Advances in Choquet Theories

by

Michele Caprio

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Sayan Mukherjee, Supervisor

_____
Peter Hoff

_____
James Berger

_____
Nicholas Cook

Dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in the Department of Statistical Science
in the Graduate School of Duke University
2022

Abstract

Advances in Choquet Theories

by

Michele Caprio

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Sayan Mukherjee, Supervisor

_____
Peter Hoff

_____
James Berger

_____
Nicholas Cook

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Statistical Science in the Graduate School of Duke University
2022

Copyright © 2022 by Michele Caprio
All rights reserved

Abstract

Choquet theory and the theory of capacities, both initiated by French mathematician Gustave Choquet, share the heuristic notion of studying the extrema of a convex set in order to give interesting results regarding its elements. In this work, we put to use Choquet theory in the study of finite mixture models and the theory of capacities in studying severe uncertainty.

In chapter 2, we show how by combining a classical non-parametric density estimator based on a Dirichlet process with techniques from Choquet theory, it is possible to retrieve the weights of a finite mixture model. We also give the rate of convergence of the Dirichlet process posterior to the Dirac measure on the weights.

In chapter 3, we introduce dynamic probability kinematics (DPK), a method for an agent to mechanically update subjective beliefs in the presence of partial information. We then generalize DPK to dynamic imprecise probability kinematics (DIPK), which allows the agent to express their initial beliefs via a set of probabilities to take ambiguity into account. We provide bounds for the lower probability associated with the updated probability sets, and we study the behavior of the latter, in particular contraction, dilation, and sure loss. Examples are provided to illustrate how the methods work. We also formulate in chapter 4 an ergodic theory for the limit of the sequence of successive DIPK updates. As a consequence, we formulate a strong law of large numbers.

Finally, in chapter 5 we propose a new, more general definition of extended prob-

ability measures ("probabilities" whose codomain is the interval $[-1, 1]$). We study their properties and provide a behavioral interpretation. We use them in an inference procedure, whose environment is canonically represented by a probability space, when both the probability measure and the composition of the state space are unknown. We develop an ex ante analysis – taking place before the statistical analysis requiring knowledge of the state space – in which we progressively learn its true composition. We describe how to update extended probabilities in this setting, and introduce the concept of lower extended probabilities. We provide two examples in the fields of ecology and opinion dynamics.

Contents

List of Figures

List of Abbreviations and Symbols

Symbols

$\mathbb{N}$	The set of natural numbers.
$\mathbb{N}_0 \equiv \mathbb{Z}_+$	The set of whole numbers.
$\mathbb{Z}$	The set of integer numbers.
$\mathbb{Q}$	The set of rational numbers.
$\mathbb{R}$	The set of real numbers.
$\sqcup$	A symbol that denotes the disjoint union between sets.
$\mathbb{I}$	The indicator function.
d_{TV}	The total variation metric.
d_H	The Hausdorff metric.
$\xrightarrow{w}$	Weak convergence.
$\xrightarrow{a.s.}$	Almost sure convergence.
$(\Omega, \mathcal{F})$	A generic measurable space constituted by set Ω and sigma-algebra $\mathcal{F}$ of subsets of Ω.
$\#A$	The cardinality of a generic set A.
A^c	The complement of a generic set A.
2^A	The power set of a generic set A.
$\mathrm{Conv}(A)$	The convex hull of a generic set A.
$\mathrm{Cl}(A)$	The closure of a generic set A.
$\mathrm{ex}(A)$	The extreme points of a generic convex set A.

δ_A	The Dirac measure at a generic set A.
$\dim(A)$	The dimension of space A.
Δ^{J-1}	The simplex in the Euclidean space $\mathbb{R}^J$.
$\mathcal{F}_i(K)$	One of the i-faces of a generic polytope K.
$\mathscr{F}_i(K)$	The collection $\{\mathcal{F}_i(K)\}$ of the i-faces of a generic polytope K.
$F_i(K)$	The cardinality of $\mathscr{F}_i(K)$.
$DP(\alpha P_0)$	A generic Dirichlet process with parameter $\alpha > 0$ and base measure P_0.
$\Delta(\Omega, \mathcal{F})$	The space of probability measures defined on measurable space $(\Omega, \mathcal{F})$.
Π	A generic set of probabilities.
$\underline{P}$	A generic lower probability.
$\mathbf{x}_t$	Data points available up to time t when performing a DPK or DIPK updating procedure.
$\text{unique}(\mathbf{x}_t)$	The unique elements in collection $\mathbf{x}_t$.
$\mathcal{P}^{\text{co}}_{\mathcal{E}_t}$	The t-th DIPK-update of set $\mathcal{P}^{\text{co}}_{\mathcal{E}_0}$ representing the agent's initial beliefs, $t \in \mathbb{N}$.
$\text{core}(\underline{P})$	The core of a generic lower probability $\underline{P}$.
$\mathcal{L}(\Omega)$	Set of all gambles on Ω.
$B(\Omega, \mathcal{F})$	The space of bounded and $\mathcal{F}$-measurable functionals on Ω.
$\Delta^{ex}(\Omega, \mathcal{F})$	The space of extended probability measures defined on measurable space $(\Omega, \mathcal{F})$.
$\Delta^{ex}_{Baire}(\Omega, \mathcal{F})$	The space of Baire extended probability measures on $(\Omega, \mathcal{F})$.
P^{ex}	A generic extended probability measure.
$\mathcal{P}^{ex}$	A generic set of extended probability measures.
$\underline{P}^{ex}$	A generic lower extended probability measure.

Abbreviations

DIPK	Dynamic impreciseprobability kinematics.
DMD	Dynamic mode decomposition.
DP	Dirichlet process.
DPK	Dynamic probability kinematics.
EM	Expectation-maximization (algorithm).
EMPF	Ergodic measure preserving flow.
LDA	Latent Dirichlet allocation.
LP	Lower probability.
MLE	Maximum likelihood estimator.
MCMC	Markov chain Monte Carlo.
PCA	Principal component analysis.
PK	Probability kinematics.
RCM	Richest cheap model.
SFA	Sparse factor analysis.
VI	Variational inference.

Acknowledgements

I would like to thank Sayan Mukherjee for being my advisor during the four years I spent at Duke. Working with him has been an honor. As an advisor, Sayan allowed me to research on the topics I found most interesting, even the most eccentric and peculiar ones (e.g. non-Diophantine arithmetics or tropical algebra). He always encouraged me and managed to find a way of guiding me and of providing profound insights. He also believed in me when even I was doubting myself. I will be forever grateful.

I would then like to thank my coauthors and friends Andrea Aveni, Jordan Bryan, Mark Burgin, Simone Cerreia-Vioglio, Shounak Chattopadhyay, Roberto Corrao, Pierpaolo De Blasi, Federico Ferrari, Ruobin Gong, P. Richard Hahn, Jürgen Jost, Jeremias Knoblauch, Paolo Leonetti, Lek-Heng Lim, Rostislav Matveev, Xiao-Li Meng, Pietro Muliere, Vittorio Orlandi, Teddy Seidenfeld, Ngoc Mai Tran, and Alessandro Zito for fruitful and interesting discussions on all sorts of topics, from foundations of probability to ideals.

I would also like to thank my defense committee members, Jim Berger, Peter Hoff and Nick Cook for their support, helpful comments and guidance. My deepest gratitude goes to the DSS staff, and in particular to Lori and Nikki, who were always very helpful and kind to me.

I would like to acknowledge the support received from the grants CCF-1934964 of the National Science Foundation of the United States of America and 1R01MH118927

-01 of the National Institute of Mental Health of the United States of America.

They say a PhD is like a marathon, and that you need the right support along the way. To this extent, I would like to thank those who shared these years in Durham with me, and that made my experience an unforgettable one. In particular, I would like to mention Alessandro, Andrea, Andy, Becky, Ed, Evan, Federico, Heather, Jordan, Shounak and Vittorio. They have been a family here at Duke, and a pretty smart one at that!

Last but not least, I want to thank Roberto, my parents and my fiancée Erica. They have been close, loving and supportive even in the most challenging times. My gratitude for them is endless.

1

Introduction

The main goal of this dissertation is to study subjective probability using techniques from Choquet theory and the theory of capacities. In particular, we show how a Dirichlet process on the distributions supported on the elements of a finite mixture model retrieves the Choquet measure, which in turn gives us the weights of the model. We also use lower probabilities, a particular type of Choquet capacities, to model belief revision under severe uncertainty.

Choquet theory, named after French mathematician Gustave Choquet, is an area of functional and convex analyses concerned with measures which have support on the extreme points of a convex set (Phelps, 2001). Its fundamental tenet is that we can represent every element in a convex set C via a weighted average of the extrema of the set. Here weighted average is to be understood as a generalization of the usual notion of convex combination to an integral taken over the set E of extreme points of C. The formal, central result to Choquet theory is the following.

Theorem 1. (Choquet, cf. (Phelps, 2001)) Let C be a metrizable compact convex subset of a locally convex space V. Pick any $c \in C$. Then, there exists a

probability measure μ on C which represents c and is supported on E, that is,

$$f(c) = \int_E f(e)\mu(\mathrm{d}e),$$

for any affine function f on C.

Choquet also characterized those compact convext sets C with the property that for every $c \in C$ there is a unique probability measure μ_c supported on E that represents c. The necessary and sufficient condition is based on the concept of Choquet simplex.

Definition 2. A nonempty convex set C (not necessarily compact) of a locally convex space V is a Choquet simplex if it has the following property. Under the embedding of V as the hyperplane $V \times \{1\}$ in the space $V \times \mathbb{R}$, the projecting cone

$$\tilde{C} := \{\alpha c \in V \times \mathbb{R} : c \in C \subset V \times \{1\},\, \alpha \geqslant 0\}$$

of C transforms the space $V \times \mathbb{R}$ into a partially ordered space P such that the space of differences $\tilde{C} - \tilde{C}$ generated by P is a vector lattice in the order induced by C. That is, each pair $c_1, c_2 \in \tilde{C} - \tilde{C}$ has at least upper bound $c_1 \vee c_2 \in \tilde{C} - \tilde{C}$.

In the case when V is finite-dimensional, a Choquet simplex is an ordinary simplex with number of vertices equal to $\dim(V) + 1$, where $\dim(V)$ denotes the dimension of space V. The characterization of C, then, is the following.

Theorem 3. (Choquet, cf. (Phelps, 2001)) Let C be a metrizable closed convex subset of a locally convex space V. Then, C is a Choquet simplex if and only if for every c in C, there exists a unique measure μ_c which represents c and is supported on E, that is,

$$f(c) = \int_E f(e)\mu_c(\mathrm{d}e),$$

for any affine function f on C.

We call μ_c the Choquet measure for c. These results entail that studying the extrema of a convex set gives us important results concerning the (elements of the) whole set. In chapter 2, we bridge the study of finite mixture models with stochastic geometry and Choquet theory. The most important result is Theorem 19. We show that if the geometric representation of a finite mixture model is a simplex, and if we place a Dirichlet process prior on the densities supported on the extrema of such simplex, then the Dirichlet process posterior converges weakly to the Dirac at the Choquet measure δ_{μ_c}. The Choquet measure retrieves the mixture weights of our model; this result is based on Theorem 3. We also give the rate of convergence of the Dirichlet process posterior to δ_{μ_c}.

Another important contribution that Choquet made to modern mathematics is the theory of capacities, which shares with Choquet theory the heuristic notion that studying the extrema of a convex set can give interesting results regarding the elements of the set. We refer to the proper Choquet theory as geometric Choquet theory, while we refer to the theory of capacities as probabilistic Choquet theory. This is because we are especially interested in the application of the theory of capacities to the foundations of subjective probability. The two Choquet theories are related by Theorem 6.

As pointed out in (Choquet, 1954), he developed the theory of capacities while reasoning on the following problem regarding Newtonian capacities:

Is the interior Newtonian capacity of an arbitrary Borelian subset X of the space $\mathbb{R}^3$ equal to the exterior Newtonian capacity of X? [1]

To solve this problem, Choquet first studied nonadditive set functions, identifying the most interesting classes among all those possible, with the goal of establishing a theory analogous to the classical theory of measurability. He discovered that New-

[1] Euristically, the Newtonian capacity is the measure of the size of a set; it is the mathematical analogue of a set's ability to hold electrical charge.

tonian capacities can be seen as the analogue of functions of a real variable whose successive derivatives are alternately positive and negative. He also discovered that there are several classes of functions that justify the interest in determining the extremal elements of convex cones of functions, and in utilizing their integral representations.

The main element of the theory of capacities is the Choquet capacity.

Definition 4. (Choquet, 1954) Given a measurable space $(\Omega, \mathcal{F})$, where $\Omega \neq \varnothing$ and $\mathcal{F}$ is a sigma-algebra of subsets of Ω, we say that a set function $\nu : \mathcal{F} \to [0, 1]$ is a Choquet capacity if $\nu(\varnothing) = 0$, $\nu(\Omega) = 1$, and $\nu(A) \leqslant \nu(B)$ for all $A, B \in \mathcal{F}$ such that $A \subset B$.

