a categorized database.

Remark 4.1
(i) Since $\tilde{C}_x$ is insensitive to repetitions of cases, we need to ensure that repeated observations in $D \in C^T$ are captured, which is guaranteed by introducing the T-replicated $\tilde{C}_x$.[17]
(ii) An immediate consequence of the minimal induced categorization property is that

[16] With slight abuse of notation we name the function and the natural categorization identical to emphasize that for each fixed categorization structure a corresponding function can be defined, i.e. that the function relies on this fixed categorization structure like a parameter.

[17] We use the total length of the database just for simplicity. One might also take the maximal amount of a case appears in D.

for database D such that $div(D) \leq 3$, we have no categorization and the framework (and later defined axioms) coincides with the framework of BGSS. Thus, it enables us to mirror their proof for these kind of databases.

Apart from the minimality condition for an activated categorization of databases, the definition makes implicitly three additional important assumptions.

First, a very important ingredient of this definition is that the evoked categories $\tilde{C}_x$ are totally unrelated to the database under consideration. Only the underlying problem activates the relevant categories. In accordance with these categories the actually available information in the database is intersected. One may argue that a database itself determines which categories are evoked (or even formed), since the database might provide some intuition and motivation how to categorize it. However, our intuition runs solely through a new problem x that activates the relevant categories in the subconsciously pre-existing natural categorization $\tilde{C}$. In this sense the induced categorization of the database occurs not directly on the level of the database.

Second, the ordering of the database D does not affect the resulting categorized database. Any reordered database $\pi(D)$ of the database D results identically activated categorized information, i.e. the content of the categorized database $\tilde{C}(x, D)$ and $\tilde{C}(x, \pi(D))$ coincides. In this sense, the definition induces some categorization invariance that is driven by our assumption that the categorization is evoked by the underlying problem x and not by the database. This precludes that a categorization of a database is affected by order effects.[18] However, since the x-evoked categories are activated independent from any database in our approach, it is quite natural that their intersections with any reordered database result in the same content.

Finally however, even though the content of differently ordered database are identical after the categorization, the order of its evoked content (cases) in the categorized databases may still be different, i.e. $\tilde{C}(x, D)$ and $\tilde{C}(x, \pi(D))$ may consist of the same content but differently ordered. However, the definition precludes this difference by assuming a specific ordering for all categorized databases according to the order on the set of basic cases (induced by the definition of $\cap$ for databases and the order on C). Yet, the reason for this assumption is not that we want to restrict the categorization process on databases in this way, but it is rather an anticipation of a property (or axiom) we would enforce for the subsequent belief formation. In the manner of BGSS, an Invariance Axiom on beliefs would say that the belief induced by a database is determined only by its content and not its ordering in the database,

[18] For instance a first impression effect might might induce a bias for the category that the first case in the database is most related to.

i.e. the beliefs induced by any reordering of the same content coincides. Thus, the assumption of a specific (seemingly restrictive) order of the categorized databases is innocent in combination with an Invariance Axiom on the belief stage, since then the order of any categorized database would be immaterial for a belief.

4.5.2 Induced category based belief

A category based belief is composed of a usual belief $P : X \times C^* \to \Delta(R)$ and a previous problem evoked categorization of the underlying information $\tilde{C} : X \times C^* \to C^*$ (i.e. $\tilde{C}(x, D) \subseteq D$), such that $(P \circ \tilde{C}) : X \times C^* \to \Delta(R)$, i.e. $(P \circ \tilde{C})(x, D) = P(x, \tilde{C}(x, D))$. Faced with a new problem $x \in X$, the agent's brain activates or evokes some appropriate categories for this problem according to the natural categorization structure $\tilde{C}$ and forms the belief based only on those pieces of information in the activated categories that are actually available in her database, i.e. subdatabase $\tilde{C}(x, D) \subseteq D$.

In the following we will fix a problem $x \in X$ and write for convenience $\tilde{C}(x, D) = \tilde{C}(D)$ and for $(P \circ \tilde{C})(x, D) = (P \circ \tilde{C})(D)$ when no confusion arises.

4.5.3 Axioms on the level of belief formation

Categorized Invariance Axiom (already implied)
For all $D \in C^*$ and all permutations π on D, i.e. $D = (c_1, ..., c_T)$, then $\pi(D) = (c_{\pi(1)}, ..., c_{\pi(T)})$ the following holds:

$$(P \circ \tilde{C})(D) = (P \circ \tilde{C})(\pi(D)).$$

The axiom basically says that the order or sequence of appearance of the cases in D is immaterial for the induced category based belief, only the content matters. Thus, the axiom is directly implied by our definition of a problem evoked categorization of databases, since we discussed already that for any two databases containing the same content, their induced categorization coincide, i.e. $\tilde{C}(D) = \tilde{C}(\pi(D))$ for all databases $D \in C^*$ and reordered database $\pi(D)$. In this sense the categorized Invariance Axiom is superfluous and indirectly substituted by the definition of a categorized database.[19]

Per se the invariance property does not allow for different impacts if a case appears earlier or later in a database. However, the order in which information is provided

[19] Technically speaking, there is no difference between restricting the categorization process to specific orders or allowing for different orderings and requiring an Invariance Axiom for beliefs.

or obtained can influence the judgment strongly and may carry information by itself. One way to cope with these order effects is to describe the cases informative enough. E.g. if one wants to capture the position or time of occurrence of a case in the categorized database, one could implement this information into the description of the cases itself. Put differently, if one challenges the consequences of an invariance property, then there must be some criteria which distinguishes the cases and paying attention explicitly to this difference in the description of the case may lead the agent to reconcile with such an invariance.

Category based Concatenation Axiom
There exists some $\lambda \in [0,1]$, such that for $\tilde{C}(D \circ E) \neq \emptyset$

$$(P \circ \tilde{C})(D \circ E) = \lambda (P \circ \tilde{C})(D) + (1-\lambda)(P \circ \tilde{C})(E),$$

where $\lambda = 0$ if and only if $\tilde{C}(D) = \emptyset$.

In the following we will call the database which emerges from concatenation of other databases as the **combined or concatenated** database, whereas the databases used for the concatenation will be called **combining or concatenating** databases.

The category based Concatenation Axiom states that a category based belief induced by a concatenated database is a weighted average of the category based beliefs induced by its respective combining databases. The axiom captures the idea that a belief based on the combination of two databases can not lie outside the interval spanned by the beliefs induced by each combining database separately. Intuitively it can be interpreted in the following way (stated from an exclusion point of view): if the information in any database induces an agent's belief not to exclude an outcome r, then the outcome r cannot be excluded by the belief induced by the combination of all these databases.[20]

However, in order to sustain the normative appealing interpretation of averaging (category based) beliefs, the categorized concatenation of two databases must coincide with the concatenation of these two categorized databases, i.e. the union of the elements surviving the categorization process for each single database should not differ from the elements surviving the categorization of the database generated

[20] Of course the axiom is stronger in the sense that it not only requires that the probability of such an r is positive, but it should lie between the minimal and maximal assigned probabilities induced by the combining (filtered) databases.

by the combination of the two. This would ensure that a category based belief induced by the concatenated database relies on information that is also employed in the category based beliefs induced by the single concatenating databases. However, this is directly achieved by the definition of a problem evoked categorization of a database, i.e. for $D \in C^T$ and $E \in C^L$

$$\tilde{C}(x, D \circ E) = (\tilde{C}_x)^{T+L} \cap (D \circ E) = ((\tilde{C}_x)^T \cap D) \circ ((\tilde{C}_x)^L \cap E) = \tilde{C}(x, D) \circ \tilde{C}(x, E)$$

The category based Concatenation Axiom assumes $\lambda = 0$ for databases such that $\tilde{C}(D) = \emptyset$. Of course $\lambda \neq 0$ would result in inconsistencies, since then $(P \circ \tilde{C})(D \circ E) = \lambda P(\emptyset) + (1 - \lambda)(P \circ \tilde{C})(E)$, which implicitly states that the category based beliefs induced by $\tilde{C}(D \circ E)$ and $\tilde{C}(E)$ would differ, even though the categorized databases coincide.

Collinearity Axiom
No three of $\{((P \circ \tilde{C})(c))_{c \in C}\}$ such that $\tilde{C}(c) \neq \emptyset$ are collinear.

Technically speaking this axiom allows to derive an unique similarity function (in combination with the other axioms), but it has also some reasonable intuition. Roughly it states that a (non trivial) estimate based on a case is never equivalent to the combined (non-trivial) estimates based on two other cases. Hence, a case is always informative in the sense that no combination of two other cases can deliver the same estimation and would make this case "redundant". By non trivial we mean that the case is activated (since a not activated case might only contribute a trivial (uninformed) uniform-like estimate).

4.5.4 Representation Theorem of category based belief formation

Theorem 4.1
Let there be a function $(P \circ \tilde{C}) : C^* \rightarrow \Delta(R)$, where $P : C^* \rightarrow \Delta(R)$ and $\tilde{C} = \{\tilde{C}_1, ..., \tilde{C}_L\}$ a categorization structure on C with corresponding induced minimal categorization function $\tilde{C} : C^* \rightarrow C^*$, i.e. for each $D \in C^*$ a categorized database $\tilde{C}(D) \subseteq C^*$ is given. Let $(P \circ \tilde{C}) : C^* \rightarrow \Delta(R)$ satisfy the Collinearity Axiom.
Then the following are equivalent:
(i) The function $(P \circ \tilde{C})$ satisfies the category based Concatenation Axiom
(ii) There exists for each $c \in C$ a unique $P^c \in \Delta(R)$, and a unique strictly positive -up to multiplication by a strictly positive number- function $s : C \rightarrow \mathbb{R}_+$, such that

for all $D \in C^*$ with $\tilde{C}(D) \neq \emptyset$

$$(P \circ \tilde{C})(D) \quad = \quad \frac{\sum_{c \in \tilde{C}(D)} s(c) P^c}{\sum_{c \in \tilde{C}(D)} s(c)}.$$

Rough sketch of the proof:
The necessity part is straightforward calculation. The sufficiency part follows the rough structure of the proof of BGSS and Bleile (2014b) (or Chapter 3), but differs in the crucial arguments. The idea is to transform the framework from the space of databases to the space of frequency vectors that is structural more tractable, i.e. the category based belief $(P \circ \tilde{C})(D) = \frac{\sum_{c \in \tilde{C}(D)} s(c) P^c}{\sum_{c \in \tilde{C}(D)} s(c)(c)}$ based on database $D \in C^*$ translates to frequency vectors f_D by $(P \circ \tilde{C})(f_D) = \frac{\sum_{j \leq m} s_j \tilde{C}_j(f_D) P^j}{\sum_{j \leq m} s_j \tilde{C}_j(f_D)}$. In order to show that this is viable we exploit the structure of the categorization procedure and the category based Concatenation Axiom.
The essential part of the proof is to derive the similarity weights $(s_i)_{i \leq m}$. This will be shown inductively over $|C| = m$ and $div(f_E) \leq m$.

Step 1: Base case for the induction, i.e. $|C| = m = 3$. Since $\tilde{C}(f) = f$ for all f such that $div(f) \leq 3$, we are exactly in the BGSS framework, which directly deliver the result for these kind of frequency vectors.

Step 2: $|C| = m > 3$ and $div(f_E) \leq m$.
As in BGSS or Chapter 2, we can show (using $\tilde{C}(f) = f$ for all f such that $div(f) \leq 3$) that the similarity weights derived in Step 1 are independent of the triplet $\{i, j, k\}$ for any set of basic cases $C = \{c^i, c^j, c^k\}$ and thus we can define for all $f \in \Delta(C)$

$$(P \circ \tilde{C})_s(f) := \frac{\sum_{j \leq m} s_j \tilde{C}_j(f) P^j}{\sum_{j \leq m} s_j \tilde{C}_j(f)}.$$

The aim is to show $(P \circ \tilde{C})_s(f) = (P \circ \tilde{C})(f)$ for all $f \in \Delta(C)$ via induction over m and using Step 1 ($m = 3$) as base case.
Let $f = \alpha q^j + (1 - \alpha) f(j)$ (*) (for some $\alpha \in (0, 1)$) where $f(j)$ denotes the point in $conv(\{(q^l)_{l \in \{1,\ldots,m\} \setminus j}\})$ that is on the line through f and the j-th unit vector q^j, as in BGSS.
(i) If there exists a $j \leq m$ such that $q^j \notin \tilde{C}(f)$, then the decomposition (*) and the category based Concatenation Axiom (and induction assumption) delivers the claim.
(ii) If $q^j \in \tilde{C}(f)$ for all $j \leq m$, then there are m many q^j such that $\tilde{C}(q^j) \neq 0^m$.

Again, the category based Concatenation Axiom applied to the m many decompositions yields that $(P \circ \tilde{C})(f)$ lies in the interior of the intervals spanned by $((P \circ \tilde{C})(q^j), (P \circ \tilde{C})(f(j)))$ for all $j \leq m$. Combined with the Collinearity Axiom this delivers a unique intersection of these lines in $(P \circ \tilde{C})(f)$ and $(P \circ \tilde{C})_s(f)$, since the elements determining the lines satisfy already the claim by the induction assumption.