In the subjective probability literature, an agent's initial beliefs about an event $A \subset \mathcal{F}$ are usually encapsulated in a single probability measure, that is then refined once new information in the form of data become available. Instead, in the imprecise probabilistic methods literature, and in particular in Bayesian sensitivity analysis (Berger, 1984), a common assumption is that at the beginning of the analysis, the analyst specifies a set $\mathcal{P}$ of probability measures called credal set. But why should they do so? In (Walley, 1991, Section 1.1.4), the author gives an exhaustive answer. The most important reason is that the available information may be scarce, vague, or conflicting, in which case a unique probability distribution may be hard to identify. We call this situation ambiguity, and it corresponds to $\mathcal{P}$ not being a singleton. The "boundary elements" of $\mathcal{P}$ (i.e. its infimum and supremum) are given by a particular type of Choquet capacities.

Definition 5. (Cerreia-Vioglio et al., 2015) Given a measurable space $(\Omega, \mathcal{F})$ as in Definition 4, call $\Delta(\Omega, \mathcal{F})$ the set of probability measures on $(\Omega, \mathcal{F})$. Then, we say that a Choquet capacity $\nu : \mathcal{F} \to [0, 1]$ is

(i) convex if $\nu(a \cup B) + \nu(a \cap B) \geqslant \nu(A) + \nu(B)$, for all $A, B \in \mathcal{F}$;

(ii) additive if $\nu(a \cup B) = \nu(A) + \nu(B)$, for all disjoint $A, B \in \mathcal{F}$;

(iii) continuous if $\lim_{n\to\infty} \nu(A_n) = \nu(A)$ whenever either $A_n \uparrow A$ or $A_n \downarrow A$;

(iv) continuous at Ω if $\lim_{n\to\infty} \nu(A_n) = \nu(\Omega)$ whenever $A_n \uparrow \Omega$;

(v) a probability measure if it is an additive Choquet capacity which is continuous at Ω;

(vi) a lower probability measure if there exists a set $\mathcal{P} \subset \Delta(\Omega, \mathcal{F})$ such that

$$\nu(A) = \inf_{P\in\mathcal{P}} P(A), \quad \forall A \in \mathcal{F}.$$

Given a Choquet capacity ν, its conjugate $\overline{\nu} : \mathcal{F} \to [0, 1]$ is given by

$$\overline{\nu}(A) := 1 - \nu(A^c), \quad \forall A \in \mathcal{F}.$$

So if ν is a lower probability, then

$$\overline{\nu}(A) = \sup_{P\in\mathcal{P}} P(A), \quad \forall A \in \mathcal{F}.$$

We call this latter an upper probability.

Upper and lower probabilities are the key elements of credal sets theory. As we have seen, they represent the "extreme elements" of a set $\mathcal{P}$ of probability measures. Consider now a generic lower probability ν, and call $\text{core}(\nu)$ the set of probability measures that setwise dominate ν, that is,

$$\text{core}(\nu) := \{P \in \Delta(\Omega, \mathcal{F}) : P(A) \geqslant \nu(A), \forall A \in \mathcal{F}\}.$$

It is convex by (Marinacci and Montrucchio, 2004a, section 2.2). Let us denote by $\text{ex}(\text{core}(\nu))$ the set of all extreme points of $\text{core}(\nu)$, that is, the elements of the core that cannot be written as convex combinations of other elements. Then, geometric and probabilistic Choquet theories are related via the following result.

Theorem 6. (Walley, 1991) Suppose $\text{core}(\nu) \neq \varnothing$. Then, the following holds.

(a) $\text{ex}(\text{core}(\nu)) \neq \varnothing$.

(b) $\text{core}(\nu)$ is the closure in the weak* topology of the convex hull of $\text{ex}(\text{core}(\nu))$.[2]

(c) If $\nu(A) = \inf_{P \in \text{core}(\nu)} P(A)$, for all $A \in \mathcal{F}$, then $\nu(A) = \inf_{P \in \text{ex}(\text{core}(\nu))} P(A)$, for all $A \in \mathcal{F}$.

So in order to define a lower probability ν that is setwise dominated by the elements of $\text{core}(\nu)$ it is enough to specify the extreme points of the core. This theorem is a corollary of the Krein-Milman theorem, which in turn is a corollary of Theorem 1. Theorem 6 makes clear how the geometric concept of extreme points is related to the (imprecise) probabilistic concept of lower probability. The importance of the core is discussed in chapters 3, 4 and 5.

In chapter 3 we develop a new procedure to update subjective beliefs, that we call dynamic imprecise probability kinematics (DIPK). It takes into account 1. the ambiguity faced by an agent in expressing their initial opinions (the agent is not able to specify a single prior) and 2. the uncertainty in the gathered data (new evidence is an uncertain perception rather than "crisp" data). Taking into account ambiguity and uncertain data allows one to produce reasonable models even when the information is imperfect or partial, and the scientist performing the analysis is not able to describe their initial beliefs about an event of interest through a single distribution. To deal with uncertainty in information gathering, DIPK is based on probability kinematics, an updating strategy that allows one to condition on uncertain events. We show that the sequence of successive DIPK updates of the set representing the agent's initial beliefs converges, and we study the properties of the updated sets. The results in chapter 3 give theoretical grounding to a highly

[2] The weak* topology is introduced in chapter 5.

applicable process for estimating and updating complex probabilities under severe uncertainty.

In chapter 4, we formulate an ergodic theory for the limit $\mathcal{P}^{\text{co}}_{\bar{\mathcal{E}}}$ of a sequence $(\mathcal{P}^{\text{co}}_{\mathcal{E}_n})$ of successive DIPK updates of a set $\mathcal{P}^{\text{co}}_{\mathcal{E}_0}$ representing the initial beliefs of an agent. As a consequence, we formulate a strong law of large numbers.

In chapter 5, we study the case in which the agent lacks knowledge about the composition of the state space Ω as well. In this situation, they need to formulate their initial beliefs via a set of extended probabilities, a generalization of regular probabilities whose codomain is $[-1, 1]$ instead of the usual unit interval $[0, 1]$. We study extended probabilities' properties and provide a behavioral interpretation. We also develop an ex ante analysis – taking place before the statistical analysis requiring knowledge of Ω – in which the true composition of Ω is progressively learned. Lastly, we describe how to update these extended probabilities sets as the composition of the state space is progressively discovered. We apply our findings to a species sampling problem and to the study of the boomerang effect (the empirical observation that sometimes persuasion yields the opposite effect: the persuaded agent moves their opinion away from the opinion of the persuading agent). The reason why this chapter is included is the following: we develop the concept of an extended Choquet capacity, and in particular of extended upper and lower probabilities (and give some of their properties). They are crucial because they completely characterize a convex and compact set of extended probability measures, which can be used to capture the ambiguity faced by the agent around which extended probability to start the analysis with. In this sense, we have a very general situation in which the agent does not fully know Ω and is incapable of selecting only one extended probability measure, but it is still able to carry out an analysis within the framework we build.

Chapter 6 concludes our work and provides some directions for future research. In appendix A we give an approximation of the joint distribution of the components

of the finite mixture model of chapter 2, and we provide the number of extrema of the convex hull within a unit simplex having the least amount of vertices. In appendix B we produce the proofs to our results. Appendix C is a note on the money bet approach to subjective probability of de Finetti.

2

Finite mixture models: a bridge with stochastic geometry and Choquet theory

2.1 Introduction

Finite mixture models go back at least to (Pearson and Erdmann III, 1894; Pearson, 1895) and have served as a workhorse in stochastic modeling (Everitt and Hand, 1981; Lindsay, 1995; Mengersen et al., 2011). Applications include clustering (McLachlan and Basford, 1988), hierarchical or latent space models (Little and Masyn, 2013), and semiparametric models (McNicholas, 2017) where a mixture of simple distributions is used to model data that is putatively generated from a complex distribution. In finite mixture models, the mixing distribution is over a finite number of components. There are also many examples of infinite mixture models in the Bayesian non-parametrics literature (Antoniak, 1974; Lavine, 1992; West et al., 1994).

In general, a finite mixture distribution of m components for a random vector Y is given by

$$Y \sim \sum_{k=1}^{m} p_k f(y; \theta_k), \quad \sum_{k=1}^{m} p_k = 1, \quad p_k \geqslant 0,$$

where the elements of the probability vector $p = (p_1, ..., p_m)^\top$ are mixture weights and θ_k denotes the parameter values for the k-th component.

Inference on the number of mixture components for finite mixture models can be difficult. In the Bayesian setting one can place a prior on the number of mixture components and use the posterior distribution to set the number of components (Miller and Harrison, 2018). In (Guha et al., 2020), the authors study the consistency of the posterior distribution of the number of clusters when a prior is placed on the number of clusters. They also propose a merge-truncate-merge procedure to consistently estimate the number of clusters from Dirichlet process mixture models. In (Fúquene et al., 2019), the authors propose using non-local priors for choosing the number of components in finite mixture models.

Another approach to inference on the number of components is to test whether the number of components is a given k or $k' > k$. The literature on testing the number of components is quite rich: classical results are summarized in (Titterington, 1990). More modern works include (Henna, 2005; Li and Chen, 2010; Chen et al., 2012). In the former, an estimator for the number of components is provided based on transformations of the observed data. The latter two propose an EM test for testing whether the number of true components in the mixture is some $k_0 > 0$, or is larger than k_0.

Another recent work of interest is (Ohn and Lin, 2020), where the authors use a data dependent prior and achieve optimal estimation of mixing measures, as well as posterior consistency for the number of clusters. They also consider a Dirichlet Process mixture to estimate a finite mixture model and show that the number of clusters can be used for consistent estimation on the number of components.

Rather than developing new tools for working with or applying finite mixture models, the main goal of this chapter is to establish connections between finite mixture models, stochastic convex geometry, and Choquet theory. We do so in the hope

that they will shed light on the workings and properties of finite mixture models. This chapter establishes a bridge between finite mixture models and stochastic geometry that allows to view finite mixture models as well-studied geometric objects. This insight allows to closely relate the number of components in a finite mixture model to the number of extrema of a convex body. Thereby, it facilitates studying the asymptotic growth rate and the asymptotic distribution of the number of components. This chapter bridges finite mixture models and (geometric) Choquet theory as well: we give a result to retrieve the weights in a finite mixture model using a uniqueness result by Gustave Choquet (Theorem 3) coupled with a Dirichlet process distribution.

The geometry of finite mixture models has primarily been studied in two contexts: differential geometry (Amari, 1983; Kass and Vos, 1997) and convex geometry (Lindsay, 1995; Marriott, 2002). The approach in this chapter is based on (stochastic) convex geometry. (Lindsay, 1995) was the first to observe that a mixture model can be seen as an element of the unit simplex in some Euclidean space $\mathbb{R}^J$. The focus was on identifiability of the weights of the mixture, a Carathéodory representation theorem for multinomial mixtures, and the asymptotic mixture geometry. (Marriott, 2002) bridges the differential and convex geometric approaches to identify restrictions for which the mixture can be written as more tractable geometric quantitites that can simplify inference problems. This chapter is similar in spirit to Lindsay's work, but uses more modern techniques from (Bárány and Buchta, 1993) and (Reitzner, 2005a).

Choquet theory in the context of finite mixture models has been inspected by (Hoff, 2003). There, the author develops an approach that uses Choquet's theorem for inference with the goal of estimating probability measures constrained to lie in a convex set, for example mixture models. The key observation in (Hoff, 2003) is that inference over a convex set of measures can be made via unconstrained inference

over the set of extreme measures. The main difference between this chapter and the approach developed in (Hoff, 2003) is that we consider a convex hull of points in a unit simplex rather than the convex hull of probability measures. Also, our goal is different: we use a result from Choquet theory to retrieve the weights in the finite mixture model at hand.

2.1.1 Setup of our work

We consider a finite mixture of multinomials. We start with the basic multinomial model where our observations X take J possible values $\{1, \ldots, J\}$ and $X \sim \text{Mult}(\pi)$, with $\pi \equiv (\pi_1, \ldots, \pi_J)^\top$ where $\pi_j = \mathbb{P}(X = j)$, with $\pi_j \geqslant 0$ for all j and $\sum_{j=1}^{J} \pi_j = 1$. A mixture of L multinomials can be specified as follows

$$X_i \sim \text{Mult}(\pi_i), \quad \pi_i = \sum_{\ell=1}^{L} \phi_{i,\ell} f_\ell,$$

where the probability vector $\phi_i \equiv (\phi_{i,1}, \ldots, \phi_{i,L})^\top$ assigns the probability of the i-th observation coming from the ℓ-th mixture component with multinomial parameters $f_\ell = (f_{\ell,1}, \ldots, f_{\ell,J})^\top$. Again $\sum_{j=1}^{J} f_{\ell,j} = 1$ with $f_{\ell,j} \geqslant 0$, and $\sum_{\ell=1}^{L} \phi_{i,\ell} = 1$ with $\phi_{i,\ell} \geqslant 0$. An important point throughout the chapter is that π_i belongs to the convex hull of probability vectors $\{f_1, \ldots, f_L\}$. The convex hull of $\{f_1, \ldots, f_L\}$ is a function of the identifiable elements of $\{f_1, \ldots, f_L\}$, that is, those elements that cannot be written as a convex combination of the other f_ℓ's. Hence, understanding the identifiable elements of this set provides information about the key model parameters.

In a Bayesian model we are interested in the posterior $\mathbb{P}(\theta \mid x_1, \ldots, x_n)$ where θ consists of the set $\{\phi_1, ..., \phi_n\}$ and $\{f_1, ..., f_L\}$, all of which are probability vectors. One can obtain point estimates of the parameters using an EM algorithm or the posterior using MCMC procedures; there are also variational approaches to compute the posterior. The finite mixture model we stated is an example of an admixture

model; the most popular admixture model is the latent Dirichlet allocation (LDA) model (Blei et al., 2003; Pritchard et al., 2000). A classic application of an admixture model is a generative process for documents. Consider a document as a collection of words; the LDA model posits that each document is a mixture of a small number of topics, and that these latter can be modeled by a multinomial distribution on the presence of a word in the topic. The hierarchical Dirichlet process (Teh et al., 2006), and generalizations thereof, may be considered as the natural nonparametric counterpart of the LDA model.

The probability vectors $\{\pi_i\}$ and the $\{f_\ell\}$ are all elements of Δ^{J-1}, the unit simplex on $\mathbb{R}^J$. Again, each of the π_i belong to the convex hull of $\{f_\ell\}$, or $\pi_i \in \text{Conv}(f_1, \ldots, f_L)$. Hence, an element of a convex hull in the Euclidean unit simplex represents (the distribution of) a finite mixture model.

Notice that the number of extrema of $\text{Conv}(f_1, \ldots, f_L)$, which we denote as M, will probably be less than L because some of the components f_ℓ are likely to be a convex combination of the others. A key concept in this chapter is that what we call the richest cheap model (RCM) representing π_i, that is, the finite mixture model representing π_i whose mixture components are $\{f_k\}_{k\in\mathcal{I}}$ such that $f_k \notin \text{Conv}(f_{\mathcal{I}\setminus\{k\}})$, for all $k \in \mathcal{I}$, $\mathcal{I} \subset \{1, \ldots, L\}$, and $\#\mathcal{I} = M$, where $\#$ denotes the cardinality operator. These conditions tell us that the M components of the richest cheap model are a subset of $\{f_1, \ldots, f_L\}$ and cannot be written as a convex combination of one another. By assuming – without loss of generality – that the identifiable elements in $\{f_1, \ldots, f_L\}$ are the first M ones, we can write the richest cheap model as

$$\pi_i = \sum_{\ell=1}^{M} \varphi_{i,\ell} f_\ell,$$

where we denote by $\varphi_i \equiv (\varphi_{i,1}, \ldots, \varphi_{i,M})^\top$ the probability vector that assigns the probability of the i-th observation coming from the ℓ-th identifiable mixture compo-

nent with multinomial parameters $f_\ell = (f_{\ell,1}, \ldots, f_{\ell,J})^\top$. Of course the $\varphi_{i,\ell}$'s are such that, for all i, $\sum_{\ell=1}^{M} \varphi_{i,\ell} = 1$, and $\varphi_{i,\ell} \geqslant 0$, for all ℓ. As we can see, the RCM captures the underlying complexity associted with the data at hand, using only the strictly necessary number of components.

We provide three main results. The first two – Theorems 11 and 13 – are very general, and state the following. Suppose we do not know what the components and the weights in our admixture model are, and we also do not know the number of components. Then, if we assume that the number of identifiable components M is a function $M(n)$ of the amount n of data we gather, we are able to tell the speed at which its expected value grows. The other main result – Theorem 19 – is more practical in nature. It states that if we know the number of identifiable components of the model, but not the components themselves nor the weights, we can place a Dirichlet process on the densities supported on Δ^{J-1} which eventually retrieves the weights. We also show how looking for the richest cheap model can be seen as an optimization problem, and we propose an algorithm to solve it.

Let us now inspect more in detail the structure of this chapter. In section 2.2, we let the number of identifiable mixture components M depend on the sample size n, that is, we let $M = M(n)$. This can be interpreted as proposing a prior on the number of identifiable mixture components that is a function of the sample size. We study the behavior of $M(n)$ as the number of observations increases. In Theorem 8 we show that if $M(n)$ is given by the cardinality of the extremal set of the convex hull of n elements sampled iid from the uniform over the simplex Δ^{J-1}, then the asymptotic growth rate of $\mathbb{E}[M(n)]$ is $(\log n)^{J-1}$. In Theorem 9 we state, retaining the same assumption on $M(n)$, a central limit theorem (CLT) for the distribution of the number of identifiable components of the admixture model. In Theorem 10 we prove that, as number of extrema of the convex hull grows to infinity, the convex

hull tends to an apeirogon, a polytope with infinitely many sides. In Theorem 11 we state that the $(\log n)^{J-1}$ asymptotic growth rate of the expectation of the number of identifiable components holds also when the n elements are drawn from a generic distribution, under a very mild assumption. We relax this latter assumption in Theorem 13.

We then consider inference when the number of identifiable admixture components is equal to J, but the admixture components and the admixture weights are unknown. In Theorem 19, we use a uniqueness result from Choquet theory (Theorem 3) to show that a Dirichlet process posterior always retrieves our admixture weights. We also give the rate of convergence of the Dirichlet process posterior to the (Dirac at the) weights.

In section 2.4, we use the idea of mixture models based on the extremal set to formulate a novel algorithm that outputs an admixture model composed of only extremal elements. We state the objective function the algorithm optimizes, and provide a two-stage algorithm. We apply this latter to the Associated Press data from the First Text Retrieval Conference (TREC-1), a large collection of terms used in 2246 documents.

2.2 Growth rates for extrema and mixture components

In this section, we build a bridge between finite mixture models and stochastic convex geometry. We give the growth rate of the expected number of identifiable components in our finite mixture model by studying the growth rate of the expected number of extrema of a convex body associated to our model. We also give a CLT for the number of identifiable components as a result of a CLT for the number of extrema of a convex body associated to our model.

We make the number of identifiable admixture components depend on the amount

n of data $x_1, \ldots, x_n$ we collect, that is, we have $M = M(n)$. In particular, let

$$S_1, \ldots, S_n \overset{iid}{\sim} \text{Uniform}(\Delta^{J-1}), \tag{2.1}$$

and call $K_n := \text{Conv}(s_1, \ldots, s_n)$, where s_j denotes the realization of S_j. Then, function M is defined as

$$M : \mathbb{N} \to \mathbb{N}, \quad n \mapsto M(n) := \#\text{ex}(K_n),$$

that is, $M(n)$ is given by the cardinality of the extremal set of K_n.

An obvious question, then, is what the asymptotic growth function based on draws from the uniform distribution on the unit simplex tells us about the asymptotic growth rate of the number of identifiable components based on draws from a generic distribution.

In a more general setting, when we drop the uniform assumption in (2.1), the number of extrema of the convex hull related to our model may be different from $M(n)$. We denote this quantity by T, and it is too going to be a function of the amount n of data we collect, that is, $T = T(n)$. In particular, suppose that $S_1, \ldots, S_n$ are now sampled iid from a generic distribution G on Δ^{J-1}. Call then $\breve{K}_n := \text{Conv}(s_1, \ldots, s_n)$. Then, function T is defined as

$$T : \mathbb{N} \to \mathbb{N}, \quad n \mapsto T(n) := \#\text{ex}(\breve{K}_n),$$

that is, $T(n)$ is given by the cardinality of the extremal set of $\breve{K}_n$.

Remark 7. Notice that $M(n), T(n) \geqslant J$ (of course, $J \geqslant 2$). If that is not the case, we can still have a convex hull, but it will be a proper subset of a smaller dimensional Euclidean space, and we are not interested in this eventuality.

2.2.1 Behavior of the extrema of K_n

We first state the growth rate of the expected number of extrema of the convex body built as the convex hull of uniform draws from the unit simplex. The growth rate is

based on results in (Bárány and Buchta, 1993; Reitzner, 2005b).

Let us briefly introduce the concept of an i-face. As pointed out in (Ziegler, 1995, Definition 2.1), in higher-dimensional geometry, the faces of a polytope are features of all dimensions. A face of dimension i is called an i-face. For example, the polygonal faces of an ordinary polyhedron are 2-faces. For any n-dimensional polytope, $-1 \leqslant i \leqslant n$, where -1 is the dimension of the empty set. Let us give a clarifying example. The faces of a cube comprise the cube itself (3-face), its facets (2-faces), the edges (1-faces), its vertices (0-faces), and the empty set (having dimension -1).

Given a generic n-dimensional polytope polytope P, we denote by $\mathcal{F}_i(P)$ one of its i-faces, $i \in \{-1, 0, \ldots, n\}$. We call $\mathscr{F}_i(P)$ the collection of its i-faces, and $F_i(P)$ the number of its i-faces, that is, $F_i(P) = \#\mathscr{F}_i(P)$, for all i. We also call a chain $\mathcal{F}_0(P) \subset \mathcal{F}_1(P) \subset \cdots \subset \mathcal{F}_n(P)$ of i-dimensional faces a tower of P.

Given these definitions, $F_0(K_n)$ denotes the number of extremal points of K_n.

Theorem 8. Let $K_n := \mathrm{Conv}(s_1, \ldots, s_n)$, where $S_1, \ldots, S_n$ are sampled as in (2.1). Then,

$$\lim_{n\to\infty} (\log n)^{-(J-1)} \mathbb{E}\left[F_0(K_n)\right] = \frac{1}{(J+1)^{J-1}(J-1)!} T(\Delta^{J-1}) =: c(J), \tag{2.2}$$

where $T(\Delta^{J-1})$ is the number of towers of Δ^{J-1}.

Notice that $F_0(K_n)$ corresponds to $M(n)$. Theorem 8 tells us that the expected number of identifiable mixture components grows at rate $(\log n)^{J-1}$.

Furthermore we can state the limiting distribution of $F_0(K_n)$; specifically, we give the following central limit theorem for $F_0(K_n)$. It immediately implies the same result for $M(n)$. We denote by $\mathbb{V}[F_0(K_n)]$ the variance of the number of extreme points of K_n.

Theorem 9. Let $K_n := \text{Conv}(s_1, \ldots, s_n)$, where $S_1, \ldots, S_n$ are sampled as in (2.1). Then,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{F_0(K_n) - \mathbb{E}[F_0(K_n)]}{\sqrt{\mathbb{V}[F_0(K_n)]}} \leqslant t \right) = \Phi(t) \tag{2.3}$$

where Φ is the cumulative distribution function of the standard normal distribution.

The last result in this section is about the shape that K_n converges to asymptotically. The next theorem states that, as the number of extreme points goes to infinity, the convex hull of these points converges to an apeirogon, a polytope with infinitely many sides.

Theorem 10. Let $K_n := \text{Conv}(s_1, \ldots, s_n)$, where $S_1, \ldots, S_n$ are sampled as in (2.1). If $F_0(K_n)$ grows to infinity, then K_n tends to an apeirogon.

2.2.2 On the expected number of components in a more general model

In this section we relax the assumption in (2.1). The main idea is that we can use the result in Theorem 8 to prove results in a more general setting, that is, a setting in which we do not require $S_1, \ldots, S_n$ to be uniformly distributed on the unit simplex Δ^{J-1}. In Theorem 11, we require that the expected number of identifiable admixture components is in a fixed linear relation with the expected number of identifiable admixture components in the simple uniform model. This can be interpreted as the stochasticity around the number of components entering the general model through the simple uniform one, and then being linearly passed on. In Theorem 13, we further weaken this already mild regularity condition. We only require that the sequence of rational numbers relating the expected number of identifiable admixture components in the uniform and in the general models does not have 0 as an accumulation point.

Theorem 11. Suppose that, for all n, we can always find a rational $\gamma \in \mathbb{Q}$ such that

$\mathbb{E}[T(n)] = \gamma \cdot \mathbb{E}[M(n)]$. Then,

$$\lim_{n\to\infty} (\log n)^{-(J-1)} \mathbb{E}[T(n)] = c'(J, \gamma). \tag{2.4}$$

The assumption that for all n we can always find $\gamma \in \mathbb{Q}$ – not depending on n – such that $\mathbb{E}[T(n)] = \gamma \cdot \mathbb{E}[M(n)]$ is mild. It means that there is a fixed (linear) relation between the expected number of identifiable admixture components of the uniform and the general models. This entails that all the stochasticity around the number of identifiable components enters the model through the number of extrema of the simple uniform model. Then, it is "passed" to the number of extrema of the more general model through coefficient γ.

Although the assumption in Theorem 11 should always be met in practice, it is still required, as we show in this next example.