Interpretation of Theorem

A category based belief formation can be interpreted as a two stage process in which in an initial step the rough categorized information is activated by the current problem and in a subsequent step the information contained in these activated categories are processed and evaluated in detail for the belief formation.
A category based belief follows exactly the experimental evidence in psychology in which individuals focus on the category(ies) that a problem belongs to or are most relevant and fitting (Murphy and Ross (1994)) and all other categories are immaterial, not retrieved and excluded. This implies that an agent does not need to retrieve or consider all potentially memorized or all past cases (as in BGSS, EG, Bleile (2014a)). In this way, such a procedure may reduce enormously the cognitive effort, since only a subset of past cases is processed in detail and all pieces of information that are members of irrelevant categories are not even needed to be retrieved.

A category based belief is not based on estimations that are associated with entire categories, but it relies on all estimates induced by the single cases in the categories activated by the problem. This is an important distinction to the axiomatized prototype based belief in the next section and reflects the disagreement in the categorization literature on how categories are actually represented. One stream of literature argues for a representation through all its members (Kruschke (1992), Medin and Schaffer (1978)), whereas another branch reasons in favor of an abstracted summary in terms of a prototype representation (e.g. Rosch and Mervis (1975) and references in Murphy (1994)).[21]

Since a category based belief relies only on the information of some activated categories, it can be interpreted as a limited attention model as in Bleile (2014b),

[21] In general, there is tradeoff between informativeness and economy involved that might be better balanced by a more intermediary representation as offered in VAM models (Vanpaemel and Storms (2008)), which allows for varying abstraction levels. It seems that people shift from using a prototype representation early in training to using an exemplar representation late in training.

where a filtered belief[22] is based only on some parts of the potentially available information (so called consideration set that survives a screening/filtering stage). For a category based belief, the information in the evoked categories can be identified as the information that is contained in such a consideration set. From this point of view, the filtering runs on the category level and the categories are roughly screened with the purpose to "determine" whether they are appropriate or not for the current problem. The "surviving" categories are examined in full detail in the further belief formation. In this way, the category based belief formation can also be interpreted as an adaption of the choice model of Manzini and Mariotti (2012b) ("Categorize then choose") to belief formation. In particular, the category based belief can be embedded into a filtered belief in the following way. A filter process $\Gamma : X \times C^* \times C^* \to C^*$ is defined for any $x \in X$ by (i) $\Gamma(x, E, E) \neq \emptyset$ for all $E \in C^*$ and (ii) $\Gamma(x, D, E) = \Gamma(x, E, E) \cap D$ for all $D \subseteq E$. A problem evoked categorization of a database $E \in C^T$, i.e. $\tilde{C}(x, E)$, can be related to a specific filter defined by

$$\Gamma(x, E, E) := \begin{cases} E \cap (c^T)_{c \in C} & \text{for } div(E) \leq k \\ (\tilde{C}_x)^T & \text{for } div(D) > k. \end{cases}$$

Then $\Gamma(x, D, E)$ coincides with $\tilde{C}(x, D)$ for all $D \subseteq E$ and all the required properties of the filter process (order invariance, equal treatment of information, ignorance of repeated information, consideration property, minimal attention span) are met by the (x-evoked) categorization function. As well, the category based Concatenation Axiom translates into its filtered version. In this way, Theorem 4.1 is a corollary of representation theorem in Bleile (2014b) or Theorem 3.1. However, for the sake of completeness and self-containedness of the chapter, we state a direct (lighter) proof of Theorem 4.1 (in Section 4.8).

4.6 Axiomatization of prototype based belief formation

4.6.1 Specific properties of the categorization procedure

For a prototype based approach we will define a specific categorization of databases in accordance with a natural categorization structure, but independent of the current problem of categorizing information. In this regard it differs from the first procedure.

[22] A filtered belief $(P \circ \Gamma)(x, D, E)$ that is based only on some filtered $\Gamma(D, E) \subseteq D$ information captured in D, i.e. $(P \circ \Gamma) : X \times C^* \times C^* \to \Delta(R)$.

Natural categorization structure

Let there be given a natural (evolutionary optimal) categorization structure $\tilde{C}$ as discussed in Section 4.4. We slightly restrict the natural categorization structure such that it satisfies some additional structural properties regarding the content of the categories.

Definition 4.5
For a set of basic cases C we call $\tilde{C} = \{\tilde{C}_1, ..., \tilde{C}_L\}$ for some $L \in \mathbb{N}$ a categorization structure, if
(i) $\tilde{C}_k \cap \tilde{C}_l = \emptyset$ and
(ii) $C = \circ_{j \leq L} \tilde{C}_l$ and
(iii) for all $l < L$ $|\tilde{C}_l| \geq 2$ and $|\tilde{C}_L| \geq 1$

In contrast to the unrestricted natural categorization structure in the section before, now the categories are explicitly defined to be disjoint and should contain sufficiently many elements. Disjointness is natural when a category is identified by a (set of) property or attribute. An object does either possess a property or it does not, which implies the disjointness. We stick to disjoint categories mainly for reasons of technical and notational simplicity, but it could be generalized.
The assumption that almost all categories contain at least two members captures the motivation to deal with "real" categories.[23] Basically, a "real" category in our definition exists whenever at least two heterogenous or distinguishable cases can be gathered in the same category according to some common criteria. Categorization is only meaningful if some pieces of information can be classified into a common genuine category. We rule out a degenerate (trivial) singleton-categorization, in which all cases get their own category. Furthermore, the optimal categorization literature (Fryer and Jackson (2008), Mohlin (2014), Peski (2011)) supports our defining properties (under some mild conditions). It shows that there exist many more cases than categories and that an optimal categorization results in few and relatively coarse categories, which implies that "optimal" categories should contain many members, i.e. $|\tilde{C}_l| \geq 2$.

The underlying reason originates from a tradeoff between benefits and disadvantages of a fine or coarse categorization.[24] A finer categorization implies more

[23] Manzini and Mariotti (2012b) consider a similar requirement for their 2nd version of categorization.
[24] This is closely relate to the problem of over-fitting in statistics, where a too close/precise fit of limited observations- i.e. using high-dimensional models- comes with risk of loosing the predictive power.

categories that contain less but more homogenous and similar members, but results in a decreasing robustness or reliability of a prediction based on these (less precise or noisy) categories with less many observations. Another compromise concerns the increasing challenge in searching and identifying the "correct" category(ies) for new objects for finer (and more narrow) categorization (e.g. Medin (1983), Jones (1983)). Thus, an agent might prefer to categorize more coarsely into larger categories, which is well known and discussed in the psychology literature and referred to as basic level categories. Basic level categories are neither the most general nor the most detailed categories.

A categorization based solely on characteristics satisfies our definition under the condition that the L categories are not empty, since if a case $c = (x, r) \in \tilde{C}_l$, then the cases $c = (x, r_i)$ would be in the category $\tilde{C}_l$ for all $i \leq n$.

The assumption that there exists at most one singleton category reflects the intuition of a category that might collect the cases that are "uncategorizable" into "real" categories.

Specific database induced categorization

Based on a natural categorization structure, we define a specific categorization procedure for given databases that transmits the idea that there is one category that contains the "uncategorizable" elements. Further it assumes - similar to the problem evoked categorization of databases in Section 4.5.1 - that a minimum amount of complexity of the database is required in order to initiate categorical thinking and processing of information. For less diverse databases categorization is not necessary and the agent considers just all cases in detail.

Definition 4.6
Let $\tilde{C} = \{\tilde{C}_1, .., \tilde{C}_L\}$ be a categorization structure on C. For all $E \in C^*$ a categorization of E or E-categorization structure $\tilde{C}^E = \{\tilde{C}_1^E, ..., \tilde{C}_{L+1}^E\}$ is given in the following way:
(i) If $div(E) < 5$, i.e. $E = (c_i^r, c_j^s, c_k^t, c_n^u)$ for distinct $i, j, k, n \leq m$ and $r, s, t, u \in \mathbb{N}_0$, then $\tilde{C}^E = \{\tilde{C}_1^E = \{c_i\}, \tilde{C}_2^E = \{c_j\}, \tilde{C}_3^E = \{c_k\}, \tilde{C}_4^E = \{c_n\}, \tilde{C}_5^E = \emptyset, ..., \tilde{C}_{L+1}^E = \emptyset\}$
Basically $\tilde{C}^E = \{\{c\}_{c \in E}\}$
(ii) If $div(E) \geq 5$, then $\tilde{C}^E := \{\tilde{C}_1^E, ..., \tilde{C}_{L+1}^E\}$, where $\tilde{C}_l^E$ for $l \leq L$ is defined as

follows:

$$\tilde{C}_l^E = \begin{cases} \tilde{C}_l \cap E & \text{if } div(\tilde{C}_l \cap E) \geq 2 \\ \emptyset & \text{if } div(\tilde{C}_l \cap E) \leq 1 \end{cases} \quad \text{and}$$

$$\tilde{C}_{L+1}^E = \bigcup_{\{l \leq L | |\tilde{C}_l \cap E| = 1\}} \tilde{C}_l \cap E$$

Note that by definition $\tilde{C}_l^E$ (for all $l \leq L+1$) does not contain repetitions of cases. The difference in processing information depending on $div(D) \lesseqgtr 5$ captures the motivation to have at least three meaningful categories, in the sense that at least two "real" non singleton categories exist, which requires $div(D) \geq 5$.
A category $\tilde{C}_{L+1}^E$ of "uncategorizable" cases is supported by an implication of optimal categorizations (e.g. Fryer and Jackson (2008)) which shows that experiences and objects in databases that are not "easy" to categorize (i.e. tend to form a singleton category) are more coarsely categorized and more often lumped together (i.e. gathered in the category of uncategorizable elements $\tilde{C}_{L+1}$).

Remark 4.2
The definition of a categorization of a database implies some sort of categorization invariance, i.e. categorizations are independent of the order, number and frequency of cases in a database. More precisely,
$\tilde{C}^E = \tilde{C}^D$ *for all* $D, E \in C^*$ *containing the same cases, i.e. for all* $c \in C$ $f_D(c) > 0$ *if and only if* $f_E(c) > 0$.
In particular $\tilde{C}^E = \tilde{C}^{E^Z}$ *and* $\tilde{C}^E = \tilde{C}^{\pi(E)}$ *for all re-orderings* $\pi(E)$ *on* E.

Basically, we require that a category (and later its prototype) is not affected by repetitions of already observed information, indirectly saying that categories are characterized by single observations of different cases and not influenced by their frequencies. Interestingly, an optimal categorization procedure (a la Jackson and Fryer (2008)) results as well in categories that remain unchanged when experiences is simply replicated (and also their prototypes), i.e. $\tilde{C}_l^{E^Z} = \tilde{C}_l^E$ for some $Z \in \mathbb{N}$.[25]

The categories of an E-categorization structure (that an agent has in mind) can be evoked or activated by cases that are also contained in another (simultaneously) available (somehow related) database D.

[25] However, the optimal categorization of Fryer and Jackson is sensitive to additional already known information if it concerns only single pieces of information. Increasing the size of only a single group of cases may lead to a shift in the categorization.

Definition 4.7

Let $D, E \in C^*$, such that $f_D(c_i) \geq 0$ only if $f_E(c_i) > 0$ for all $i \leq m$. Then, the E-categories evoked by D result from an D-induced E-categorization function $\tilde{C} : C^* \times C^* \to P(C^*)$

$$\tilde{C}(D, E) := \tilde{C}^E(D) = (\tilde{C}_1^E(D), .., \tilde{C}_{L+1}^E(D)), \quad \textit{where for all } l \leq L+1$$

$$\tilde{C}_l^E(D) \quad := \quad \begin{cases} \tilde{C}_l^E & \textit{if } D \cap \tilde{C}_l^E \neq \emptyset \\ \emptyset & \textit{otherwise.} \end{cases}$$

The definition is basically some consistency condition (note that $\tilde{C}_l^E(E) = \tilde{C}_l^E$), but one can interpret it as well in the following way. An agent having already categorized a database E and is "simultaneously" faced with processing the less rich database D (consisting only of already categorized cases in E) will not forget her already "internalized" E-categorization structure. In particular, the cases in D activate some categories in the richer E-categorization structure $\tilde{C}^E$ and it might only happen that some of these categories are not activated by cases in D, i.e. if for a $l \leq L+1$ and all $c \in D$ $c \notin \tilde{C}_l^E$, then $\tilde{C}_l^E$ is not evoked by D. However, this interpretation does only apply to "simultaneously" available and actively categorized databases. It does not hold for the natural categorization structure that is subconsciously (evolutionary and automatically) anchored in the brain and thus not actively formed in a process. An important implication of the definition is that for any reordering $\pi(D)$ of the database D, we have the same list of evoked categories in terms of content as well as in terms of order, i.e. $\tilde{C}^E(D) = \tilde{C}^E(\pi(D))$. In combination with the Remark 4.2, this implies a categorization invariance, i.e. for any $(D, E) \in C^* \times C^*$ such that $f_D(c) \geq 0$ if $f_E(c) > 0$ and appropriate permutations π, π'

$$\tilde{C}^E(D) = \tilde{C}^{E^L}(D^Z) = \tilde{C}^{\pi(E)}(\pi'(D)). \tag{4.1}$$

This property of a database evoked categorization of information is restrictive, since the order of cases in D might affect the order in which the E-categories are activated.[26] However, similar as in Section 4.5.1, this specific assumption is not a property we want to enforce explicitly, but it is an anticipation of an Invariance Axiom on the belief level, we would have enforced if the evoked categorization structure would be order sensitive.