Example 12. Let $\mathbb{E}[M(n)]$ be some function $g(n)$ of n, taking values on $\mathbb{N}$ for all n, and let

$$\mathbb{E}[T(n)] = \begin{cases} J + \lceil \log\log n \rceil & \text{, for all } n \leqslant \tilde{n} \\ J + \lceil \log n \rceil & \text{, for all } n > \tilde{n} \end{cases},$$

for some $\tilde{n} \in \mathbb{N}$. Then, we can find $\gamma \in \mathbb{Q}$ such that $J + \lceil \log\log n \rceil = \gamma \cdot g(n)$, for all $n \leqslant \tilde{n}$, and also $\gamma' \in \mathbb{Q}$ such that $J + \lceil \log n \rceil = \gamma' \cdot g(n)$, for all $n > \tilde{n}$. But we may have that $\gamma \neq \gamma'$. In this case, we cannot use the previous theorem to determine the growth rate of $\mathbb{E}[T(n)]$. △

In light of Example 12, we present a generalization of Theorem 11.

Theorem 13. Consider the sequence $(\gamma_n) \in \mathbb{Q}^{\mathbb{N}}$ such that $\mathbb{E}[T(n)] = \gamma_n \mathbb{E}[M(n)]$, for all $n \in \mathbb{N}$. Suppose that (γ_n) is such that 0 is not an accumulation point, and call r_{γ_n} a tight bound of (γ_n), that is, $\gamma_n / r_{\gamma_n} \to 1$, as $n \to \infty$. Then,

$$\lim_{n\to\infty} \frac{1}{r_{\gamma_n} (\log n)^{J-1}} \mathbb{E}[T(n)] = c(J). \tag{2.5}$$

Hence, the expected number of identifiable mixture components in this more general case grows at rate $r_{\gamma_n}(\log n)^{J-1}$. The growth rate, then, depends on r_{γ_n}, the limiting behavior of the sequence of rationals that links the simple uniform model to the general one.

In this section, we used ideas from stochastic convex geometry to state properties of a finite admixture model. This constitutes a novelty with respect to the standard approach in Bayesian analysis. These techniques have the potential to uncover other properties of finite mixture models that might otherwise be inaccessible.

Remark 14. It is immediate to see that there is a universal upper bound for the Euclidean distance between two points in a unit simplex: for all $x, y \in \Delta^{J-1}$, $d(x,y) \equiv \|x - y\| \leqslant 2$. This gives us an interesting result: the Hausdorff distance between K_n and $\breve{K}_n$ has a universal upper bound as well. Indeed,

$$d_H(K_n, \breve{K}_n) = \max\left\{ \sup_{x \in K_n} \inf_{y \in \breve{K}_n} d(x,y), \sup_{y \in \breve{K}_n} \inf_{x \in K_n} d(x,y) \right\} \leqslant 2.$$

Notice also that in Theorem 11 if – instead of requiring $\mathbb{E}[T(n)] = \gamma \cdot \mathbb{E}[M(n)]$ – we are willing to make the slightly stronger assumption that $T(n) = \rho \cdot M(n)$, $\rho \in \mathbb{Q}$ possibly different from γ, then we retrieve Theorem 9. This because, since $F_0(K_n) \equiv M(n)$, we have that

$$\frac{T(n) - \mathbb{E}[T(n)]}{\sqrt{\mathbb{V}[T(n)]}} = \frac{\rho F_0(K_n) - \mathbb{E}[\rho F_0(K_n)]}{\sqrt{\mathbb{V}[\rho F_0(K_n)]}} = \frac{F_0(K_n) - \mathbb{E}[F_0(K_n)]}{\sqrt{\mathbb{V}[F_0(K_n)]}},$$

and so Theorem 9 follows. In a similar fashion, if in Theorem 13 we require that, for all n, $T(n) = \rho_n \cdot M(n)$, $(\rho_n) \in \mathbb{Q}^{\mathbb{N}}$ possibly different from (γ_n), then we retrieve Theorem 9.

2.3 The Choquet measure and a prior on extremal points

In this section we build a bridge between finite mixture models and Choquet theory. We show how, thanks to a uniqueness result by Gustave Choquet, a Dirichlet process can be used to retrieve the mixture weights in a finite admixture model. We also give the rate of convergence of the Dirichlet process posterior to the Dirac measure on the weights.

A famous result by Choquet (Theorem 3) states that for every element p in a simplex C, there exists a unique measure – that we call the Choquet measure associated with p, and denote by ν_p – supported on the extrema $E = ex(C)$ such that $p = \sum_{e \in E} e \cdot \nu_p(e)$. In our analysis, p corresponds to π_i, the elements e in $E = ex(C)$ correspond to the identifiable f_ℓ's, and the $\nu_p(e)$'s correspond to the weights of the identifiable f_ℓ's.

In Theorem 19, we show that if we only assume that the number M of components is known and equal to J, a Dirichlet process retrieves ν_p. This result is important because $\nu_{\pi_i}(f_\ell) = \varphi_{i,\ell}$. The $\varphi_{i,\ell}$'s represent the weights of the richest cheap finite mixture model representing π_i, so $\pi_i = \sum_{\ell=1}^{M} f_\ell \nu_{\pi_i}(f_\ell)$.

By retrieving, we mean that given a Dirichlet process prior $DP(\alpha P_0)$ specified on the distributions supported on Δ^{J-1} having parameter $\alpha > 0$ and P_0 as base measure, its posterior converges weakly to the Dirac at ν_{π_i}. We also give the rate of convergence.

Remark 15. Recall that $\pi_i = \sum_{\ell=1}^{M} f_\ell \varphi_{i,\ell} = \sum_{\ell=1}^{L} f_\ell \phi_{i,\ell}$, where we labeled the unidentifiable components as $f_{M+1}, \ldots, f_L$, $M \leqslant L$. This is without loss of generality.

2.3.1 Choquet theory and extrema of convex bodies

Let us denote by $\mathcal{K}_M := \text{Conv}(f_1, \ldots, f_M) = \text{Conv}(f_1, \ldots, f_L)$ the convex hull generated by the M identifiable components of our finite mixture model. An example

of a simplex within the unit 2-simplex in $\mathbb{R}^3$ is given in Figure 2.1. Our first goal is to learn about distributions supported on the extrema of $\mathcal{K}_M$, $E_M := ex(\mathcal{K}_M)$.

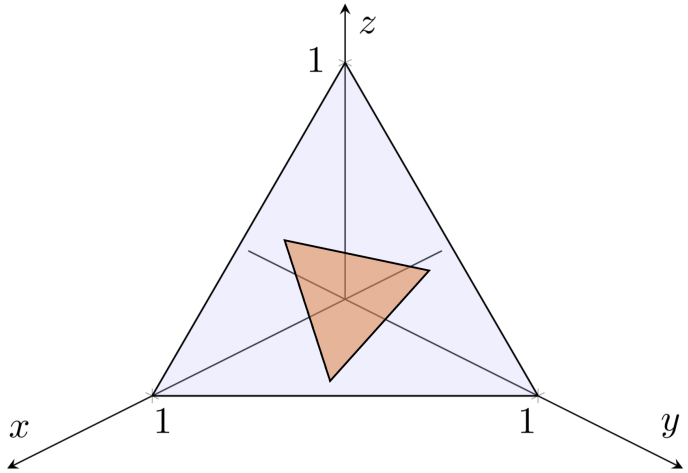

FIGURE 2.1: A triangular-shaped convex hull within the unit 2-simplex in $\mathbb{R}^3$. It is a simplex because it is the convex hull of its vertices.

Since Δ^{J-1} is locally convex, and $\mathcal{K}_M \subset \Delta^{J-1}$ is a metrizable compact convex set, then thanks to Theorem 1, we know that for every $p \in \mathcal{K}_M$, there exists μ supported on E_M such that $p = \sum_{f_j \in E_M} f_j \cdot \mu(f_j)$. In addition, if $\mathcal{K}_M$ is a simplex, then $\mu = \nu_p$ is the unique Choquet measure of Theorem 3.

Proposition 16. If $\mathcal{K}_M$ is a simplex, then every element in $p \in \mathcal{K}_M$ can be represented by a unique measure ν_p (the Choquet measure representing p) supported on E_M.

Remark 17. For any convex set built as the convex hull of M points in the simplex Δ^{J-1}, we can always find a subset that is of particular interest. Indeed, consider M points $f_1, \ldots, f_M$ in Δ^{J-1} that cannot be written as a convex combination of one another; call $\mathcal{K}_M$ their convex hull and $E_M := ex\mathcal{K}_M$. Let us also refer to the set of probability measures on $(E_M, \sigma(E_M))$ by $\Delta(E_M, \sigma(E_M))$, where $\sigma(E_M) = 2^{E_M}$. Then by Proposition 16, if $\mathcal{K}_M$ is a simplex, we have that for all $p \in \mathcal{K}_M$, there exists a unique measure $\nu_p \in \Delta(E_M, \sigma(E_M))$ such that $p = \sum_{f_j \in E_M} f_j \cdot \nu_p(f_j)$. Hence, we can

always retrieve the set $N_{\mathcal{K}_M} \subset \Delta(E_M, \sigma(E_M))$ of Choquet measures associated with the (elements of the) convex set $\mathcal{K}_M$. It is immediate to see that there is a bijection between $N_{\mathcal{K}_M}$ and $\mathcal{K}_M$. An important consequence of Proposition 16 is that the elements of $\mathcal{K}_M$ inherit the properties of the elements of E_M. The Choquet measures are the "channel" through which the properties are "carried" from the extrema to the points in the convex hull.

2.3.2 Choquet theory for mixture weights

The Dirichlet process (DP) is of fundamental importance in Bayesian nonparametrics (Ferguson, 1974; Lavine, 1992, 1994; Ghosh and Ramamoorthi, 2003; Ghosal and van der Vaart, 2017). It is a default prior on spaces of probability measures, and a building block for priors on other structures. Because it "selects" almost surely discrete distributions, it is a sensible choice for our case, since the Choquet measure is supported on the finite set E_M. The DP possesses the conjugacy property.

Theorem 18. (Ferguson, 1974) Suppose that $P \sim DP(\alpha P_0)$, for some $\alpha > 0$ and base measure P_0, and we collect iid observations $X_1, \ldots, X_n \mid P \sim P$. Then, (a version of) the posterior for P is given by $DP(\alpha P_0 + \sum_{j=1}^{n} \delta_{X_j}) \equiv DP(\alpha P_0 + nP_n)$.

Here $P_n = 1/n \sum_{j=1}^{n} \delta_{X_j}$ denotes the empirical distribution of the observations. An extensive treatment of DPs is given in (Ghosal and van der Vaart, 2017, Chapter 4).

The following theorem states that, given convex body $\mathcal{K}_M$, if $\mathcal{K}_M$ is a simplex, then for all $p \in \mathcal{K}_M$ we can select a DP that always retrieves the Choquet measure ν_p. This is a useful result because, despite Proposition 16 tells us that every $p \in \mathcal{K}_M$ can be represented by a unique ν_p supported on the extrema of $\mathcal{K}_M$, it does not give the analytic form of ν_p.

The Choquet measure ν_{π_i} accruing to π_i, is important because it gives us the weights $\varphi_{i,\ell}$'s, that is $\nu_{\pi_i}(f_\ell) = \varphi_{i,\ell}$, for all $\ell \in \{1, \ldots, M\}$. This means that, using

the DP approach described in the next theorem, we are able to identify the mixture weights of the richest cheap finite mixture model.

Theorem 19. Let $\mathcal{K}_M$ be a simplex. If $e_1, \ldots, e_k$ are an iid sample of elements of E_M from ν_{π_i}, then

$$DP\left(\alpha P_0 + \sum_{j=1}^{k} \delta_{e_j}\right) \xrightarrow[k\to\infty]{w} \delta_{\nu_{\pi_i}} \quad \nu_{\pi_i}\text{-a.s.} \tag{2.6}$$

where α is a positive real, P_0 is a base measure supported on Δ^{J-1}, $\xrightarrow{w}$ denotes the weak convergence, and $\delta_{\nu_{\pi_i}}$ is the Dirac measure at ν_{π_i}. In addition, the rate of convergence is given by $k^{-1/2}$.[1]

The idea is that we recover the Choquet measure ν_{π_i}. Because the DP is a measure over measures, the formal statement is convergence to $\delta_{\nu_{\pi_i}}$. We need base measure P_0 to be supported on Δ^{J-1} because it has to give positive probability to all the elements of the unit simplex.

2.4 A procedure to find the richest cheap model

In this section we use the idea of a mixture model based on the extremal set to provide a procedure which finds the richest cheap admixture model. We apply the procedure to a document-term matrix (Harman, 1992) and examine the number of topics inferred and the word frequency distribution of the topics.

In an admixture model, there are two sets of parameters:

1. the mixing weights for each individual, which we denote as a matrix Φ where $\Phi_{i,j}$ is the probability that the i-th sample is drawn from the j-th component. Each row of Φ is the mixture vector of the i-th observation $\phi_i = (\phi_{i,1}, \ldots, \phi_{i,L})$;

[1] Relative to the total variation metric.

2. the probability vectors parameterizing each mixture component, which we can write as a matrix F whose j-th column is f_j.

The relation between admixture modeling and sparse factor analysis (SFA) was explored in detail in (Engelhardt and Stephens, 2010). There, conditions were provided when SFA and LDA have very similar results, and the implications for population genetics were discussed. The key insight in (Engelhardt and Stephens, 2010) is that given an observation matrix $X = [x_1, ..., x_n]$ from a binomial admixture model, learning an admixture model amounts to the following minimization procedure

$$\min_{F,\Phi} \|\mathbb{E}[X] - \Phi F\|^2. \tag{2.7}$$

The SFA framework can be summarized as minimizing (2.7) with the constraint that many of the elements of Φ will be zero, or that every observation is a sparse combination of each component. The spirit behind the algorithm proposed in this section is to think of sparsity as the extremal set: we want to find a set of components that are extremal yet still accurately solves the above minimization.

We first state the likelihood for the admixture model, assuming a maximum of L components,

$$\mathcal{L}(X_1, ..., X_n; \{\phi_1, ..., \phi_n\}, \{f_1, ..., f_L\}) = \prod_{i=1}^{n} \text{Mult}\left(\pi_i = \sum_{\ell=1}^{L} \phi_{i,\ell} f_\ell\right).$$

The maximum likelihood estimator for for the above model is

$$\{\{\widehat{\phi}_1, ..., \widehat{\phi}_n\}, \{\widehat{f}_1, ..., \widehat{f}_L\}\} \equiv \underset{\{\phi_1,...,\phi_n\},\{f_1,...,f_L\}}{\arg\max} \mathcal{L}(X_1, ..., X_n; \{\phi_1, ..., \phi_n\}, \{f_1, ..., f_L\}). \tag{2.8}$$

A notion of sparsity related to the SFA framework is to maximize the likelihood subject to the the constraint that components are identifiable, that is, no component

can be represented as a convex combination of other components. We consider the procedure that will maximize the following objective function

$$\begin{aligned} &\underset{\mathcal{I},\{\phi_1,\ldots,\phi_n\},\{f_k\}_{k\in\mathcal{I}}}{\operatorname{argmax}} \quad \prod_{i=1}^{n} \operatorname{Mult}\left(\pi_i = \sum_{k\in\mathcal{I}} \phi_{i,k} f_k\right). \\ &\text{subject to} \qquad\qquad f_k \notin \operatorname{Conv}(f_{\mathcal{I}\setminus\{k\}}),\ \forall k \in \mathcal{I}, \end{aligned} \tag{2.9}$$

where $\mathcal{I}$ is a subset of the set $\{1, ..., L\}$ and is the collection of the indices of the extremal set. Constraint $f_k \notin \operatorname{Conv}(f_{\mathcal{I}\setminus\{k\}})$, for all $k \in \mathcal{I}$, ensures that no mixture component is contained in the convex combination of the others. Notice that the cardinality of $\mathcal{I}$ represents the number of components M of the richest cheap model, as we described in section 2.1.1.

The maximization specified by equation (2.9) is non-convex and finding the global optima is difficult. We propose a two-step procedure to solve equation (2.9). We first set the number of components L to be arbitrarily large and compute the standard maximum likelihood estimator specified in (2.8). The result of the first step are the parameters $\{\{\widehat{\phi}_1, ..., \widehat{\phi}_n\}, \{\widehat{f}_1, ..., \widehat{f}_L\}\}$. In the second step we compute the convex hull of $\{\widehat{f}_1, ..., \widehat{f}_L\}$ and we consider the cardinality M of its extremal set

$$M := \#ex(\operatorname{Conv}(\{\widehat{f}_1, ..., \widehat{f}_L\}),$$

where we denote by # the cardinality operator. If M is smaller than L, we rerun the MLE in (2.8) with M components, otherwise we stop and keep the current parameters. The parameters obtained from the second iteration of the MLE are estimates of the parameters of the richest cheap model. Notice that computing the convex hull is evocative of the Choquet procedure described in section 2.3. If $M = J$, at the end of this algorithm, we have an estimate $\widehat{\nu}_{\pi_i}$ of the Choquet measure for π_i, since $\widehat{\nu}_{\pi_i}(\widehat{f}_\ell) = \widehat{\varphi}_{i,\ell}$, for all $\ell \in \{1, \ldots, M\}$, where we denote the estimates of the weights of the richest cheap model as $\widehat{\varphi}_{i,\ell}$, for all ℓ.

We applied our two-step minimization procedure to a well studied dataset (Bail, 2018) which is a document-term matrix consisting of term frequencies of 10473 terms in 2246 documents collected from Associated Press documents (Harman, 1992). We used the latent Dirichlet allocation (LDA) function in the R package `topicmodels` (Grün and Hornik, 2011) to compute the MLE. We used the convex hull function in the R package `geometry` to compute the convex hull. Computing the convex hull over the full topic frequency vectors – elements belonging to simplex Δ^{10472} – is prohibitive and also does not make sense when the number of topics are less than 10472. We used principal components analysis (PCA) to project the frequency vectors of the topics onto a lower dimensional space and then computed the convex hull of the projections. We used a simple scree plot to notice that $3-5$ dimensions are sufficient to capture 30% of the variation when we carry out our analysis specifying $L = 200$ initial topics. If we choose $L < 200$ initial topics, we have that $3 - 5$ dimension explain more than 30% of the variation. We only need to compute the number of extrema of the convex hull and not the extremal elements themselves in our procedure, so it suffices to compute the convex hull in the low dimensional space.

We ran the above procedure on the document-term matrix specifying the initial number of topics as 50, 100, 150, and 200. Given the results in the PCA step, we projected down to 5 dimensions. The number of extremal points – i.e. the number of topics – we obtained were 12, 11, 9, 8 with initialization 50, 100, 150, and 200, respectively. The number of topics we obtained is similar to the number obtained in previous studies: the majority of these latter use 9 to 12 topics (Bail, 2018; Hou, 2017).

3

Dynamic Precise and Imprecise Probability Kinematics

3.1 Introduction

Updating an opinion on the likelihood of an event when new data becomes available is one of the most natural tasks we perform daily. The goal of this paper is to introduce a method to update mechanically an agent's subjective beliefs in the presence of severe uncertainty.

In particular, we will consider an agent facing ambiguity and collecting partial information. With the former, we mean that a single probability measure is not enough to encapsulate the agent's initial beliefs, a very common and well documented situation (Walley, 1991, Section 1.1.4). We inspect ambiguity in Section 3.1.1. Partial information means that the agent cannot collect crisp evidence; rather, they gather information whose nature is probabilistic. Our updating mechanism is based on probability kinematics (PK), an updating rule expressly conceived to deal with partial information. We inspect probability kinematics and its relation with the procedure we present in Section 3.1.2.

We call the method we propose dynamic imprecise probability kinematics, whose acronym is DIPK. It is framed within the credal sets theory paradigm. In this field, a set of probability measures (called a credal set) is used to capture either the agent's initial uncertainty, or inconsistency/imprecision in the process of collecting data. To derive DIPK, we first assume that the agent does not face ambiguity. We come up with a simpler updating technique that we call dynamic probability kinematics (DPK), and then we generalize it by requiring the agent to specify a set $\mathcal{P}$ of probability measures representing their initial beliefs. DIPK is especially useful because it allows the update to be performed mechanically: the agent only needs to specify $\mathcal{P}$. To the best of our knowledge, this is the first time a PK-rooted mechanical procedure to update subjective beliefs in the presence of ambiguity and partial information within the credal sets theory paradigm is presented.

3.1.1 Ambiguity

Precise probabilities are widely employed as the central vocabulary of many modes of uncertainty reasoning, nearly exclusively so in statistical inference, for example. In the subjective probability literature, the agent's initial beliefs about an event $A \subset \Omega$ are usually encapsulated in a single probability measure, that is then refined once new information in the form of data become available. As Walley points out in (Walley, 1991, Section 1.1.4), though, missing information and bounded rationality may prevent the agent from assessing probabilities precisely in practice, even if doing so is possible in principle. This may be due to the lack of information on how likely events of interest are, lack of computational time or ability, or because it is extremely difficult to analyze a complex body of evidence. We call this condition faced by agent *ambiguity* (Ellsberg, 1961). Often times agents do not realize they face ambiguity, as observed in (Berger, 1984) and in the de Finetti lecture delivered at ISBA 2021. There, Berger points out how most people tend to under-report variance; the folklore

says by a factor of 3. People simply think that they know more than they actually do.

In the presence of ambiguity, the agent may only be able to specify a set $\mathcal{P}$ of probability measures that seem "plausible" or "fit" to express their initial opinion on the events of interest. Generally speaking, the farther apart (e.g. in the total variation distance) the "boundary elements" of $\mathcal{P}$ (i.e. its infimum and supremum), the higher the agent's uncertainty. This way of proceeding, called the *sensitivity analysis approach*, is inspected in depth in Remark 29.

As Section 3.6 will discuss, the infima of the sets updated according to our DIPK procedure – that, as we shall see, are called lower probabilities – completely characterize the sets. That is why in Section 3.7 we give lower and upper bounds for the updated lower and upper probabilities, respectively, and in Section 3.8 we study the behavior of the updated sets (contraction, dilation, sure loss) by giving sufficient conditions involving lower (and upper) probabilities.

3.1.2 Probability kinematics

DPK and DIPK are rooted in probability kinematics (PK), also known as Jeffrey's rule of updating. PK can be seen as a generalization of Bayesian updating, the most famous and widely used technique to describe updating of beliefs. This latter prescribes the scholar to form an initial opinion on the plausibility of the event A of interest, where A is a subset of the state space Ω, and to express it by specifying a probability measure P, so that $P(A)$ can be quantified. Once some data E is collected, the Bayesian updating mechanism revises the initial opinion by applying the Bayes rule

$$P^\star(A) \equiv P(A \mid E) = \frac{P(A \cap E)}{P(E)} = \frac{P(E \mid A)P(A)}{P(E)} \propto P(E \mid A)P(A),$$

provided that $P(E) \neq 0$.[1] In (Jeffrey, 1957, 1965, 1968), Richard Jeffrey makes a compelling case of the fact that Bayes rule is not the only reasonable way of updating. For example, its use presupposes that both $P(E)$ and $P(A \cap E)$ have been quantified before event E takes place: this can be a very challenging task, for example when E is not anticipated. As we can see, Bayes rule is not well-suited for the agent to face partial information. The following example illustrates a situation in which Bayes rule is not *directly* applicable to compute the updated probability of an event, but Jeffrey's rule can be applied.

Example 20. (Diaconis and Zabell, 1982, Section 1.1) Three trials of a new surgical procedure are to be conducted at a hospital. Let 1 denote a successful outcome, and 0 an unsuccessful one. The state space has the form

$$\Omega = \{000, 001, 010, 011, 100, 101, 110, 111\}.$$

A colleague informs us that another hospital performed this type of procedure 100 times, registering 80 successful outcomes. This information is relevant and should influence our opinion about the outcome of the three trials, but it cannot be put in direct terms of the occurrence of an event in the original Ω, thus Bayes rule is not directly applicable.

Since the description contains no information about the order of the three trials, our initial opinion P assumes that they are exchangeable. That is, consider the partition $\{E_0, E_1, E_2, E_3\}$ of Ω where E_j is the set of all outcomes with exactly j successes, exchangeability implies that we assign equal probabilities to atomic events within each partition. In other words, $P(\{001\}) = P(\{100\}) = P(\{010\})$ and $P(\{110\}) = P(\{101\}) = P(\{011\})$.

[1] Although conditioning on a zero probability event is technically possible, see e.g. literature on lexicographic probability (Blume et al., 1991) and layers of zero probabilities (Coletti and Scozzafava, 2002), we do not consider this eventuality in the present work.

The success rate at the other hospital informs our opinion over the partition $\{E_j\}$ only, and nothing more. In relation to our old opinion P, our updated opinion $P^\star$ satisfies $P(A \mid E_j) = P^\star(A \mid E_j)$ for all $A \subset \Omega$ and all $j \in \{0, \ldots, 3\}$. Upon specifying a new subjective assessment of the $P^\star(E_j)$'s, the updated probability measure $P^\star$ can be fully reassessed by the relation

$$P^\star(A) = \sum_{j=0}^{3} P^\star(A \mid E_j) P^\star(E_j) = \sum_{j=0}^{3} P(A \mid E_j) P^\star(E_j).$$

It is within our liberty to reassess the $P^\star(E_j)$'s. We may, for example, regard the three trials as a random subsample of size three from those of the other hospital. This would equate $P^\star(E_j)$ to the probability of obtaining j successes from a Hypergeometric$(100, 80, 3)$ distribution. $\triangle$

The rule $P^\star(A) = \sum_{E_j \in \mathcal{E}} P(A \mid E_j) P^\star(E_j)$ is known as Jeffrey's rule of conditioning. It is valid when there is a partition $\mathcal{E}$ of the sample space Ω such that

$$P^\star(A \mid E_j) = P(A \mid E_j), \quad \forall A \subset \Omega, \forall E_j \in \mathcal{E}. \tag{3.1}$$

As pointed out in (Walley, 1991, Section 6.11.8), under assumption (3.1), Jeffrey's rule is a consequence of coherence. It is useful when new evidence cannot be identified with the occurrence of an event, but has the effect of changing the probabilities we assign to the events in partition $\mathcal{E}$. It has the practical advantage of reducing the assessment of $P^\star$ to the simpler task of assessing $P^\star(E_j)$, for all $E_j \in \mathcal{E}$. In the above example, instead of a full reassessment of probabilities on Ω, the agent only needs to deliberate new assessment of the four probabilities $P^\star(E_0)$ through $P^\star(E_3)$ based on the given information.