[26]With additional effort on notation and definitions we could take care of orders.

Admissability of database based categorization

We defined an environment in which pairs of (somehow nested) databases $(D, E) \in C^* \times C^*$ affect the belief formation. The richer database E induces some categorization, where the cases of D activate the E-categories for the actual process of belief formation. In such a framework not all potential combinations of databases are plausible and meaningful for a belief formation based on categorized information. Our admissibility condition specifies the circumstances under which categorization of information is normatively and descriptively reasonable.

As discussed already, a categorization heuristics is useful for sufficiently complex and diverse databases. However, a sufficiently diverse database is not directly complex, e.g. if it is classified into a single (or very few) large category(ies) or into almost singleton (very fine) categories. Basically, only if sufficiently many meaningful categories are evoked an agent starts to think and process information categorical and feels confident in relying on (summarized) information on the category level. For databases that involve only very few activated databases an agent may not want to rely on only coarse (imprecise) summaries of these few categories, but might go through the information case by case in order to be sufficiently informed.[27] Thus, in such a situation a categorization of information does not offer some advantage to an approach of just taking into account all single pieces of information directly.

Our admissibility condition (i.e. (ii) and (iii)) restricts the pairs of databases for which an agent starts categorical thinking and processing. It requires that a minimum number of "real" categories of $\tilde{C}^E$ (namely three) are in some sense activated by a database D and considered for its evaluation.

Definition 4.8
The admissible pairs $(D, E) \in C^* \times C^*$ are given by the set A as follows

$$A := \Big\{(D,E) \in C^* \times C^* | \quad \begin{array}{ll} (i) & f_D(c) > 0 \textit{ then } f_E(c) > 0 \textit{ for all } c \in C \\ (ii) & \textit{if } div(D) = 2 \textit{ then there must exist } c \in E \backslash D \\ & \quad \textit{such that } |\tilde{C}^E(D \circ c)| = 3 \\ (iii) & \textit{if } div(D) \geq 3 \textit{ then } |\tilde{C}^E(D)| \geq 3 \end{array} \Big\}$$

Note that $(D, E) \in A$ if and only if $(D^Z, E^L) \in A$.
A necessary condition for a "real" categorization is $div(E) \geq 5$. For less diverse E each contained case is interpreted as singleton category and all conditions in A are

[27] As in the setup of a category based belief or the original belief formation without a categorization as taken in BGSS, EG, Bleile (2014a).

naturally satisfied. Thus, we need to discuss the admissibility conditions only for more diverse databases that induce non trivial singleton categories.

For condition (ii) and (iii) to be satisfied we need $|\tilde{C}^E| \geq 3$, which captures our motivation to employ only sufficiently "rich" categorizations. Consequently, those pairs (D, E) are ruled out such that $|\tilde{C}^E| < 3$. Part (iii) captures explicitly the intuition of "satisfactorily" many activated E-categories, by enforcing that there must exist at least three different cases in D that evoke three different E-categories.

For database D such that $div(D) = 2$ part (ii) requires that both contained different cases need to belong to different categories according to the E-categorization structure. The underlying idea is that D activates two different E-categories and thus triggers (makes aware) some categorical thinking and processing. However, if the database D evokes just one category $\tilde{C}_l^E$ in the E-categorization, then an agent might not initiate any categorical thinking and processing at all or is just not aware of different categories, but might rely on each single case directly. This is exactly ruled out by condition (ii).[28]

The admissibility condition is justifiable in general, but our initial motivation originates in the most interesting situation $D = E$. The condition (i) is directly met. A pair (D, D) with $div(D) = 2$ is not admissible. This matches our desire that no categorization is induced for databases that has no sufficiently complex categorization - i.e. not at least three categories- such that a categorization procedure offers (summarized) information on the category level for acceptable many categories. For more diverse databases, $div(D) \geq 3$, the admissibility just requires that the D-categorization structure consists of at least three different categories, as we desire.

4.6.2 Induced prototype based belief

An agent forms a prototype based belief based on her available admissible pair of information $(D, E) \in C^* \times C^*$ in the following way. Based on a natural categorization structure $\tilde{C}$, a categorization of a database E result in the categorization $\tilde{C}^E$. An agent evaluates the simultaneously available database D by exploiting the categorized information $\tilde{C}^E$ contained in the richer database E that is activated by cases in D.

Thus, a prototype based belief relies on categories in E that are evoked/activated by

[28]Note that for all $D \subseteq E \in C^*$ with $div(D) = 1$ (D, E) is admissible, in particular for any $E \in C^*$ and all $c \in E$ $(c, E) \in A$. This requires an exception of the given interpretation, since it is obvious that only one E-category can be activated. However, this is driven by the situation $D = E = (c^T)$, which implies a belief only based on information c (and single category member), which is acceptable in the light of the desired representation in Theorem 4.2.

the D induced categorization function $\tilde{C} : C^* \times C^* \to C^*$ (i.e. $\tilde{C}(D, E) = \tilde{C}^E(D)$) such that $(P \circ \tilde{C}) : X \times C^* \times C^* \to \Delta(R)$, i.e. $(P \circ \tilde{C})(x, D, E) = P(x, \tilde{C}(D, E)) = P(x, \tilde{C}^E(D))$.[29]
$(P \circ \tilde{C})$ is a belief induced by the categories in the richest database available (i.e. E) that are activated by single pieces of information in the database under consideration (i.e. D). The most intuitive situation is where $E = D$, in which all D-categories are employed.

Throughout the paper the problem x is fixed, therefore x is often suppressed in the following, i.e. $(P \circ \tilde{C})(x, D, E) = (P \circ \tilde{C})(D, E)$.

4.6.3 Axioms on the level of belief formation

As already mentioned above, our (restrictive) assumptions on the categorization procedure (see Remark and 4.2 equation (4.1)) replaces the otherwise imposed Invariance Axiom.

Implied Invariance Axiom
For all admissible pairs $(D, E) \in A$ and $D \in C^T$ and all permutations π on T for all T, we have

$$(P \circ \tilde{C})(D, E) = (P \circ \tilde{C})(\pi(D), E)$$

The Invariance Axiom says that the order of the information in the database is immaterial for the induced belief. Only their content is important. For a discussion of the axiom see Section 4.2.

Remark 4.3
Note that in particular $(P \circ \tilde{C})(D, E) = (P \circ \tilde{C})(\pi(D), \pi'(E))$ *by Remark 4.2.*

Thus the order invariance accounts for both databases, which is important for the proof.

Prototype based Concatenation Axiom:
For all $D, E, F \in C^*$ such that $(D, F) \in A$, $(E, F) \in A$ and $(D \circ E, F) \in A$ are admissible pairs, then there exist $\lambda \in (0, 1)$ such that:

$$(P \circ \tilde{C})(D \circ E, F) = \lambda(P \circ D)(D, F) + (1 - \lambda)(P \circ \tilde{C})(E, F).$$

[29] I.e. $P : X \times P(C^*) \to \Delta(R)$ relies only on categories.

The interpretation of the axiom is similar to the Concatenation Axiom in BGSS and as in Section 4.5.3. In order to keep the normatively desirable spirit of averaging, we need to ensure that the information employed in the belief formation for the concatenation $D \circ E$ is meaningfully related to the single databases D and E. Intuitively, the categories evoked by $(D \circ E)$ need to be covered by the categories evoked by either D or E. This can only hold in general if a common categorization structure underlies all involved activation processes, i.e. for a common $\tilde{C}^F$. A determination of a belief as an average of two other beliefs would be hard to justify if the underlying beliefs rely on different categorizations. If so, there might be very different categories involved in the different beliefs that would prevent an easy averaging, since no common evaluation basis exists. Of course, to be able to activate some categories from this common categorization structure, all observed cases in D and E need to be categorizable with regard to this common basis. Thus, the F-categorization ought to cover at least the available information in $D \circ E$, i.e. $F \supseteq D \circ E$. Moreover, having this categorization structure in mind, it appears reasonable to employ it in the belief formation process and not to shift to another less rich categorization, e.g. like moving to a E-categorization for database E, i.e. $\tilde{C}^E(E)$. Thus, the assumed structure ensures that $\tilde{C}^F(D \circ E) = \tilde{C}^F(D) \cup \tilde{C}^F(E)$ and therefore the induced beliefs rely on the same F-categories and are only distinct in the way which categories their contained cases evoke.
The prefix "prototype based" Concatenation Axiom or belief will become clear in the representation in Theorem 6.1, which suggests an interpretation of the axiom in terms of category related prototypes.

Identity Axiom
Let $(D_1, E), (D_2, E) \in A$ and related categorization structure $\tilde{C}^E = \{\tilde{C}_1^E, .., \tilde{C}_{L+1}^E\}$. For all D_1, D_2 such that $\tilde{C}^E(D_1) = \tilde{C}^E(D_2) = \tilde{C}_l^E$ for some $l \leq L+1$, then $(P \circ \tilde{C})(D_1, E) = (P \circ \tilde{C})(D_2, E)$.

The axiom says that the prototype based beliefs induced by databases (given an E-categorization structure) coincide if the databases evoke only one identical E-category. In this situation, the specific content of the information is immaterial for the induced belief, since only the activated category is relevant.

Collinearity Axiom
For all databases $D \in C^*$, no three distinct vectors of $\{((P \circ \tilde{C})(c, D))_{c \in D}\}$ are collinear.

The interpretation is the same as in Section 4.5 or BGSS.
The only difference is the requirement of distinctiveness, since $(P \circ \tilde{C})(c, D)$ is identical for cases in D that are contained in (and activate) the same category. Basically, no three (prototypical) estimates based on different categories are collinear.

4.6.4 Representation Theorem of prototype based belief formation

Theorem 4.2
Let $(P \circ \tilde{C})$ be a function $(P \circ \tilde{C}) : C^* \times C^* \to \Delta(R)$, where $\tilde{C} : C^* \times C^* \to P(C^*)$ is an induced categorization function with underlying E-categorization structures $\tilde{C}^E = \{\tilde{C}_1^E, ..., \tilde{C}_{L+1}^E\}$ for all $E \in C^*$. Let the prototype based belief $(P \circ \tilde{C})$ satisfy the Collinearity Axiom.
Then the following are equivalent
(i) $(P \circ \tilde{C})$ satisfies the prototype based Concatenation and Identity Axiom
(ii) For all $E \in C^*$ and any $l \leq L + 1$ such that $\tilde{C}_l^E \neq \emptyset$ there exists a unique $P^{\tilde{C}_l^E} \in \Delta(R)$ and a unique -up to multiplication by a strictly positive number- strictly positive function $\tilde{s} : C^* \times C^* \to \mathbb{R}_+$, such that for all $(D, E) \in A$ the following representation holds:

$$(P \circ \tilde{C})(D, E) \quad = \quad \frac{\sum_{l=1}^{L+1} \tilde{s}(D, \tilde{C}_l^E) P^{\tilde{C}_l^E}}{\sum_{l=1}^{L+1} \tilde{s}(D, \tilde{C}_l^E)}.$$

Moreover, $\tilde{s}(D \circ c, \tilde{C}_l^E) > \tilde{s}(D, \tilde{C}_l^E)$ for all $c \in \tilde{C}_l^E$.

Rough sketch of the proof:
The necessity part is straightforward calculation. The sufficiency part follows the rough structure of the proof of BGSS and Bleile (2014b) (or Chapter 3), but differs in the crucial arguments. Again, the first step is to reason, why it is viable to transform $(P \circ \tilde{C})(D, E)$ for $(D, E) \in A$ into $(P \circ \tilde{f})(f_D, f_E)$ for appropriate adjusted definitions of admissible pairs (f_D, f_E).
The essential part of the proof is to derive the similarity weights $(s_i)_{i \leq m}$. This will be shown inductively over $div(f_D) \leq m$.

Step 1: Base case for the induction, take any triplet $\{i, j, k\} \subset \{1, .., m\}$ such that $f_D \in conv(\{(q^v)_{v \in \{i,j,k\}}\})$, i.e. $div(f_D) \leq 3$, and $(f_D, f_E) \in A$.
(i) For $(f_D, f_E) \in A$ such that $2 \leq div(f_D) \leq 3$ and $3 \leq div(f_E) \leq 4$ the categorization procedure is vanished and coincide with the BGSS framework.

(ii) For $(f_D, f_E) \in A$ such that $div(f_D) \leq 3$ and $div(f_E) \geq 4$.
(a) For $f_D \in (q^i, q^j)$ and $(f_D, f_E) \in A$, there exist some k-th unit vector q^k such that admissibility condition (ii) holds. Then take the simplex spanned by $\{q^i, q^j, q^k\}$ and adopt Step 1 of BGSS, Bleile (2014a,b), i.e. find s_i, s_j, s_k, define $(P \circ \tilde{f})_s(f_D, f_E)$ and run the recursive procedure to cover all simplicial points on this simplex for the fixed f_E. This yields that for $(P \circ \tilde{f})_s(f_D, f_E) = (P \circ \tilde{f})(f_D, f_E)$ for all $f_D \in conv(\{q^i, q^j, q^k\})$ and fixed f_E.
(b) Since for an admissible pair $(f_D, f_E) \in A$ from (a) there might exist many q^k such that the admissibility condition (ii) holds. Repeat (a) for all such k.
(c) Since for $f_D \in (q^i, q^j)$ in (a) there exist many f_E such that (f_D, f_E) is admissible, repeat (a) and (b) for all these f_E.
(d) Repeat (a), (b) and (c) for any pair (q^i, q^j) with distinct $i, j \leq m$.
Thus, we have that $(P \circ \tilde{f})_s(f_D, f_E) = (P \circ \tilde{f})(f_D, f_E)$ for all admissible pairs (f_D, f_E) such that $div(f_D) \leq 3$.

Step 2: For all $(f_D, f_E) \in A$ such that $div(f_D) > 3$
By the properties of the categorization, we know that at least two f_E-categories contain a least two cases and another category contains at least one member, e.g. $\tilde{f}_1^{f_E} \supseteq \{q^1, q^i\}$, $\tilde{f}_2^{f_E} \supseteq \{q^2, q^j\}$, etc.
Let $f_D = \alpha q^j + (1 - \alpha) f(j)$ (for some $\alpha \in (0, 1)$) where $f(j)$ denotes the point in $conv\{(q^l)_{l \in \{1,\ldots,m\} \setminus j}\}$ that is on the line through f_D and q^j, as in BGSS.
Then the prototype based Concatenation Axiom and the induction assumption delivers

$$(P \circ \tilde{f})(f_D, f_E),\ (P \circ \tilde{f})_s(f_D, f_E) \in \bigcap_{t=1,2} ((P \circ \tilde{f})(q^t, f_E),\ (P \circ \tilde{f})(f(t), f_E)).$$

Applying the Collinearity Axiom yields the uniqueness of the intersection and the desired result.

Interpretation of Theorem

The theorem is described for general pairs of admissible databases $(D, E) \in A$. However, the most interesting and natural situation is given by identical information (D, D), which motivated our examination of a belief formation based on categorized information. For $(D, D) \in A$, an agent only employs the categorized information in database D and is not involved in the process of case based activation of categories as necessary for pairs $(D, E) \in A$. For (D, D), the admissibility condition for a "real" categorization of the information D requires that the database D is categorized at least into three different categories (which is in principle only possible for databases with at least five different cases). Thus, the most interesting and meaningful situation is for (large) databases that allow sufficiently rich categorization structures.

A prototype based belief formation does not focus on employing the information contained in the most appropriate evoked (target) categories for a problem, but it takes into account the entire categorized information in a database. The belief uses the summarized category based information across all categories. This tries to compensate for potential misassignments if the actual problem does not allow for straightforward most appropriate "target" category(ies). The process is not based on a detailed piece by piece evaluation of all cases and their induced estimates separately, but relies on the summarized coarse information on the category level. This is in line with the procedure in Anderson (1991). In particular, an agent only needs to compare and balance the categories (as an entity) at large and use their category specific predictions. Thus not all single pieces of information need to be evaluated, which is a severe (cognitive) simplification and captures the underlying aspect and motivation of a categorization (heuristic). The category based estimation $P^{\tilde{C}_l^D}$ is the main ingredient and eponymous for our belief, since it can be interpreted as representative or prototypical estimate associated with the category $\tilde{C}_l^D$. Each category has a unique representative prototypical estimate, which does not distinguish between between cases in the same category, i.e. for all $c \in \tilde{C}_l^E$ $(P \circ \tilde{C})(c, E) = P^{\tilde{C}_l^E}$. Implicitly, this means that a category is understood in terms of a prototypical element that captures, compresses, aggregates, summaries and abstracts the essence and central tendency of a category (for prototype theory see e.g. Posner and Keele (1968), Reed (1972)). A specific representation of such an aggregated prototype based estimate is not implied, but a very natural prototype is simply the mean across previously experienced objects in the category. But also other statistics can serve as prototypes.

The weights $(\tilde{s}(\tilde{C}_l^D, D))_l$ that are assigned to each category specific estimate $P^{\tilde{C}_l^D}$ reflect the relevance of category l in the database D for the current problem. The weights do not only measure the relevance of the categories $\tilde{C}_l^D$, but also incorporate how often (in similarity weighted terms) this category is activated by the specific database D.[30] Thus, the specific content and structure (frequencies) of the (activating) databases are taken care of, i.e. two databases that are categorized in identical categories can induced different beliefs if they contain differently many

[30] Remember that a category does not contain repetitions of cases, i.e. $\tilde{C}_l^D = \tilde{C}_l^{D \circ c}$ for any $c \in \tilde{C}_l^D$.

cases of specific types, since then the relative relevance of the same categories can be altered.

In sum, a prototype based belief formation facilitates fast and cognitive less demanding predictions compared to "smoother" forms of similarity based reasoning as for instance provided by kernel-based predictions or BGSS, EG and Bleile (2014a). These approaches need to incorporate all single pieces of information. For the prototype based approach, an agent simply evaluates a problem in terms of prototypical thinking and reasoning, by averaging the categories' prototypical estimates, which can be derived, stored and retrieved solely on the category level, independent of the problem. A kernel based or BGSS belief requires a higher and more complex cognitive load and task by the need to store a large amount of information and generate many more (conditional) estimations that complicates the belief formation.

4.7 Conclusion

Chapter 4 examines how beliefs are formed by agents that use a categorization procedure in order to process, store and retrieve the available information. The cognitive science literature emphasizes the role of categorical processing, thinking and reasoning. Based on this insight we axiomatize two stage belief formation procedures in which agents employ categorized information and do not incorporate the available information piece by piece. We assume that an agent is equipped (or has acquired (evolutionary)) subconsciously with some intuition or heuristic how to (optimally) categorize the entire set of possible pieces of information. Based on such a naturally given categorization heuristic, we introduce a procedure that consistently categorizes databases. We consider two well known and observed procedures of categorizations differing in how categories are activated for a new problem and how they are represented. One procedure relies only on the information in specific "target" categories that are the most appropriate categories for the current problem (our axiomatized category based belief formation). Another procedure relies on all categories of the database. However, they are not represented by their contained single pieces of information, but rest on so called prototypical elements that represent a summary or central tendency of the category (our axiomatized prototype based belief formation). These two versions of belief formation based on categorizations are represented as weighted sums of estimates induced by past respectively categorized and represented information. The weights, that are assigned to the different estimates, measure the similarity of the current problem with the single piece of information that induced the estimate or respectively its relevance with the particular category (or its proto-

type).
For both procedures, we require a minimum amount of complexity/diversity of the underlying information such that an agent really engages into categorical processing and thinking. Otherwise an agent sticks just to piece by piece evaluation of the information.
The axioms on the belief level are closely related to the axioms introduced in BGSS and Bleile (2014b) (or Chapter 3) and modified in a way to capture the categorization of information and their consequences for induced beliefs.
Compared to the beliefs axiomatized in BGSS, EG and Bleile (2014a), both belief formations based on categorized information reduce the cognitive effort extensively. For the category based belief an agent only needs to consider, evaluate and estimate each single piece of information within the target categories, i.e. only some subset of the available information. In the prototype based belief, an agent even thinks entirely categorical or in prototypes and thus treats information always on an aggregate level.

4.8 Proof of Theorem 4.1 (Category based belief formation)

It is straightforward to show that the representation satisfy the axioms.
The difficult part is the sufficiency direction, i.e. axioms imply representation. As before, the essential step in the proof will be to identify database with their frequency vectors, which allows to exploit the more tractable structure of the space of frequencies on C and to adopt the approach taken in BGSS (and use the mechanism of Bleile (2014a) (or Chapter 2)). However, since in addition a categorization step is involved, the crucial steps in the inductive proof require different arguments.
The notation and definitions on the space of frequency vectors are identical as in Section 3.9.2.

4.8.1 Translating the database framework into frequencies

Why is it viable?

Remember that we fixed a categorization structure $\tilde{C}$ and a problem $x \in X$.

In the following, we show that a consistent transformation from databases to frequencies is viable. Roughly, we want to identify a problem evoked categorization

$\tilde{C}(D)$ of a database D by its frequency vector in $\Delta(C)$ such that $\tilde{C}(f_D) \in \Delta(C)$ corresponds to $\tilde{C}(D) \in C^*$ within the category based belief formation, i.e. such that $(P \circ \tilde{C})(D)$ corresponds to $(P \circ \tilde{C})(f_D)$. For this purpose, we exploit the structure of the categorization procedure and the axioms on the belief formation stage.
We need to show that the two stage procedure is independent of the order of the involved database and its length. However, as already discussed in Section 4.5.1, for the specific categorization procedure only the content matters and the order of cases is irrelevant, i.e. $\tilde{C}(D) = \tilde{C}(\pi(D))$ for any reordering $\pi(D)$ of the database D. Furthermore, the length of the database is immaterial, since the category based Concatenation Axiom implies that $(P \circ \tilde{C})(D^Z) = \sum_{i \leq Z} \lambda_i (P \circ \tilde{C})(D) = (P \circ \tilde{C})(D)$ for all $Z \in \mathbb{N}$ and appropriate $\lambda \in \Delta^Z$. Consequently, for a category based belief we can identify any database $D \in C^*$ by its frequency vector f_D, i.e. the category based belief translates from $(P \circ \tilde{C}) : C^* \to \Delta(R)$ to $(P \circ \tilde{C}) : \Delta(C) \to \Delta(R)$ by $(P \circ \tilde{C})(f_D) := (P \circ \tilde{C})(D)$.
We need to reformulate the categorization, axioms and results from databases to frequency vectors, given a fixed problem $x \in X$.

Categorization in frequency terms

Definition 4.9
(i) Given a natural categorization $\tilde{C} = (\tilde{C}_1, .., \tilde{C}_L)$, the list of x-evoked categories on C, i.e. $\tilde{C}_x = (\tilde{C}_l)_{l \in M_x} \subseteq \tilde{C} \subseteq C^*$ for a corresponding $M_x \subseteq \{1, .., L\}$ translates to a x-evoked categorized frequency vector $\tilde{C}^x \in \Delta(C)$:

$$\tilde{C}^x \;=\; \Big(\frac{1}{|\tilde{C}_x|} \sum_{l \in M_x} 1_{\{\tilde{C}_l\}}(c^1), ..., \frac{1}{|\tilde{C}_x|} \sum_{l \in M_x} 1_{\{\tilde{C}_l\}}(c^m)\Big) \in \Delta(C),$$

which describes how often the ordered cases in C appear in the list of evoked categories.

(ii) A x-evoked categorization function for a database $D \in C^*$, i.e. $\tilde{C}(x, D) \in C^*$ translates to a x-evoked categorization function for a frequency vector $f = f_D$, i.e. $\tilde{C} : X \times \Delta(C) \to \Delta(C) \cup 0^m$ such that $\tilde{C}(x, f) = (\tilde{C}_1(x, f), ..., \tilde{C}_m(x, f)) \in \Delta(C) \cup 0^m$ is defined for $j \leq m$ by

$$\tilde{C}_j(x, f) = f_j 1_{\{div(f) \leq 3\}}(f) + \frac{f_j 1_{\{\tilde{C}^x\}}(q^j)}{\sum_{i \leq m} f_i 1_{\{\tilde{C}^x\}}(q^i)} 1_{\{div(f) > 3\}}(f).$$

$\tilde{C}(x, f)$ is the frequency vector of a categorized database D (represented by frequency

vector f) evoked by a problem x, i.e. describes the resulting frequencies of the cases in D that are also contained in the x-evoked categories.

Note that $\tilde{C}(f_D)$ represents $\tilde{C}(E)$ for all $E = \pi(D^Z)$ and any $Z \in \mathbb{N}$.
As before, we will suppress a fixed x, i.e. $\tilde{C}(x, f) = \tilde{C}(f)$.