To see that Jeffrey's rule of conditioning is a generalization of Bayes rule, consider partition $\{E, E^c\}$, for some $E \subset \Omega$. Then if $P^\star(E) = 1$, we have that $P^\star(A) = P(A \mid E)P^\star(E) + P(A \mid E^c)P^\star(E^c) = P(A \mid E)$, which is Bayes rule. In addition, as

studied in (Diaconis and Zabell, 1982, Section 2), if we are given the couple $\{P, P^\star\}$ of probability measures, we can always reconstruct a partition $\{E_j\}$ for which $\{P, P^\star\}$ could have arisen via Jeffrey's updating rule, unlike Bayesian conditionalization.

Let us now discuss the relation between DPK and Jeffrey's updating. The three main tasks in PK are:

(1) Collecting a partition $\mathcal{E}$ of state space Ω;

(2) Subjectively assess the probability $P^\star(E)$ to attach to the elements E of partition $\mathcal{E}$;

(3) Compute the update $P^\star(A) = \sum_{E \in \mathcal{E}} P(A \mid E) P^\star(E)$.

In DPK, we:

(1') Collect data points belonging to a generic set $\mathcal{X}$ that induce a partition $\mathcal{E}$ of state space Ω;

(2') Mechanically attach probabilities to the elements of the induced partition;

(3') Compute the update as in "regular" PK.

We allow the evidence observed by the agent to belong to a general set $\mathcal{X}$; data points are regarded as the realization of a random variable $X : \Omega \to \mathcal{X}$. Notice that if the distribution P_X of X were to be known, the elements of $\mathcal{X}$ would induce a unique partition $\mathcal{E} = \{E_j\}$ of Ω, where $E_j = \{\omega \in \Omega : X(\omega) = x_j\}$ and $P^\star(E_j) = P_X(\{x_j\})$, for all $x_j \in \mathcal{X}$. Instead, to further capture the idea of partial information, we consider the case where P_X is unknown. As we shall see, given data points $x_1, \ldots, x_n \in \mathcal{X}$, they induce a partition $\mathcal{E}_1 = \{E_j\}_{j=1}^{m+1}$, $m \leqslant n$, where m is the number of unique elements in $\{x_1, \ldots, x_n\}$, $E_j = \{\omega \in \Omega : X(\omega) = x_j\}$ for $j \in \{1, \ldots, m\}$, and $E_{m+1} = (\cup_{j=1}^{m} E_j)^c$. The relative frequency of $x_1, \ldots, x_n$ will induce the probability that the agent assigns

to the elements of $\mathcal{E}_1$, making the update from P to $P^\star$ mechanical. We inspect subsequent DPK updates in Section 3.5.

3.1.3 Structure of the paper

The paper is organized as follows. In Section 3.2, we discuss the connection between our work and the existing literature. In Section 3.3, we introduce dynamic probability kinematics (DPK). In Section 3.4, we provide the mechanical version of DPK that we use throughout the rest of the work. It is mechanical in the sense that it does not require the agent to subjectively specify the probability to attach to the elements of the partition, but it does so mechanically. In Section 3.5, we explain how to subsequently update probability measure P as more and more data become available. In Section 3.6, we introduce dynamic imprecise probability kinematics (DIPK). In Section 3.7, we give bounds for the upper and lower probabilities associated with the updated probability set. In Section 3.8, we study the behavior of updated sets of probabilities, namely contraction, dilation, and sure loss. Section 3.9 presents two examples that illustrate how to implement DPK and DIPK. Section 6 concludes our work. Appendix B contains the proofs of our results.

3.2 Related literature

In this Section, we present some papers that deal with Jeffrey's updating in the context of imprecise probability models. Probability kinematics has been generalized to be put to use in the context of Dempster-Shafer theory, evidence theory, neighborhood models theory, possibility theory, maximum entropy theory, and credal sets theory. DIPK belongs to this last category.

In (Shafer, 1981), Shafer discusses Jeffrey's updating from a philosophical perspective, and is the first to consider its application to the context of Dempster-Shafer theory, for which belief functions – functions representing the degree of belief of the

agent on a given event – and Dempster's rule of combination play a central operational role. In (Ichihashi and Tanaka, 1989) and (Smets, 1993) the authors further study the generalization of Jeffrey's updating for belief functions defined on a finite sample space. In (Ichihashi and Tanaka, 1989), the authors point out how Shafer's approach is different from the normative Bayesian approach and is not a straight generalization of Jeffrey's rule, so they propose rules of conditioning for which Jeffrey's rule is a direct consequence of a special case. In (Smets, 1993), the author generalizes the results in (Ichihashi and Tanaka, 1989). He shows that several forms of Jeffrey's updating rule can be defined so that they correspond to the geometrical rule of conditioning and to Dempster's rule of conditioning, respectively.

In (Ma et al., 2011), the authors provide a generalization of both Jeffrey's rule and Dempster conditioning to propose an effective revision rule in the field of evidence theory. This is very interesting since when one source of evidence is less reliable than another, the idea is to let prior knowledge of an agent be altered only by some of the input information. The change problem is thus intrinsically asymmetric. To this extent, their model takes into account inconsistency between prior and input information. Other works that deal with a generalization of Jeffrey's rule within the framework of evidence theory are (Tang et al., 2004), in which the authors propose a generalization of probability kinematics where a priori knowledge and new evidence are all modeled by independent random sets, and (Lv et al., 2007) in which a priori knowledge and evidences are modelled by a probability distribution and a collection of multi-dimensional random sets, respectively.

In (Škulj, 2006), the author discusses the application of Jeffrey's rule to neighborhood models theory. In this field, uncertainty is captured by neighborhood of a classical probability measure P, presented in the form of interval probabilities $[L, U]$. This means that $P(A) \in [L(A), U(A)]$, for all $A \subset \Omega$, where Ω is the state space of interest. The author shows that a neighborhood $[L, U]$ of a probability measure P

whose lower envelope L is convex or bi-elastic with respect to the base probability measure (Škulj, 2006, Definitions 3 and 4) is closed with respect to Jeffrey's rule of conditioning. This means that Jeffrey's posterior for $Q \in [L, U]$ still belongs to the interval.

Possibility theory (Zadeh, 1978) is a framework alternative to probability theory that is suitable for handling uncertain, imprecise and incomplete knowledge. In possibility theory, there are two different ways to define the conditioning depending on how possibility degrees are interpreted, one called quantitative possibility and the other called qualitative possibility. In (Benferhat et al., 2011), the authors investigate the existence and uniqueness of the posterior probabilities computed according to a possibilistic counterpart of Jeffrey's rule in both the quantitative and qualitative possibilistic frameworks.

In (Marchetti and Antonucci, 2018), the authors generalize Jeffrey's rule to credal sets theory. The authors introduce imaginary kinematics (Marchetti and Antonucci, 2018, Definition 7). They combine Jeffrey's rule with Lewis' imaging (Lewis, 1976) for credal sets to be able to update beliefs when possibly inconsistent probabilistic evidence is gathered. Evidence on some variables is called *inconsistent* when it contradicts certainty (or impossibility) in the agent's knowledge base. There are two main differences between our work and (Marchetti and Antonucci, 2018):

1. We consider an agent facing ambiguity who specifies a set of probability measures that encapsulates their initial beliefs, while (Marchetti and Antonucci, 2018) do not;

2. In (Marchetti and Antonucci, 2018) the authors consider the instance in which gathered evidence is partial and possibly inconsistent, while we only deal with the former.

In the future we will generalize DIPK by relaxing the (tacit) assumption that the

gathered evidence is consistent.

It is worth noting that in (Caprio and Mukherjee, 2021b) the authors provide an ergodic theory for the limit of a sequence of successive DIPK updates of a set representing the initial beliefs of an agent. As a consequence, they formulate a strong law of large numbers.

3.3 A new way of updating subjective beliefs

In this Section, we describe a new way of updating subjective beliefs based on Jeffrey's rule of conditioning (Diaconis and Zabell, 1982; Jeffrey, 1957, 1965, 1968), which we call dynamic probability kinematics (DPK). Let Ω be the state space of interest, and assume it is at most countable. The version of DPK with uncountable Ω will be the subject of a future work. Suppose that P is a probability measure on $(\Omega, \mathcal{F})$ representing an agent's initial beliefs around the elements of $\mathcal{F} = 2^{\Omega}$, and that we want to update it after collecting some data. The agent observes data points $x_1, \ldots, x_n$ that are realizations of a random quantity $X : \Omega \to \mathcal{X}$ whose distribution is unknown. Notice that collecting $x_1, \ldots, x_n$ is equivalent to observing $\omega_1, \ldots, \omega_n \sim Q$, where Q is unknown, and then computing $X(\omega_i) = x_i$. Consider now the collection $\mathcal{E}' := \{E_i\}_{i=1}^{n}$, where $E_i \equiv X^{-1}(x_i) := \{\omega \in \Omega : X(\omega) = x_i\}$. It induces partition $\mathcal{E} = \{E_j\}_{j=1}^{m+1}$ of Ω, $m \leqslant n$, whose first m elements are the unique elements of $\mathcal{E}'$, and $E_{m+1} = \Omega \backslash \cup_{j=1}^{m} E_j$.

As an update to P, we propose the following

$$
\begin{aligned}
P_{\mathcal{E}} : \mathcal{F} \to [0,1], \quad A \mapsto & P_{\mathcal{E}}(A) := \sum_{E_j \in \mathcal{E}} P(A \mid E_j) P_{\mathcal{E}}(E_j) \\
& \text{such that } P_{\mathcal{E}}(E_j) \geqslant 0, \forall E_j \in \mathcal{E}, \text{ and } \sum_{E_j \in \mathcal{E}} P_{\mathcal{E}}(E_j) = 1.
\end{aligned}
\tag{3.2}
$$

We have the following.

Proposition 21. $P_{\mathcal{E}}$ is a probability measure, and it is a Jeffrey's posterior for P.

In general, Jeffrey's rule of conditioning – as presented in (Diaconis and Zabell, 1982, Equation 1.1) – is given by $P^\star(A) = \sum_j P(A \mid E_j)P^\star(E_j)$, where $P^\star$ is Jeffrey's posterior for P. It is valid when Jeffrey's condition is met, that is, when there is a given partition $\{E_j\}$ of the state space Ω such that $P(A \mid E_j) = P^\star(A \mid E_j)$ is true for all $A \in \mathcal{F}$ and all j. Specifically, this condition is met by $P_\mathcal{E}$. Since $P_\mathcal{E}$ is a probability measure by Proposition 21, it is true that, for all $A \in \mathcal{F}$, $P_\mathcal{E}(A) = \sum_{E_j \in \mathcal{E}} P_\mathcal{E}(A \mid E_j)P_\mathcal{E}(E_j)$. But given our definition for $P_\mathcal{E}$, we also have that $P_\mathcal{E}(A) = \sum_{E_j \in \mathcal{E}} P(A \mid E_j)P_\mathcal{E}(E_j)$. This implies that there is a partition $\mathcal{E}$ for which $P(A \mid E_j) = P_\mathcal{E}(A \mid E_j)$ is true for all $A \in \mathcal{F}$ and all $E_j \in \mathcal{E}$.

3.4 A mechanical procedure to compute $P_\mathcal{E}$

In this Section, we show how computing $P_\mathcal{E}(A)$ can be performed as a mechanical procedure, if the analyst is unwilling or unable to make full subjective probabilistic assessment for the elements of $\mathcal{E}$. Recall that $\mathcal{E}' = \{E_i\}_{i=1}^n = \{X^{-1}(x_i)\}_{i=1}^n$, and $\mathcal{E} = \{E_j\}_{j=1}^{m+1}$, where $E_1, \ldots, E_m$ are the unique elements of $\mathcal{E}'$, and $E_{m+1} = (\cup_{j=1}^m E_j)^c$. Then, consider the empirical probability measure $P^{emp} \in \Delta(\Omega, \mathcal{F})$ such that, if $E_{m+1} \neq \varnothing$,

$$P^{emp}(E_j) = \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{I}(E_j = E_i), \quad \text{for all } j \in \{1, \ldots, m\}, \tag{3.3}$$

where $\mathbb{I}$ denotes the indicator function, and

$$P^{emp}(E_{m+1}) = 1 - \sum_{j=1}^{m} P^{emp}(E_j). \tag{3.4}$$

If instead $E_{m+1} = \varnothing$,

$$P^{emp}(E_j) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(E_j = E_i), \quad \text{for all } j \in \{1, \ldots, m\} \tag{3.5}$$

and

$$P^{emp}(E_{m+1}) = 0. \tag{3.6}$$

We require that

$$P_{\mathcal{E}}(E_j) = \beta(n)P(E_j) + [1 - \beta(n)]\, P^{emp}(E_j), \quad \forall E_j \in \mathcal{E}, \tag{3.7}$$

where $\beta(n)$ is a coefficient in $[0, 1]$ depending on n: the posterior probability $P_{\mathcal{E}}$ assigned to the elements E_j of partition $\mathcal{E}$ is a mixture of the prior P and the empirical probability measure P^{emp}. Performing the update in (3.2) becomes then a mechanical procedure. Jeffrey's updating procedure becomes easier (the analyst does not have to subjectively specify the probability assigned to all the elements of $\mathcal{E}$, a task that can be mentally and mathematically challenging). In the remainder of this document, we are going to use the mechanical procedure we just described to assign updated probabilities to the elements of $\mathcal{E}$.