Axioms in frequency terms

Category based Concatenation Axiom
For all $T_i \in \mathbb{N}$ $(i = 1, 2)$ and any $f_i \in \Delta_{T_i}$, there exists $\lambda \in [0, 1]$, such that for $f = \frac{T_1}{T_1+T_2} f_1 + \frac{T_2}{T_1+T_2} f_2$

$$(P \circ \tilde{C})(f) = \lambda (P \circ \tilde{C})(f_1) + (1 - \lambda)(P \circ \tilde{C})(f_2),$$

where $\lambda = 0$ if and only if $\tilde{C}(f_1) = 0^m$.

Collinearity Axiom
No three of $\{((P \circ \tilde{C})(q^j))_{j \leq m}\}$ such that $\tilde{C}(q^j) \neq 0^m$ are collinear.

4.8.2 Theorem 4.1, sufficiency part in frequency terms

Proposition 4.1
Let there be a function $(P \circ \tilde{C}) : \Delta(C) \to \Delta(R)$, where $P : \Delta(C) \to \Delta(R)$ and $\tilde{C} : \Delta(C) \to \Delta(C)$ a categorization function on the set of frequency vectors.
If $(P \circ \tilde{C})$ satisfies the category based Concatenation and Collinearity Axiom, then there exist unique probability vectors $(P^j)_{j \leq m} \in \Delta(R)$, and unique -up to multiplication by a strictly positive number- strictly positive numbers $(s_j)_{j \leq m} \in \mathbb{R}$, such that for all $q \in \Delta(C)$ such that $\tilde{C}(q) \neq 0^m$

$$(P \circ \tilde{C})(q) = \frac{\sum_{j \leq m} s_j \tilde{C}_j(q) P^j}{\sum_{j \leq m} s_j \tilde{C}_j(q)}, \tag{4.2}$$

where $\tilde{C}_j(q)$ denotes the frequency of case c_j in $\tilde{C}(q)$.

4.8.3 Proof of Theorem 4.1, sufficiency part in frequency terms

Step 0:
Obviously, by the definition of the categorization on frequency vectors we need to choose $P^j = (P \circ \tilde{C})(q^j)$ for $j \leq m$, since $\tilde{C}_j(q) \in \{q^j, 0^m\}$ and for some q (e.g. for q s.th. $div(q) \leq 3$)) $\tilde{C}_j(q) = q^j$.

The aim is to find numbers $(s_j)_{j\leq m}$ such that representation (4.2) holds for all $q \in \Delta(C)$.

Step 1: $|C| = m = 3$

By the definition of the categorization function we have that $\tilde{C}(q) = q$ for all $q \in \Delta(C)$, since $div(q) \leq 3$. In such a situation a categorization of information does not take place and thus the framework coincides with the one in BGSS. Basically $(P \circ \tilde{C})(q)$ coincides with $P(q)$ in the BGSS framework for $div(q) \leq 3$ and therefore Step 1 of the proof in BGSS can be directly adopted for these frequency vectors.

Step 2: Now we consider $|C| = m > 3$.
Step 2.1: Defining the similarity weights

Using the considerations from Step 1 for all triplets $\{i, j, k\}$ and $C = \{c^i, c^j, c^k\}$ we can derive the similarity weights $(s_v^{\{i,j,k\}})_{v\in\{i,j,k\}}$ and we know that for all $q \in \Delta(\{q^i, q^j, q^k\})$, the following representation holds

$$(P^{\{i,j,k\}} \circ \tilde{C})(q) = \frac{\sum_{v\in\{i,j,k\}} s_v^{\{i,j,k\}} \tilde{C}_v(q) P^v}{\sum_{v\in\{i,j,k\}} s_v^{\{i,j,k\}} \tilde{C}_v(q)},$$

where for all $v \in \{i, j, k\}$ P^v are independent of the triplet $\{i, j, k\}$ (by Step 0) and $(s_v^{\{i,j,k\}})_{v\in\{i,j,k\}}$ are unique up to multiplication by a positive number.

With a similar reasoning as in the proof in BGSS Step 2.1 (or Section 2.9.7) and using again that $\tilde{C}(q) = q$ for q such that $div(q) \leq 3$, we can show that the similarity values $s_v^{\{i,j,k\}}$ are independent of the choice of i, j and k for all v.
Thus, given these $(s_v)_{v\leq m}$ we can define for all $q \in \Delta$

$$(P \circ \tilde{C})_s(q) := \frac{\sum_{j\leq m} s_j \tilde{C}_j(q) P^j}{\sum_{j\leq m} s_j \tilde{C}_j(q)}.$$

Obviously, $(P \circ \tilde{C})_s$ satisfies the category based Concatenation Axiom.

Step 2.2: Completion to all $q \in \Delta(C)$**, i.e. show that for all** $q \in \Delta(C)$ $(P \circ \tilde{C})_s(q) = (P \circ \tilde{C})(q)$

We proof this by induction over k for $q \in \Delta(C)$ such that $div(q) = k$.
By Step 1 we know that the claim $(P \circ \tilde{C})_s(q) = (P \circ \tilde{C})(q)$ is true for all $q \in \Delta(C)$ such that $div(q) \leq 3$. This serves as the base case for the induction. Now we assume that $(P \circ \tilde{C})_s(q) = (P \circ \tilde{C})(q)$ for $q \in \Delta(C)$ such that $div(q) = k - 1$ and we will show it for $q \in \Delta(C)$ such that $div(q) = k$.

A similar construction as in BGSS, but with different reasoning yields the result. Let $q = \sum_{l \in K} \alpha_l q^l$ with $\alpha_l > 0$ and $K \subseteq \{1, ..., m\}$ such that $|K| = k$.
Define for all $l \in K$ the frequency vector $q(l)$ to be the vector in $conv(\{(q^j)_{j \in K \setminus l}\})$ such that q lies on the line generated by $(q(l), q^l)$. By the category based Concatenation Axiom there exists some $\lambda \in [0, 1]$ such that for all $j \in K$

$$(P \circ \tilde{C})(q) = \lambda (P \circ \tilde{C})(q^j) + (1 - \lambda)(P \circ \tilde{C})(q(j)).$$

We distinguish between two cases:

(i) If there exists a $j \in K$ such that $q^j \notin \tilde{C}(q)$, then the category based Concatenation Axiom implies for $q = \alpha q^j + (1 - \alpha) q(j)$ for an appropriate $\alpha \in (0, 1)$ that $(P \circ \tilde{C})(q) = (P \circ \tilde{C})(q(j))$, since $(P \circ \tilde{C})(q^j)$ receives zero-weight. The same is true for $(P \circ \tilde{C})_s$, since it satisfies also the category based Concatenation Axiom. However, since $div(q(j)) = k - 1$, the induction assumption yields $(P \circ \tilde{C})_s(q(j)) = (P \circ \tilde{C})(q(j))$ and we get the desired result:

$$(P \circ \tilde{C})(q) = (P \circ \tilde{C})(q(j)) = (P \circ \tilde{C})_s(q(j)) = (P \circ \tilde{C})_s(q).$$

(ii) If for all $j \in K$ $q^j \in \tilde{C}(q)$, then there are $k > 3$ many q^j such that $\tilde{C}(q^j) \neq 0^m$. Again, the category based Concatenation Axiom applied to the k many decompositions of $q = \alpha^j q^j + (1 - \alpha^j) q(j)$ for appropriate $\alpha^j \in (0, 1)$ for all $j \in K$ yields that $(P \circ \tilde{C})(q)$ lies in the interior of the intervals spanned by $((P \circ \tilde{C})(q^j), (P \circ \tilde{C})(q(j)))$ for all $j \in K$. Since for no three different $j \in K$ these intervals can lie on the same line by the Collinearity Axiom (since no three of $\{((P \circ \tilde{C})(q^j) = P^j)_{j \leq C}\}$ are collinear), there must exist some intersections of the lines. However, since $(P \circ \tilde{C})(q)$ lies in all these intervals, the intersection must be unique and exactly equal to $(P \circ \tilde{C})(q)$. However, also $(P \circ \tilde{C})_s(q)$ lies on all these intervals, since by induction assumption $(P \circ \tilde{C})(f) = (P \circ \tilde{C})_s(f)$ for all $f \in \Delta(C)$ such that $div(f) \leq k - 1$ (which q^j and $q(j)$ satisfy). Thus the unique intersection can only be $(P \circ \tilde{C})_s(q)$, which shows the equivalence of $(P \circ \tilde{C})(q) = (P \circ \tilde{C})_s(q)$. This completes the proof for $|C| > 3$ and eventually the Proposition 4.1 and Theorem 4.1. □

4.9 Proof of Theorem 4.2 (Prototype based belief formation)

It is straightforward to show that the representation satisfies the axioms.
To show that the axioms imply the representation requires some work. As before, we identify databases with their frequency vectors in order to adopt the approach taken in BGSS (and use the mechanism of Bleile (2014a) (or Chapter 2)). However, the additional categorization procedure alters the reasoning in the inductive proof significantly.
The notation and definitions necessary for the framework of frequencies are identical as in Section 3.9.2.

4.9.1 Translating the database framework into frequencies

Why is it viable?

We want to identify the prototype based belief induced by categorized databases, i.e. $(P \circ \tilde{C}) : C^* \times C^* \to \Delta(R)$ by a belief $(P \circ \tilde{f}) : \Delta(C) \times \Delta(C) \to \Delta(R)$ based on frequency vectors and their induced categorization structures $\tilde{f} \in P(\Delta(C))$, i.e. $\tilde{C}$ is represented by $\tilde{f}$ and $(P \circ \tilde{C})(D, E)$ by $(P \circ \tilde{f})(f_D, f_E)$.
Let $(D, E) \in A$.
(1) In a first step we exploit Remark 4.3, i.e. $(P \circ \tilde{C})(D, E) = (P \circ \tilde{C})(\pi(D), \pi'(E))$, where π, π' are permutations that reorder the cases in D and respectively in E arbitrarily. Basically it says that orders of databases are totally immaterial for the induced prototype based belief, i.e. only frequency vectors matter.
(2) a) The definition of a prototype based belief and database related categorization structures yields

$$(P \circ \tilde{C})(D, E) = P(\tilde{C}^E(D)) = P(\tilde{C}^{E^Z}(D)) = (P \circ \tilde{C})(D, E^Z) \text{ for all } Z \in \mathbb{N}$$

b) In addition the prototype based Concatenation Axiom implies for $D^Z = D \circ ... \circ D$ for all $Z \in \mathbb{N}$.

$$(P \circ \tilde{C})(D^Z, E) = (P \circ \tilde{C})(D, E).$$

Combining (2)a) and (2)b) yields for all $(D, E) \in A$ and $V, Z \in \mathbb{N}$

$$(P \circ \tilde{C})(D^V, E^Z) = (P \circ \tilde{C})(D, E) \tag{4.3}$$

and thus the lengths of the involved databases are immaterial as well.
Combining (1) and equation (4.3) shows that we can identify any $D, E \in C^*$ by their frequency vectors $f_D, f_E \in \Delta(C)$ for a prototype based belief formation.

Consequently, we can rewrite our framework into a frequency framework.

Categorization structures in frequency terms

Definition 4.10
A categorization structure $\tilde{f} = \{\tilde{f}_1, ..., \tilde{f}_L\}$ on $\Delta(C)$ satisfies the following properties
(i) $(\tilde{f}_l)_i := \frac{1}{|\tilde{C}_l|} 1_{\tilde{C}_l}(c_i)$ for all $l \le L$ and $i \le m$
(ii) $(\tilde{f}_k)_i > 0$ if and only if $(\tilde{f}_j)_i = 0$ for all $i \le m$ and any distinct $j, k \le L$ (i.e. disjointness)
(iii) $\sum_{l \le L} (\tilde{f}_l)_i > 0$ for all $i \le m$ (i.e. all cases are categorized)
(iv) for all $l \le L$ $(\tilde{f}_l)_i \le 1/2$ and $\tilde{f}_L(i) \le 1$ for all $i \le m$ (i.e. specific content structure)

Define for all $l \le L+1$ and any $q \in \Delta(C)$ $A(l, q) := \{j \le m | (\tilde{f}_l)_j > 0 \text{ and } q_j > 0\}$, i.e. the indices j such that case q^j contained in database q belongs to category $\tilde{f}_l$.

Definition 4.11
Let $\tilde{f} = \{\tilde{f}_1, ..., \tilde{f}_L\}$ be a categorization structure on $\Delta(C)$. A categorization structure $\tilde{f}^q = \{\tilde{f}^q_1, ..., \tilde{f}^q_{L+1}\}$ on q is defined by
(i) if $div(q) \le 4$, i.e. $q \in conv(\{q^h, q, q^j, q^k\})$, then $\tilde{f}^q = \{\tilde{f}_1 = q^h, \tilde{f}_2 = q^i, \tilde{f}_3 = q^j, \tilde{f}_4 = q^k, \tilde{f}_5 = 0^m, ..., \tilde{f}_{L+1} = 0^m\}$

(ii) if $div(q) > 4$, then $\tilde{f}^q$ is given for all $l \le L+1$ by

$$\begin{aligned}
\tilde{f}^q_l : &= \begin{cases} \left(\frac{1_{A(l,q)}(1)}{|A(l,q)|}, ..., \frac{1_{A(l,q)}(m)}{|A(l,q)|}\right) & \text{if } |A(l,q)| \ge 2 \\ \emptyset = 0^m & \text{if } |A(l,q)| \le 1 \end{cases} \\
\tilde{f}^q_{L+1} &= \frac{\sum_{l \le L} 1_{\{l \le L || A(l,q)|=1\}}(l) q^{A(l,q)}}{\sum_{l \le L} 1_{\{l \le L || A(l,q)|=1\}}(l)}
\end{aligned}$$

By the equation (4.1) it is also possible to define $\tilde{C}^E(D)$ in frequency term.

Definition 4.12
For any $q, e \in \Delta(C)$ such that $q_i \ge 0$ only if $e_i > 0$:
$\tilde{f}^e(q) = \{\tilde{f}^e_1 1_{\{q | \exists i \le m\ q_i > 0 \text{ and } \tilde{f}^e_1(i) > 0\}}(q), ..., \tilde{f}^e_{L+1} 1_{\{q | \exists i \le m\ q_i > 0 \text{ and } \tilde{f}^e_{L+1}(i) > 0\}}(q)\} \in P(\Delta(C))$
is the result of an q-induced e-categorization function $\tilde{f} : \Delta(C) \times \Delta(C) \to P(\Delta(C))$.

Admissibility condition in frequency terms

Definition 4.13
Define the set of admissible pairs $(q,e) \in \Delta(C) \times \Delta(C)$ by the following

$$A := \Big\{(q,e)| \quad \begin{array}{l} (i) \text{ for all } i \leq m \; q_i > 0 \text{ only if } e_i > 0 \\ (ii) \text{ if } div(q) = 2 \text{ then } \exists i \leq m \; q_i = 0 \text{ and } e_i > 0 \\ \quad s.\ th.\ \text{for } \alpha \in (0,1) \; |\tilde{f}^e(\alpha q + (1-\alpha)q^i)| = 3 \\ (iii) \text{ if } div(q) \geq 3 \text{ then } |\tilde{f}^e(q)| \geq 3 \end{array} \Big\}$$

Axioms in frequency terms

Prototype based Concatenation Axiom
For all $(\alpha q + (1-\alpha)q', e) \in A$ for some $\alpha \in (0,1)$, there exists $\lambda \in (0,1)$

$$(P \circ \tilde{f})(\alpha q + (1-\alpha)q', e) = \lambda (P \circ \tilde{f})(q,e) + (1-\lambda)(P \circ \tilde{f})(q',e).$$

Identity Axiom
For all $(q,e),(q',e) \in A$ such that there exists a unique e-category $\tilde{f}^e_l(q) \neq 0^m$ and $\tilde{f}^e_l(q') \neq 0^m$ and $\tilde{f}^e_j(q) = 0^m = \tilde{f}^e_j(q)$ for all $j \neq l \leq L+1$ it holds

$$(P \circ \tilde{f})(q_1, e) = (P \circ \tilde{f})(q_2, e).$$

Collinearity Axiom
No three distinct vectors of $\{((P \circ \tilde{f})(q^j,e))_{j \leq m}\}$ are collinear for any $(q^j, e) \in A$.

4.9.2 Theorem 4.2, sufficiency part in frequency terms

Proposition 4.2
Let there be given a function $(P \circ \tilde{f}) : \Delta(C) \times \Delta(C) \to \Delta(R)$, where $\tilde{f}$ a categorization function $\tilde{f} : \Delta(C) \times \Delta(C) \to P(\Delta(C))$ with related categorization structure $\tilde{f}^e$ for all $e \in \Delta(C)$. If the prototype based belief $(P \circ \tilde{f})$ satisfies the prototype based Concatenation, Identity and Collinearity Axiom.
Then for all $e \in \Delta(C)$ and each $l \leq L+1$ such that $\tilde{f}^e_l \neq 0^m$ there exist $P^{\tilde{f}^e_l} \in \Delta(R)$ and a strictly positive -and unique up to multiplication with a strictly positive number- values $s = (s_j)_{j \leq m}$ such that for all admissible $(q,e) \in A$

$$(P \circ \tilde{f})(q,e) = \frac{\sum_{l \leq L+1} \tilde{s}(\tilde{f}^e_l, q) P^{\tilde{f}^e_l}}{\sum_{l \leq L+1} \tilde{s}(\tilde{f}^e_l, q)}, \tag{4.4}$$

where $\tilde{s}(\tilde{f}_l^e, q) := \sum_{\{i \leq m | q^i \subseteq q \,\wedge\, q^i \subseteq \tilde{f}_l^e\}} q_i s_i$.

4.9.3 Proof of Theorem 4.2, sufficiency part in frequency terms

Step 0:
We get directly that for all $l \leq L+1$ and all $e \subseteq \Delta(C)$ $P^{\tilde{f}_l^e}$ must be chosen by $(P \circ \tilde{f})(q^j, e)$ for an appropriate $j \leq m$ such that $q^j \subseteq \tilde{f}_l^e$. By the Identity Axiom this is unique.

The aim is to find the $(s_j)_{j \leq m}$ such that the representation (4.4) holds for all admissible pairs $(q, e) \in A$. We proceed in two steps, where the first step considers $(q, e) \in A$ such that $div(q) \leq 3$ and in a second step we inductively generalize it to q with larger diversity.

Step 1: $(q, e) \in A$ **such that** $div(q) \leq 3$**,** i.e. take any triplet $\{i, j, k\} \subset \{1, .., m\}$ such that $q \in conv(\{(q^v)_{v \in \{i,j,k\}}\})$.

Step 1.1: $(q, e) \in A$ **such that** $div(q) = 1$ is covered in Step 0.

Step 1.2: $(q, e) \in A$ **such that** $2 \leq div(q) \leq 3$ **and** $3 \leq div(e) \leq 4$

Note that (q, e) with $div(e) = 2$ are not admissible.
By the definition of an e-categorization structure the categorization step vanishes for all such pairs (q,e). This means that the prototype based framework (and axioms) directly amounts to the BGSS framework, i.e. $(P \circ \tilde{f})(q, e)$ coincides with the belief $P(q)$ in BGSS. Applying their proof yields the desired representation for all (q, e) with the above properties.

Step 1.3: $(q, e) \in A$ **such that** $div(q) \leq 3$ **and** $div(e) > 4$

Step 1.3.1: Determine the similarity weights $(s_v)_{v \in \{i,j,k\}}$

Define for all triplets $\{i, j, k\} \subset \{1, .., m\}$ $q^*_{\{i,j,k\}} := \frac{1}{3}(q^i + q^j + q^k)$
Obvious $div(q^*_{\{i,j,k\}}) = 3$ and to fulfill the admissability take e such that $|\tilde{f}^e(q^*_{\{j,k,l\}})| = 3$ and hence each q^v for $v \in \{i, j, k\}$ needs to be contained in a different category of $\tilde{f}^e$.
We **assume for convenience that** $q^v \in \tilde{f}_v^e$ (it simplifies the notational effort extensively) for $v \in \{i, j, k\}$ which is possible after renaming the categories appro-

priately.
Observe that, since $(q^*_{\{i,j,k\}}, e)$ is admissible, also $(q, e) \in A$ for any $q \in conv(\{q^i, q^j, q^k\})$ and $q \subseteq e$.

Now, we use $q^*_{\{i,j,k\}}$ to determine the values $(s_v)_{v\in\{i,j,k\}}$ given in the representation of the theorem. By the prototype based Concatenation Axiom there exist $\lambda = (\lambda_v)_{v\in\{i,j,k\}} \in int(\Delta^3)$ such that

$$(P \circ \tilde{f})(q^*_{\{i,j,k\}}, e) = \sum_{v\in\{i,j,k\}} \lambda_v P^{\tilde{f}^e_v},$$

where we have used that $(P \circ \tilde{f})(q^v, e) = P^{\tilde{f}^e_v}$, as shown in Step 0.
The representation of the theorem (plugging in the definition of $\tilde{s}$) delivers

$$(P \circ \tilde{f})(q^*_{\{i,j,k\}}, e) = \frac{\sum_{v\in\{i,j,k\}} s_v 1/3 P^{\tilde{f}^e_v}}{\sum_{v\in\{i,j,k\}} s_v 1/3}$$

Equating these two equations and using that $(P^{\tilde{f}^e_v})_{v\in\{i,j,k\}}$ are not collinear yields a solution for $(s_v)_{v=i,j,k}$, (i.e. $s_v = \frac{\lambda_v}{\sum_{v\in\{i,j,k\}} \lambda_v}$). These s_v might depend on the specific triplet used in $q^*_{\{i,j,k\}}$, whereas obviously $P^{\tilde{f}^e_v} = (P \circ \tilde{f})(q^v, e)$ is independent of $\{i, j, k\}$, since only depending on e and q^v. However, similar as in Step 2.1 of the previous proof (or as in BGSS) we can show that the similarity values can be chosen independently of the particular triplet.
Thus, given these $s = (s_1, ..., s_m)$ we define for any triplet $\{i, j, k\} \subset \{1, ..., m\}$ and $q \in conv(\{q^i, q^j, q^k\})$ and e such that $(q, e) \in A$ the following prototype based belief

$$(P \circ \tilde{f})_s(q, e) := \frac{\sum_{v\in\{i,j,k\}} s_v q_v P^{\tilde{f}^e_v}}{\sum_{v\in\{i,j,k\}} s_v q_v}. \tag{4.5}$$

Recall that we assumed for convenience that $q^v \in \tilde{f}^e_v$, which allows this easy representation.
Furthermore, observe that P_s satisfies the prototype based Concatenation Axiom.

Step 1.3.2: Completion to all $(q, e) \in A$ such that $div(q) \leq 3$

Define $E := \{(q, e) \in A \mid (P \circ \tilde{f})_s(q, e) = (P \circ \tilde{f})(q, e)\}$. In the following we want to derive that for all $q \in conv(\{q^i, q^j, q^k\})$ and all e such that $(q, e) \in A$ also $(q, e) \in E$. As in BGSS, the idea is to show that for all simplicial points g of any simplicial partition (see Section 2.9.6 for a definition) such that $(g, e) \in A$ as well $(g, e) \in E$

holds. We apply the mechanism as in Bleile (2014a,b) to cover all simplicial points recursively. Based on this, it is possible to cover each $q \in conv(\{q^i, q^j, q^k\})$ by sequences of (appropriate) simplicial points.
In order to illustrate the intuition, we will describe only the first step, i.e. the simplicial points of the first simplicial partition. The further steps are analogously modified versions (to the prototype based setup) of arguments used in Bleile (2014b).

Remark 4.4
For all $e \in \Delta(C)$ such that $q^i \subset e$, the pair (q^i, e) is admissible. However, for q such that $div(q) \in \{2, 3\}$ this does not hold true in general. Nevertheless the sets of frequency vectors $e \in \Delta(C)$ that makes (q, e) admissible for $div(q) = 3$ coincides with the set of frequency vectors that "make" (q', e) admissible for $q' \subset q$ and $div(q') = 2$. This follows directly from the definition of the admissibility condition (ii). This guarantees that there exists a common set of vectors e that make any t-th $(t \geq 1)$ simplicial point admissible, when it is paired with such an e. This is important for our recursive procedure of combining all (differently diverse) simplicial points (as well become clear below).

Consider q such that $2 \leq div(q) \leq 3$
Step A:
Consider any two distinct q^i, q^j $(i, j \leq m)$, w.l.o.g let $i = 1, j = 2$. Define $q_1^1 := \frac{1}{2}q^1 + \frac{1}{2}q^2$ and obviously $div(q_1^1) = 2$. For all e such that $(q, e) \in A$, there exist some q^k (which might be different for different e) such that for $\alpha \in (0, 1)$ $|\tilde{f}^e(\alpha q_1^1 + (1 - \alpha)q^k)| = 3$. q_1^1 is a simplicial point of the first simplicial partition of the triangle spanned by $conv(\{q^1, q^2, q^k\})$.
The prototype based Concatenation Axiom delivers the existence of $\beta, \gamma \in (0, 1)$ such that

$$
\begin{aligned}
(P \circ \tilde{f})(q^*_{\{1,2,k\}}, e) &= \beta(P \circ \tilde{f})(q_1^1, e) + (1 - \beta)(P \circ \tilde{f})(q^k, e) \\
(P \circ \tilde{f})(q_1^1, e) &= \gamma(P \circ \tilde{f})(q^1, e) + (1 - \gamma)(P \circ \tilde{f})(q^2, e)
\end{aligned}
$$

Hence we get (using the notation that (a,b) indicates the line running through a and b) that $(P \circ \tilde{f})(q_1^1, e)$ is contained in

$$((P \circ \tilde{f})(q^1, e), (P \circ \tilde{f})(q^2, e)) \cap ((P \circ \tilde{f})(q^k, e), (P \circ \tilde{f})(q^*_{\{1,2,k\}}, e)). \tag{4.6}$$

The same holds true for $(P \circ \tilde{f})_s$, since it also satisfies the prototype based Concatenation Axiom. Moreover, by Step 0 we know already that for all $(q^j, e) \in A$, also $(q^j, e) \in E$ and that for any triplet $\{i, j, k\}$ and all $(q^*_{\{i,j,k\}}, e) \in A$ also $(q^*_{\{i,j,k\}}, e) \in E$.