Remark 22. There is a subtlety in moving from P to $P_{\mathcal{E}}$. Let $\breve{P} := \beta(n)P + [1 - \beta(n)]P^{emp}$. Requiring that $P_{\mathcal{E}}(E_j) = \beta(n)P(E_j) + [1-\beta(n)]P^{emp}(E_j)$, for all $E_j \in \mathcal{E}$, means that the restriction $P_{\mathcal{E}}|_{\sigma(\mathcal{E})}$ of $P_{\mathcal{E}}$ agrees with the restriction $\breve{P}|_{\sigma(\mathcal{E})}$ of $\breve{P}$ on the sigma algebra $\sigma(\mathcal{E})$ generated by the elements of $\mathcal{E}$. $P_{\mathcal{E}}|_{\sigma(\mathcal{E})}$ is then extended to $(\Omega, \mathcal{F})$ through $P(\cdot \mid E_j)$, for all j, via

$$P_{\mathcal{E}}(A) = \sum_{E_j \in \mathcal{E}} P(A \mid E_j) P_{\mathcal{E}}(E_j).$$

3.5 Subsequent updates

Let us denote the amount of data available at time $t = 1$ by n_1. Once at time $t = 2$ we observe new data points $x_{n_1+1}, \ldots, x_{n_2}$, we update $P_{\mathcal{E}} \equiv P_{\mathcal{E}_1}$ to $P_{\mathcal{E}_1\mathcal{E}_2}$ via the same mechanical procedure depicted in Section 3.4. With this, we mean the following. We now have observed data $x_1, \ldots x_{n_1}, x_{n_1+1}, \ldots, x_{n_2}$. Then, we consider partition $\mathcal{E}_2 =$

$\{E_j\}_{j=1}^{k+1}$, where $E_1, \ldots, E_k$ are the unique elements in the collection $\mathcal{E}'' = \{E_i\}_{i=1}^{n_2} = \{X^{-1}(x_i)\}_{i=1}^{n_2}$, and $E_{k+1} = (\cup_{j=1}^{k} E_k)^c$. We equate $P_{\mathcal{E}_1\mathcal{E}_2}(E_j) = \beta(n_2)P_{\mathcal{E}_1}(E_j) + [1 - \beta(n_2)]P_2^{emp}(E_j)$, for all $E_j \in \mathcal{E}_2$, where the $P_2^{emp}(E_j)$'s are computed similarly to (3.3)–(3.6), so we have

$$P_{\mathcal{E}_1\mathcal{E}_2}(A) = \sum_{E_j \in \mathcal{E}_2} P_{\mathcal{E}_1}(A \mid E_j) P_{\mathcal{E}_1\mathcal{E}_2}(E_j).$$

Clearly, Proposition 21 is true also for $P_{\mathcal{E}_2}$.

Call $(P_{\mathcal{E}_1 \cdots \mathcal{E}_t})$ the sequence of successive updates of probability measure P representing the initial subjective beliefs of the agent around the elements of Ω, and $\mathbf{x}_t = \{x_i\}_{i=1}^{n_t}$ the collection of data points available at time t. Notice that

$$\#\mathcal{E}_t = \#\text{unique}(\mathbf{x}_t) + 1,$$

where # denotes the cardinality operator. That is, the number of elements of partition $\mathcal{E}_t$ is a function of the collected observations up to time t; in particular, it is equal to the number of unique observations x_i plus 1. In the remainder of the paper, for notational convenience, we write $P_{\mathcal{E}_t}$ in place of $P_{\mathcal{E}_1 \cdots \mathcal{E}_t}$, for all $t \in \mathbb{N}$.

Remark 23. Notice that, for all $t \in \mathbb{N}$, $n_t > n_{t-1}$, and $n_0 = 0$. That is, the amount of data points available at time t is always larger than that at time $t-1$; this implies that as $t \to \infty$, then $n_t \to \infty$. In addition, we have that $P_{\mathcal{E}_t}$ depends on $n_1, \ldots, n_t$ and $P_{\mathcal{E}_0}$; we denote this by $P_{\mathcal{E}_t} \equiv P_{\mathcal{E}_t}(n_1, \ldots, n_t, P_0)$. To show this, we write $P_{\mathcal{E}_2}$ in

terms of n_1, n_2, and P_0. We assume that $E_{k+1} \neq \varnothing$, so the following holds

$$P_{\mathcal{E}_2}(A) = \sum_{E \in \mathcal{E}_2} P_{\mathcal{E}_1}(A \mid E) \left[\beta(n_2) P_{\mathcal{E}_1}(E) + (1 - \beta(n_2)) \frac{1}{n_2 + 1} \sum_{s=1}^{n_2} \mathbb{I}(\tilde{E}_s = E) \right]$$

$$= \sum_{E \in \mathcal{E}_2} \left\{ \frac{\sum_{E' \in \mathcal{E}_1} P_{\mathcal{E}_0}(A \cap E \mid E') \left[\beta(n_1) P_{\mathcal{E}_0}(E') + (1 - \beta(n_1)) \frac{1}{n_1+1} \sum_{i=1}^{n_1} \mathbb{I}(\check{E}_i = E') \right]}{\sum_{E' \in \mathcal{E}_1} P_{\mathcal{E}_0}(E \mid E') \left[\beta(n_1) P_{\mathcal{E}_0}(E') + (1 - \beta(n_1)) \frac{1}{n_1+1} \sum_{i=1}^{n_1} \mathbb{I}(\check{E}_i = E') \right]} \right.$$

$$\cdot \left[\beta(n_2) \sum_{E' \in \mathcal{E}_1} P_{\mathcal{E}_0}(E \mid E') \left(\beta(n_1) P_{\mathcal{E}_0}(E') + (1 - \beta(n_1)) \frac{1}{n_1 + 1} \sum_{i=1}^{n_1} \mathbb{I}(\check{E}_i = E') \right) \right.$$

$$\left. \left. + (1 - \beta(n_2)) \frac{1}{n_2 + 1} \sum_{s=1}^{n_2} \mathbb{I}(\tilde{E}_s = E) \right] \right\}, \, \check{E}_i \in \mathcal{E}', \tilde{E}_s \in \mathcal{E}''.$$

It is easy to see how this can be generalized to any $t > 2$. In the remainder of the paper, for notational convenience we write $P_{\mathcal{E}_t}$ in place of $P_{\mathcal{E}_1 \cdots \mathcal{E}_t}(n_1, \ldots, n_t, P_0)$, and $P_{\mathcal{E}_t}(A)$ in place of $P_{\mathcal{E}_1 \cdots \mathcal{E}_t}(n_1, \ldots, n_t, P_0; A)$, for all $A \in \mathcal{F}$.

A consequence of how we build partitions is that, for any t, $\mathcal{E}_t$ is not coarser than $\mathcal{E}_{t-1}$. To see this, suppose $\mathcal{E}_{t-1}$ has $\ell + 1$ many elements, that is, $\mathcal{E}_{t-1} = \{E_1^{\mathcal{E}_{t-1}}, \ldots, E_\ell^{\mathcal{E}_{t-1}}, E_{\ell+1}^{\mathcal{E}_{t-1}}\}$. As we know, this means that $E_{\ell+1}^{\mathcal{E}_{t-1}} = (\cup_{j=1}^{\ell} E_j^{\mathcal{E}_{t-1}})^c$. Now suppose that in the next updating step we only observe one element x. If it is not a "novelty", then $\mathcal{E}_t = \mathcal{E}_{t-1}$. If instead x is a new element, we have that $\mathcal{E}_t$ has $\ell + 2$ many elements. In particular, $E_j^{\mathcal{E}_{t-1}} = E_j^{\mathcal{E}_t}$, for all $j \in \{1, \ldots, \ell\}$, and $E_{\ell+1}^{\mathcal{E}_{t-1}} = E_{\ell+1}^{\mathcal{E}_t} \cup E_{\ell+2}^{\mathcal{E}_t}$. Of course, if we observe more elements, we further refine $E_{\ell+1}^{\mathcal{E}_{t-1}}$.

Proposition 24. There exists a partition $\tilde{\mathcal{E}}$ that cannot be refined as a result of the updating process described in Sections 3.3 and 3.4.

We now show how, under mild standard assumptions, the sequence of successive subjective beliefs updated according to the DPK procedure converges. Call $Q_{\tilde{\mathcal{E}}}$ the restriction of probability measure Q introduced in Section 3.3 to the sigma algebra

$\sigma(\tilde{\mathcal{E}})$ generated by the elements of $\tilde{\mathcal{E}}$. That is, $Q_{\tilde{\mathcal{E}}} := Q|_{\sigma(\tilde{\mathcal{E}})}$, $Q_{\tilde{\mathcal{E}}} : \sigma(\tilde{\mathcal{E}}) \to [0,1]$. Call then $\mathscr{Q}$ the collection of extensions of $Q_{\tilde{\mathcal{E}}}$ from $\sigma(\tilde{\mathcal{E}})$ to $\mathcal{F} = 2^\Omega$. Notice that $\mathscr{Q} \neq \varnothing$ and that $\mathscr{Q}$ is a singleton if and only if $\overline{\mathcal{F}}^{Q_{\tilde{\mathcal{E}}}} = \mathcal{F}$, where

$$\overline{\mathcal{F}}^{Q_{\tilde{\mathcal{E}}}} := \left\{A \in 2^\Omega : Q_{\tilde{\mathcal{E}}\star}(A) = Q^\star_{\tilde{\mathcal{E}}}(A)\right\}$$

is the $Q_{\tilde{\mathcal{E}}}$-completion of $\sigma(\tilde{\mathcal{E}})$, and $Q_{\tilde{\mathcal{E}}\star}$ and $Q^\star_{\tilde{\mathcal{E}}}$ are the inner and outer measures induced by $Q_{\tilde{\mathcal{E}}}$, respectively. Let d_{TV} denote the total variation distance

$$d_{TV}(\pi, \gamma) = \sup_{A \in \mathcal{F}} |\pi(A) - \gamma(A)|,$$

for all $\pi, \gamma \in \Delta(\Omega, \mathcal{F})$.

Theorem 25. If the data points are sampled independently, $\mathbb{E}(X) < \infty$, and $\beta(n_t) = o(1/n_t)$, then $P_{\mathcal{E}_t}$ converges to an element of $\mathscr{Q}$ with probability 1 as $n_t \to \infty$ in the total variation distance.

Because as n_t grows to infinity the partition induced by collection $\{X^{-1}(x_i)\}_{i=1}^{n_t}$ approaches $\tilde{\mathcal{E}}$, we denote by $P_{\tilde{\mathcal{E}}}$ the limit we find in Theorem 25.

Remark 26. If we have an element E of the partition $\mathcal{E}$ that we are considering whose probability is 0, computing the conditional probability of any $A \subset \Omega$ given E is a problem, in that it gives rise to an indeterminate form. Indeed, for a generic probability measure P, we have that $P(A \mid E) = \frac{P(A \cap E)}{P(E)} = \frac{0}{0}$, an indeterminate form. We solve this indeterminacy by requiring that for dynamic probability kinematics this ratio is equal to 0. This because the information conveyed by the data collected on the event A is already encapsulated in all the other elements of the partition.