Thus, it only remains to check, whether the induced prototype based beliefs of $(q^1,e),(q^2,e),(q^k,e)$ and $q^*_{\{1,2,k\}}$ are not collinear. Then, the two lines involved in (4.6) have an unique intersection. This implies that $(P\circ\tilde{f})(q_1^1,e)$ and $(P\circ\tilde{f})_s(q_1^1,e)$ must coincide, since both lie on both lines, i.e. $(q_1^1,e)\in E$. However, the non-collinearity can be easily seen, since (q^v,e) for $v=1,2,k$ induce three different prototype based beliefs $P^{\tilde{f}_v^e}$ (since $|\tilde{f}^e(\alpha q_1^1+(1-\alpha)q^k)|=3$ for any $\alpha\in(0,1)$) that are not collinear by the Collinearity Axiom.
Analogously, we can analyze all simplicial points of the first simplicial partition of $conv(\{q^1,q^2,q^k\})$, i.e. for $q_1^2=\frac{1}{2}(q^1+q^k)$ and $q_1^3=\frac{1}{2}(q^2+q^k)$. Following the (slightly modified) reasoning/method as in Sections 2.9.6 or 3.9.3 (or Bleile (2014a,b) Step 1), we can show that for the chosen e, all simplicial points of any t-th simplicial partition are in E. Finally, one can find a sequence of simplicial points that converges to any $q\in\Delta(\{q^1,q^2,q^k\})$, where their induced prototype based belief converge. For the details see the proofs of BGSS or the above mentioned sections. Thus, for a given e such that $(q,e)\in E$ (which induces q^k as above) we have for all $q\in conv(\{q^1,q^2,q^k\})$ as well $(q,e)\in A$.

Step B:
Observe that for each admissible pair $(q,e)\in A$ such that $div(q)=2$ the frequency vector e induces a set of frequency vectors $e_{ad}:=\{q^k\mid \exists k\leq m \text{ s. th. } |\tilde{f}^e(\alpha q_1^1+(1-\alpha)q^k|=3\}$.
Hence keeping q^1,q^2 fixed, the same procedure as in Step A can be applied to the triangle $conv(\{q^1,q^2,q^k\})$ for all $k\leq m$ such that $q^k\in e_{ad}$ for this specific e.

Step C:
Now, apply Step A and B to all e such that $(q,e)\in A$ for $div(q)\in\{2,3\}$.

Step D:
Finally, applying the procedure to all possible pairs q^i,q^j, instead of $i=1,j=2$, we get that for all admissible $(q,e)\in A$ such that $div(q)\leq 3$ the claim $(q,e)\in E$, i.e. $A=E$ for all q such $div(q)\leq 3$, which concludes the proof of Step 1.

Now, we consider the situation for $(q,e)\in A$ such that $div(q)>3$. Therefore we need an extended definition of $(P\circ\tilde{f})_s$ for $(q,e)\in\Delta(C)\times\Delta(C)$:

$$(P\circ\tilde{f})_s(q,e):=\frac{\sum_{i\leq m}s_iq_i(\sum_{l\leq L+1}1_{\tilde{f}_l^e}(q^i)P^{\tilde{f}_l^e})}{\sum_{i\leq m}s_iq_i}.$$

The indicator function appears in comparison to definition (4.5) since at this point, it is not anymore clear to which category $\tilde{f}_l^e$ a specific unit vector q^i belongs. Basically, $(P \circ \tilde{f})_s$ is a reformulation of the representation (4.4).

Step 2: Show that $(q,e) \in E$ **for all** $(q,e) \in A$ **such that** $div(q) > 3$

We prove inductively over k that for all q such that $div(q) = k$ with $3 \leq k \leq m$ and all admissible pairs $(q,e) \in A$ it holds, that $(P \circ \tilde{f})_s(q,e) = (P \circ \tilde{f})(q,e)$, i.e. $(q,e) \in E$.
We take the situation $div(q) = k = 3$, which was shown in Step 1, as the basis of the induction and assume that the claim is true for all $(q,e) \in A$ such that $div(q) = k-1$ for $k > 3$.
Take any $(q,e) \in A$ such that $div(q) = k$, w.l.o.g. $q \in int(conv(\{q^1,..,q^k\}))$. By admissibility $|\tilde{f}^e(q)| \geq 3$ holds, i.e. there are at least two categories $\tilde{f}_l^e$ for some $l \leq L+1$ containing at least two different cases q^j for some $j \leq m$ and another category containing at least one different case. W.l.o.g. let these categories be given in the following way (for distinct $i,j,k \neq 1,2$)

$$\begin{aligned} \tilde{f}_1^e(q) &\supseteq \{q^1, q^i\} \\ \tilde{f}_2^e(q) &\supseteq \{q^2, q^j\} \\ \tilde{f}_{L+1}^e(q) &\supseteq \{q^k\} \\ \tilde{f}_l^e(q) &\supseteq \emptyset \text{ for all } l \neq 1,2 \leq L \end{aligned}$$

Now, let q be decomposed by $q = \alpha_t q^t + (1-\alpha_t)q(t)$, where for $t = 1,2$ $q(t)$ is the point in $conv(\{q^j | j \leq k, j \neq t\})$ that is on the line connecting q^t and q and $\alpha_t \in (0,1)$ accordingly. By the prototype based Concatenation Axiom we know that there exist $\lambda_t \in (0,1)$ such that

$$(P \circ \tilde{f})(q,e) = \lambda_t (P \circ \tilde{f})(q^t,e) + (1-\lambda_t)(P \circ \tilde{f})(q(t),e)$$

This is possible since $(q^t,e) \in A$ and $(q(t),e) \in A$, since $\tilde{f}^e(q(t)) = \tilde{f}^e(q)$ by construction. Since in addition $(q^t,e),(q(t),e) \in E$ by the induction assumption (since $div(q(t)) = k-1$) and $(P \circ \tilde{f})_s$ satisfies the prototype based Concatenation Axiom, we have

$$(P \circ \tilde{f})(q,e),\ (P \circ \tilde{f})_s(q,e) \in \bigcap_{t=1,2} ((P \circ \tilde{f})(q^t,e),\ (P \circ \tilde{f})(q(t),e))$$

This intersection is unique since for $t = 1, 2$ it can be shown that the following holds:

$$(P \circ \tilde{f})(q^2, e) = P^{\tilde{f}_2^e} \notin ((P^{\tilde{f}_1^e}, (P \circ \tilde{f})(q(t), e)) =: h \tag{4.7}$$

If this would not be true, i.e. if $P^{\tilde{f}_2^e}$ would be on this line h, it would require that $(P \circ \tilde{f})(q(t), e)$ is on the line between $P^{\tilde{f}_1^e}$ and $P^{\tilde{f}_2^e}$. However, by construction $(P \circ \tilde{f})(q(t), e) \in int(conv(\{P^{\tilde{f}_1^e}, P^{\tilde{f}_2^e}, P^{\tilde{f}_k^e}, ...\}))$, which implies that it cannot lie on $(P^{\tilde{f}_1^e}, P^{\tilde{f}_2^e})$, since by the Collinearity Axiom no three of $(P^{\tilde{f}_l^e})_l$ are collinear. Thus in sum, claim (4.7) is true, which implies that for $t = 1, 2$ the lines based on $((P \circ \tilde{f})(q^t, e), (P \circ \tilde{f})(q(t), e))$ are distinct and intersect uniquely in $(P \circ \tilde{f})(q, e) = (P \circ \tilde{f})_s(q, e)$, i.e. $(q, e) \in E$.
This completes the entire proof. $\square$

Bibliography

[1] Ahn, D. (2008). Ambiguity Without a State Space. *Review of Economic Studies, 71, 3-28.*

[2] Al-Najjar, N. I. and Pai, M. (2014). Coarse decision making and overfitting. *Journal of Economic Theory, 150, 467-486.*

[3] Allport, G.W. (1954). The Nature of Prejudice. *Addison Wesley, Cambridge.*

[4] Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98-3, 409-429.*

[5] Anderson. J.R. (2005). Cognitive Psychology and Its Implications. *Worth, New York.*

[6] Andrews, R.L., Srinivasan, T.C. (1995). Studying Consideration Effects in Empirical Choice Models Using Scanner Panel Data. *Journal of Marketing Research, 32-1, 30-41.*

[7] Apesteguia, J., Ballester, M. (2013). Choice by Sequential Procedures. *Games and Economic Behavior, 77, 90-99.*

[8] Arad, A., Gayer, G. (2012). Imprecise Data Sets as a Source of Ambiguity: A Model and Experimental Evidence. *Management Science, 58-1, 188-202.*

[9] Bargh, J.A. (1994). The Four Horsemen of Automaticity: Awareness, Intention, Efficiency, and Control in Social Cognition. *Handbook of Social Cognition, 1, 1-40.*

[10] Bargh, J.A. (1997). The Automaticity of Everyday Life. *The Automaticity of Everyday Life: Advances in Social Cognition, 10, 1-61.*

[11] Bargh, J.A. (1999). The Cognitive Monster: The Case Against the Controllability of Automatic Stereotype Effects. *Dual Process Theories in Social Psychology, 361-382.*

[12] Bettman, J.R. (1979). Issues in Research on Consumer Choice. *Advances in Consumer Research, 6, 214-217.*

[13] Billot, A., Gilboa, I., Schmeidler, D. (2004). An Axiomatization of an Exponential Similarity Function. *Mimeo, University of Tel Aviv.*

[14] Billot, A. Gilboa, I., Samet, D., Schmeidler, D. (2005). Probabilities as Similarity-Weighted Frequencies. *Econometrica, 73, 1125-1136.*

[15] Bleile, J. (2014a). Cautious Belief Formation. *Working Paper No. 507, Center for Mathematical Eonomics, Bielefeld.*

[16] Bleile, J. (2014b). Limited attention in Case-based Belief Formation. *Working Paper, Center for Mathematical Eonomics, Bielefeld.*

[17] Bleile, J. (2014c). Categorization based Belief Formation. *Working Paper, Center for Mathematical Eonomics, Bielefeld.*

[18] Broadbent, D. E. (1958). Perception and Communication. *Pergamon Press, New York.*

[19] Caplin, A., Dean, M., Martin, D. (2011). Search and Satisficing. *American Economic Review, 101-7, 2899-2922.*

[20] Cherepanov, V., Feddersen, T., Sandroni, A. (2013). Rationalization. *Theoretical Economics, 8, 775-800.*

[21] Chiang, J., Chib, S., Narasimhan, C. (1998). Markov chain monte carlo and models of consideration set and parameter heterogeneity. *Journal of Econometrics, 89-1-2, 223-248.*

[22] de Clippel, G., Eliaz, K. (2012). Reason-based choice: A bargaining rationale for the attraction and compromise effects. *Theoretical Economics, 7, 125-162.*

[23] Dekel, E., Lipman, B. (2012). Costly Self-Control and Random Self-Indulgence. *Econometrica, 80-3, 1271-1302.*

[24] Dulleck, U., Hackl, F., Weiss, B. Winter-Ebmer, R. (2011). Buying online: An Analysis of Shopbot Visitors. *German Economic Review, 12-4, 395-408.*

[25] Eichberger J., Guerdjikova, A. (2010). Case-Based Belief Formation under Ambiguity. *Mathematical and Social Sciences, 60-3, 161-177.*

[26] Eliaz, K., Spiegler, R. (2011a). Consideration sets and competitive marketing. *Review of Economic Studies, 78, 235-262.*

[27] Eliaz, K., Spiegler, R. (2011b). On the Strategic Use of Attention Grabbers. *Theoretical Economics, 6, 127-155.*

[28] Ellsberg, D. (1961). Risk, ambiguity and the Savage axioms. *Quartely Journal of Economics, 75, 643-669.*

[29] Epstein, L.G., Schneider, M. (2007). Learning Under Ambiguity. *Review of Economic Studies, 74, 1275-1303.*

[30] Fryer, R., Jackson, M. O. (2008). A categorical model of cognition and biased decision making. *The B.E. Journal of Theoretical Economics (Contributions), 8-1, 1-42.*

[31] Gärdenfors, P. (2000). Conceptual Spaces: The Geometry of Thought. *MIT Press, Cambridge.*

[32] Gelman, S. A., Markman, E. M. (1986). Categories and induction in young children. *Cognition, 23, 183-209.*

[33] Gensch, D. H. (1987). A Two Stage Disaggregate Attribute Choice Model. *Marketing Science, 6, 223-231.*

[34] Gerasimou, G. (2013). A Behavioural Model of Choice in the Presence of Decision Conflict. *School of Economics and Finance Discussion Paper No. 1302, University of St Andrews.*

[35] Gigerenzer, G., Todd, P., ABC Research Group (1999). Simple Heuristics That Make Us Smart. *Oxford University Press, New York.*

[36] Gilboa I., Schmeidler, D. (1995). Case-Based Decision Theory *Quarterly Journal of Economics, 110, 605-639.*

[37] Gilboa, I., Schmeidler, D. (2001). A Theory of Case-Based Decisions.*Cambridge University Press, Cambridge, UK.*

[38] Gilboa, I., Schmeidler, D. (2003). Inductive Inference: An Axiomatic Approach. *Econometrica, 71, 1-26.*

[39] Gilboa, I., Lieberman, O., Schmeidler, D. (2006). Empirical Similarity. *Review of Economics and Statistics, 88, 433-444.*

[40] Gilboa, I. (2009). Theory of Decision under Uncertainty. *Econometric Society Monographs. Cambridge University Press, Cambridge.*

[41] Gilboa, I., Schmeidler, D. (2010). Case based Predictions: Introduction.

[42] Gilboa, I., Lieberman, O., Schmeidler, D. (2011). A similarity-based approach to prediction. *Journal of Econometrics, 162-1, 124-131.*

[43] Gilboa I., Postlewaite, A., Schmeidler, D. (2012). Rationality of belief or: why savages axioms are neither necessary nor sufficient for rationality. *Synthese, 187-1, 11-31.*

[44] Goldstone, R. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition, 52-2, 125-157.*

[45] Goldstone, R.L., Son, J.Y. (2005). Similarity. *The Cambridge handbook of thinking and reasoning. Cambridge University Press, 13-36.*

[46] Gul, F. Pesendorfer, W. (2001). Temptation and Self-Control. *Econometrica, 69, 1403-1435.*

[47] Gul, F., Natenzon, P., Pesendorfer, W. (2012). Random Choice as Behavioral Optimization. *Working Paper.*

[48] Hau, R., Pleskac, T.J., Hertwig, R. (2010). Decisions From Experience and Statistical Probabilities: Why They Trigger Different Choices Than a Priori Probabilities. *Journal of Behavioral Decision Making, 23, 48-68.*

[49] Hauser, J. R., Wernerfelt, B. (1990). An evaluation cost model of evoked sets. *Journal of Consumer Research, 16, 383-408.*

[50] Hauser, J. R. (2013). Consideration Set Heuristics. *Forthcoming, Journal of Business Research.*

[51] Hayne, H. (1996). Categorization in Infancy. *Advances in Infancy Research, 10.*

[52] Heit, E. . Features of Similarity and Category-Based Induction. *Proceedings of the Interdisciplinary Workshop on Categorization and Similarity, University of Edinburgh, 115-121.*

[53] Heit, E., Rubinstein, J. (1994). Similarity and Property Effects in Inductive Reasoning. *Journal of Experimental Psychology: Learning, Memory and Cognition, 20-2, 411-422.*

[54] Horan, S. (2013). A Simple Model of Biased Choice. *Working paper, forthcoming: Journal of Economic Theory.*

[55] Houy, N. (2007). Rationality and order-dependent sequential rationality. *Theory and Decision, 62, 119-134.*

[56] Houy, N., Tadenuma, K. (2009). Lexicographic compositions of multiple criteria for decision making. *Journal of Economic Theory, 144, 1770-1782.*

[57] Houy, N. (2010). A characterization of prudent choices. *Social Choice and Welfare, 34-2, 181-192.*

[58] Jarvis, L., Wilcox, J. (1977). A model for predictive measurements of advertising effectiveness. *Industrial Marketing Management, 6, 9-14.*

[59] Jones, G. Y. (1983). Identifying basic categories. *Psychological Bulletin, 94, 423-428.*

[60] Krueger, J., Clement, R. (1994). Memory-based judgments about multiple categories. *Journal of Personality and Social Psychology, 67, 35-47.*

[61] Kruschke, J.K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99, 22-44.*

[62] Lleras, J.S., Masatlioglu, Y., Nakajima, D., Ozbay, E.Y. (2010) and (2012). When more is less: Limited consideration. *Unpublished paper.*

[63] Lombardi, M. (2009). Reason-Based Choice Correspondences. *Mathematical Social Sciences, 57, 58-66.*

[64] Malt, B. C., Ross, B. H., Murphy, G. L. (1995). Predicting features for members of natural categories when categorization is uncertain. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21, 646-661.*

[65] Manzini, P., Mariotti, M. (2007). Sequentially Rationalizable Choice. *American Economic Review, 97, 1824-1839.*

[66] Manzini, P., Mariotti, M. (2012a). Choice by lexicographic semiorders. *Theoretical Economics, 7, 1-23.*

[67] Manzini, P., Mariotti, M. (2012b). Categorize Then Choose: Boundedly Rational Choice and Welfare. *Journal of the European Economic Association, 10-5, 1141-1165.*

[68] Manzini, P., Tyson, C. J., Mariotti, M. (2013a). Two-stage threshold representations *Theoretical Economics, 8-3.*

[69] Manzini, P., Mariotti, M. (2013b). Moody Choice. *Working Paper.*

[70] Marinacci, M. (2002). Learning about Ambiguous Urns. *Statistical Papers, 43, 143-151.*

[71] Masatlioglu, Y., Ok, E. A. (2005). Rational Choice with Status Quo Bias. *Journal of Economic Theory, 121, 1-29.*

[72] Masatlioglu, Y., Nakajima, D., Ozbay, E.Y. (2012). Revealed attention. *American Economic Review, 102, 2183-2205.*

[73] Medin, D. L., Schaffer, M.M. (1978). Context theory of classification learning. *Psychological Review, 85, 207-238.*

[74] Medin, D. L. (1983). Structural principles of categorization. *Interaction: Perception, Development and Cognition, 203-230.*

[75] Medin, D. L., Aguilar, C. (1999). Categorization. *MIT Encyclopedia of the Cognitive Sciences.*

[76] Miller, G. A. (1956). The magic number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review, 63, 81-97.*

[77] Mohlin, E. (2014). Optimal Categorization. *Journal of Economic Theory, 152, 356-381.*

[78] Mullainathan, S. (2002). Thinking Through Categories. *MIT Working Paper.*

[79] Murphy, G. L., Medin, D.L. (1985). The Role of Theories in Conceptual Coherence. *Psychological Review, 92, 289-316.*

[80] Murphy, G. L., Ross, B. H. (1994). Predictions from uncertain categorizations. *Cognitive Psychology 27, 148-193.*

[81] Murphy, G. L. (2002). The Big Book of Concepts. *MIT Press, Cambridge.*

[82] Nedungadi, P. (1990). Recall and Consumer Consideration Sets: Influencing Choice Without Altering Brand Evaluations. *Journal of Consumer Research, 17-3, 263-276.*

[83] Osherson, D. N., Wilkie, O., Smith, E. E., Lopez, A., Shafir, E. (1990). Category-Based Induction. *Psychological Review, 97, 185-200.*

[84] Papi, M. (2012). Satisficing choice procedures. *Journal of Economic Behavior and Organization, 84-1, 451-462.*

[85] Peski, M. (2011). Prior symmetry, similarity-based reasoning, and endogenous categorization. *Journal of Economic Theory 146, 111-140.*

[86] Pessemier, E. A. (1978). Stochastic properties of changing preferences. *American Economic Review, 68-2, 380-385.*

[87] Posner, M. I., Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77, 353-363.*

[88] Pothos, E. M., Chater, N. (2002). A simplicity categorization. *Cognitive Science, 26, 303-343.*

[89] Quinn, P.C., Eimas, P.D. (1996). Perceptual Organization and Categorization. *Young Infants: Advances in Infancy Research, 10, 1-36.*

[90] Reed, S. K. (1972). Pattern Recognition and Categorization. *Cognitive Psychology, 3, 382-407.*

[91] Reutskaja, E., Nagel, R., Camerer, C. F., Rangel, A. (2011). Search Dynamics in Consumer Choice Under Time Pressure: An Eye-Tracking Study. *American Economic Review, 101-2, 900-926.*

[92] Rips, L. J. (1975). Inductive Judgements About Natural Categories. *Journal of Verbal learning and Verbal Behavior, 14, 665-681.*

[93] Rips, L. J. (1989). Similarity, typicality, and categorization. *Similarity and analogical reasoning, 21-59.*

[94] Roberts, J. H., Lattin, J. (1991). Development and Testing of a Model of Consideration Set Composition *Journal of Marketing Research, 28, 429-440.*

[95] Roberts, J. H., Lattin, J. (1997). Consideration: Review of research and prospects for future insights. *Journal of Marketing Research, 34, 406-410.*

[96] Rosch, E., Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7, 573-603.*

[97] Rosch, E., Lloyd, B. (1976). Cognition and categorization. *Lawrence Erlbaum Associates, Hillsdale.*

[98] Rosch, E. (1978). Principles of Categorization. *Cognition and Categorization, 27-48.*

[99] Rubinstein, A., Salant, Y. (2006). A model of choice from lists. *Theoretical Economics, 1, 3-17.*

[100] Salant, Y., Rubinstein, A. (2008). (A,f): Choice with Frames *Review of Economic Studies, 75, 1287-1296.*

[101] Savage, L.J. (1954). The foundations of statistics. *New York, John Wiley and Sons.*

[102] Schank, R. C. (1986). Explanation Patterns: Understanding Mechanically and Creatively. *Hillsdale, NJ: Lawrence Erlbaum Associates.*

[103] Schwartz, B. (2004). The Paradox of Choice: Why More is Less. *Harper Perennial, New York.*

[104] Shafir, E., Simonson, I., Tversky, A. (1993). Reason-based choice. *Cognition, 49, 11-36.*

[105] Shugan, S. M. (1980). The cost of thinking. *Journal of Consumer Research, 27-2, 99-111.*

[106] Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics, 69-1, 99-118.*

[107] Simon, H. A. (1957). Models of Man: Social and Rational: Mathematical Essays on Rational Human Behavior in a social setting. *New York: John Wiley and Sons, Inc.*

[108] Simon, H. A. (1959). Theories of Decision-Making in Economics and Behavioral Science. *American Economic Review, 49-3. 253-283.*

[109] Smith, L. B. (1989). From global similarity to kinds of similarity: The construction of dimensions in development. *Similarity and analogical reasoning, 146-178.*

[110] Smith, E.E., Patalano, A.L., Jonides, J. (1998). Alternative strategies of categorization. *Cognition, 65, 167-196.*

[111] Stigler, G.J. (1961). The economics of information. *Journal of Political Economy, 69-3, 213-225.*

[112] Tversky, A. (1972). Elimination by Aspects: A Theory of Choice. *Psychological Review 79, 281-299.*

[113] Tversky, A. (1977). Features of Similarity. *Psychological Review, 84-4, 327-352.*

[114] Tversky, A., Sattath, S., Slovic, P. (1988). Contingent weighting in judgment and choice. *Psychological Review, 95, 371-384.*

[115] Tyson, C.J. (2008). Cognitive constraints, contraction consistency, and the satisficing criterion. *Journal of Economic Theory, 127-1, 51-70.*

[116] Tyson, C. J. (2013). Satisficing Behavior with a Secondary Criterion. *Working Paper.*

[117] Vanpaemel, W., Storms, G. (2008). In search of abstraction: The varying abstraction model of categorization. *Psychonomic Bulletin and Review, 15-4, 732-749.*

[118] Wright, P., Barbour, F. (1977). Phased Decision Strategies: Sequels to an Initial Screening. *In Multiple Criteria decision Making, 91-109. Amsterdam: North-Holland Publishing Co.*

Kurzer Lebenslauf

Jörg Bleile

Geboren am 26.12.1980 in Müllheim.

Abitur 2000 am Markgräfler Gymnasium in Müllheim.

Zivildienst in Bad Krozingen von 10/2000 bis 08/2001.

Diplom in Mathematik an der Albert-Ludwigs-Universität Freiburg von 10/2001 bis 11/2007 mit Auslandsaufenthalt an der University of British Columbia, Vancouver.

Diplom in Volkswirtschaftslehre an der Albert-Ludwigs-Universität Freiburg von 10/2006 bis 04/2009.

Wissenschaftlicher Mitarbeiter (Forschung und Lehre), Lehrstuhl für Wirtschafts-theorie Albert-Ludwigs-Universität Freiburg von 07/2009 bis 08/2010.

Wissenschaftlicher Mitarbeiter (Lehre) in der Justizvollzugsanstalt Freiburg im Rahmen der Fernuniversität Hagen (und KIT) von 10/2009 bis 07/2010.

Doktorand (Dr. rer. pol.), International Research Training Group: Economic Behavior and Interaction Models (EBIM), Universität Bielefeld und Paris 1 Pantheon-Sorbonne seit 10/2010.