Remark 27. Dynamic probability kinematics is not commutative. With this we mean the following. Consider an initial probability P and compute its dynamic probability kinematics update $P_{\mathcal{E}_1}$ based on partition $\mathcal{E}_1$; then compute the DPK

update of $P_{\mathcal{E}_1}$ based on partition $\mathcal{E}_2$, and call this update $P_{\mathcal{E}_1\mathcal{E}_2}$. If we proceed in the opposite direction, that is, if we first update P to $P_{\mathcal{E}_2}$, and then update this latter to $P_{\mathcal{E}_2\mathcal{E}_1}$, we have that, in general, $P_{\mathcal{E}_1\mathcal{E}_2} \neq P_{\mathcal{E}_2\mathcal{E}_1}$. To see this, consider the following scenario. Let $\mathcal{E}_1$ be the partition induced by observations $x_1, \ldots, x_{n_1}$, and $\mathcal{E}_2$ the partition induced by observations $x_1, \ldots, x_{n_1}, x_{n_2=n_1+1}$, where $x_{n_2} = x_{n_1}$. This means that $\mathcal{E}_1 = \mathcal{E}_2$, but $P_{\mathcal{E}_1\mathcal{E}_2}(E) \neq P_{\mathcal{E}_2\mathcal{E}_1}(E)$, for all $E \in \mathcal{E}_1 = \mathcal{E}_2$. To illustrate this, let $\#\mathcal{E}_1 = \#\mathcal{E}_2 = m + 1$, $m \leqslant n_1$, assume $E_{m+1} \neq \varnothing$, and notice that

$$P_{\mathcal{E}_1\mathcal{E}_2}(E_j) = \beta(n_1+1)P_{\mathcal{E}_1}(E_j) + \frac{1-\beta(n_1+1)}{n_1+2}\sum_{i=1}^{n_1+1}\mathbb{I}(E_j = X^{-1}(x_i)), \quad j \in \{1, \ldots, m\} \tag{3.8}$$

and

$$P_{\mathcal{E}_1\mathcal{E}_2}(E_{m+1}) = 1 - \sum_{j=1}^{m} P_{\mathcal{E}_1\mathcal{E}_2}(E_j). \tag{3.9}$$

Instead, suppose that we first update according to $\mathcal{E}_2$ and then according to $\mathcal{E}_1$. This may happen if we lose data point x_{n+1}, for example because of a transcription error. Then we have

$$P_{\mathcal{E}_2\mathcal{E}_1}(E_j) = \beta(n_1)P_{\mathcal{E}_1}(E_j) + \frac{1-\beta(n_1)}{n_1+1}\sum_{i=1}^{n_1}\mathbb{I}(E_j = X^{-1}(x_i)), \quad j \in \{1, \ldots, m\} \tag{3.10}$$

and

$$P_{\mathcal{E}_2\mathcal{E}_1}(E_{m+1}) = 1 - \sum_{j=1}^{m} P_{\mathcal{E}_2\mathcal{E}_1}(E_j). \tag{3.11}$$

As we can see, $P_{\mathcal{E}_1\mathcal{E}_2}(E_j) \neq P_{\mathcal{E}_2\mathcal{E}_1}(E_j)$, $j \in \{1, \ldots, m\}$, and also $P_{\mathcal{E}_1\mathcal{E}_2}(E_{m+1}) \neq P_{\mathcal{E}_2\mathcal{E}_1}(E_{m+1})$.

In (Diaconis and Zabell, 1982, Section 3), the authors study when Jeffrey's update is commutative. As we shall see, their results cannot be directly applied to DPK. In (Diaconis and Zabell, 1982, Theorem 3.1), the authors show that, given two generic

partitions $\mathcal{E}$ and $\mathcal{G}$, if

$$P_{\mathcal{E}\mathcal{G}}(E) = P_{\mathcal{E}}(E) \quad \text{and} \quad P_{\mathcal{G}\mathcal{E}}(G) = P_{\mathcal{G}}(G), \tag{3.12}$$

for all $E \in \mathcal{E}$ and all $G \in \mathcal{G}$, then $P_{\mathcal{E}\mathcal{G}} = P_{\mathcal{G}\mathcal{E}}$. We give now a simple counterexample to show that the sufficient condition does not hold for DPK.

Suppose that we observe $x_1 = 1$, $x_2 = x_3 = 5$, $x_4 = 7$, and $x_5 = 8$. They induce partition $\mathcal{E}_1 = \{E_j^{\mathcal{E}_1}\}_{j=1}^5$ whose elements are $E_1^{\mathcal{E}_1} = X^{-1}(1)$, $E_2^{\mathcal{E}_1} = X^{-1}(5)$, $E_3^{\mathcal{E}_1} = X^{-1}(7)$, $E_4^{\mathcal{E}_1} = X^{-1}(8)$, and $E_5^{\mathcal{E}_1} = (\cup_{j=1}^4 E_j^{\mathcal{E}_1})^c$. The empirical probabilities assigned to the elements of $\mathcal{E}_1$ according to (3.3) and (3.4) are $P_1^{emp}(E_j^{\mathcal{E}_1}) = 1/6$, $j \in \{1, 3, 4, 5\}$, and $P_1^{emp}(E_2^{\mathcal{E}_1}) = 1/3$. Now, suppose that we observe a new data point $x_6 = 11$, so that $x_1, \ldots, x_6$ induce a new partition $\mathcal{E}_2 = \{E_j^{\mathcal{E}_2}\}_{j=1}^6$ whose elements are such that $E_j^{\mathcal{E}_2} = E_j^{\mathcal{E}_1}$ for $j \in \{1, \ldots, 4\}$, $E_5^{\mathcal{E}_2} = X^{-1}(11)$, and $E_6^{\mathcal{E}_2} = (\cup_{j=1}^5 E_j^{\mathcal{E}_2})^c$. As we can see, $E_5^{\mathcal{E}_2} \cup E_6^{\mathcal{E}_2} = E_5^{\mathcal{E}_1}$, so $\mathcal{E}_2$ is a refinement of $\mathcal{E}_1$. The empirical probabilities assigned to the elements of $\mathcal{E}_2$ according to (3.3) and (3.4) are $P_2^{emp}(E_j^{\mathcal{E}_2}) = 1/7$, $j \in \{1, 3, 4, 5, 6\}$, and $P_2^{emp}(E_2^{\mathcal{E}_2}) = 2/7$. Then, we have that

$$\begin{aligned} P_{\mathcal{E}_1\mathcal{E}_2}(E_1^{\mathcal{E}_1}) = P_{\mathcal{E}_1\mathcal{E}_2}(E_1^{\mathcal{E}_2}) &= \beta(6) P_{\mathcal{E}_1}(E_1^{\mathcal{E}_1}) + [1 - \beta(6)]1/7 \\ &\neq P_{\mathcal{E}_1}(E_1^{\mathcal{E}_1}) = \beta(5) P(E_1^{\mathcal{E}_1}) + [1 - \beta(5)]1/6, \end{aligned}$$

which does not meet condition (3.12).

In (Diaconis and Zabell, 1982, Theorem 3.2), the authors show that $P_{\mathcal{E}\mathcal{G}} = P_{\mathcal{G}\mathcal{E}}$ if and only if $\mathcal{E}$ and $\mathcal{G}$ are Jeffrey-independent, that is, if and only if $P_{\mathcal{E}}(G) = P(G)$ and $P_{\mathcal{G}}(E) = P(E)$, for all $E \in \mathcal{E}$ and all $G \in \mathcal{G}$. The underlying implicit assumption to this result, though, appears to be the fact that $P(E \cap G) > 0$, for all E and all G. As it is immediate to see, this does not hold in our case, so we cannot use this result to check the commutativity of DPK updates.

Should the lack of commutativity worry the agent that intends to update their beliefs using DPK? The answer is no. Since successive partitions are induced by an

increasing amount of collected data points, commutativity would mean that losing data yields no loss of information on the likelihood of the event $A \subset \Omega$ of interest. This is undesirable: the more we know about the composition of Ω, the better we want our assessment to be on the plausibility of event A. As Diaconis and Zabell point out in (Diaconis and Zabell, 1982, Section 4.2, Remark 2), "noncommutativity is not a real problem for successive Jeffrey updating"; it is not a real problem for DPK either.

Before concluding this Remark, we mention how, despite DPK is not in general commutative, the limit probability $P_{\tilde{\mathcal{E}}}$ is the same regardless of the order in which data is collected. Suppose we collect observations in a different order in two different procedures. Call $(\mathcal{E}_t)$ and $(\mathcal{E}'_t)$ the sequences of successive partitions in the first and second procedures, respectively, and $\tilde{\mathcal{E}}$ and $\tilde{\mathcal{E}}'$ the limit partitions for the first and second procedures, respectively.

Proposition 28. Suppose that data points are sampled independently, $\mathbb{E}(X) < \infty$, and $\beta(n_t) = o(1/n_t)$. Call $P_{\tilde{\mathcal{E}}}$ the almost sure limit of $(P_{\mathcal{E}_t})$ and $P_{\tilde{\mathcal{E}}'}$ the almost sure limit of $(P_{\mathcal{E}'_t})$ in the total variation metric as n_t goes to infinity. Then, $P_{\tilde{\mathcal{E}}} = P_{\tilde{\mathcal{E}}'}$.

3.6 Working with sets of probabilities

In this Section, we generalize dynamic probability kinematics to dynamic imprecise probability kinematics (DIPK). To do so, we first need to introduce the concepts of lower probability, upper probability, and core of a lower probability.

Consider a generic set of probabilities Π on a measurable space $(\Omega, \mathcal{F})$. The lower probability $\underline{P}$ associated with Π is defined as

$$\underline{P}(A) := \inf_{P \in \Pi} P(A), \quad \forall A \in \mathcal{F}.$$

The upper probability $\overline{P}$ associated with Π is defined as the conjugate to $\underline{P}$, that is,

$$\overline{P}(A) := 1 - \underline{P}(A^c) = \sup_{P' \in \Pi} P'(A), \quad \forall A \in \mathcal{F}.$$

Let us denote by $\Delta(\Omega, \mathcal{F})$ the set of all probability measures on $(\Omega, \mathcal{F})$. Lower probability $\underline{P}$ completely characterizes the convex set

$$\begin{aligned} \text{core}(\underline{P}) &:= \{P \in \Delta(\Omega, \mathcal{F}) : P(A) \geqslant \underline{P}(A), \forall A \in \mathcal{F}\} \\ &= \{P \in \Delta(\Omega, \mathcal{F}) : \overline{P}(A) \geqslant P(A) \geqslant \underline{P}(A), \forall A \in \mathcal{F}\}, \end{aligned}$$

where the second equality is a characterization (Cerreia-Vioglio et al., 2015, Page 3389). Notice that the core is convex (Marinacci and Montrucchio, 2004a, Section 2.2) and weak*-compact (Marinacci and Montrucchio, 2004a, Proposition 3). Recall that in the weak* topology, a net $(P_\alpha)_{\alpha \in I}$ converges to P if and only if $P_\alpha(A) \to P(A)$, for all $A \in \mathcal{F}$.

By complete characterization, we mean that it is sufficient to know $\underline{P}$ to be able to completely specify core($\underline{P}$). To emphasize this aspect, some authors say that $\underline{P}$ is *compatible* with core($\underline{P}$) (Gong and Meng, 2021).

To generalize DPK to DIPK, we first prescribe the agent to specify a set of probabilities $\mathcal{P}$, then to compute the lower probability associated with it. The core of such lower probability represents the agent's initial beliefs. To update their beliefs, the agent computes the DPK update of the extrema of the core, that is, of the elements of the core that cannot be written as a convex combination of other elements. Their updated beliefs are represented by the convex hull of the updated extrema, which coincides with the core of the updated lower probability by Theorem 6.

We require the agent's beliefs to be represented by the core for two main reasons. The first, mathematical, one is to ensure that the belief set can be completely characterized by the lower probability. The second, philosophical, one is to further hedge against ambiguity. Indeed, it is immediate to see that $\mathcal{P}$ is contained in the core, so

requiring this latter to represent the agent's beliefs allows them to take into account a wider range of plausible probabilities.[2]

Remark 29. The way the agent specifies the set of probability measures representing their initial beliefs is equivalent to them performing what is known as *sensitivity analysis*. That is, at the beginning of the analysis, the agent specifies a set of possible (or plausible) candidates for the true or ideal probability measure P_T governing the events of interest (Berger, 1984). As (Walley, 1991, Section 5.9) points out, this way of proceeding assumes the *axiom of ideal precision*: there exists a true probability measure P_T governing the random events, but it cannot be precisely known e.g. because we would need an infinitely long reflection to elicit it. Nevertheless, the sensitivity analysis approach is mathematically equivalent to:

1. assessing the lower probability $\underline{P}'$ of some events $\{A_k\} \subset \mathcal{F}$ of interest, defined as the supremum price the agent is willing to pay to enter a bet that pays \$1 if event A_k takes place, and \$0 otherwise,[3]

2. considering its natural extension $\underline{P}$ to all the elements in $\mathcal{F}$ (Walley, 1991, Section 3.1),

3. considering the set of regular probability measures that setwise dominate lower probability $\underline{P}$.

For this reason, we deem the sensitivity analysis approach satisfactory.

[2] Notice that from an information theoretic perspective, performing an analysis starting with set $\mathcal{P}$ or with the core of the lower probability associated with $\mathcal{P}$ are two different problems. In particular, the second entails a higher initial degree of ignorance of the agent.

[3] In the imprecise probabilities literature, agents are often required to specify coherent lower (and upper) probabilities (Walley, 1991, Chapter 2). A way of doing so is to require that no Dutch books can be made against the punter (Caprio and Mukherjee, 2021a, Definition 5.1), i.e. that we cannot find a finite collection $\{B_j\}_{j=1}^n \subset \mathcal{F}$ along with numbers $\{s_j\}_{j=1}^n \subset \mathbb{R}$ such that, for all $\omega \in \Omega$, $\sum_{j=1}^n s_j \left[\mathbb{1}_{B_j}(\omega) - \underline{P}(B_j)\right] < 0$. In (Walley, 1991, Section 3.3.3) the author shows that $\underline{P}$ is coherent if and only if it can be written as the infimum of a set $\mathcal{P}$ of regular probability measures.

In addition, it is worth to notice that lower probabilities are a special case of *lower previsions*. To define these latter, we need to first introduce the concept of *gambles*. A gamble Y is a bounded real-valued function on Ω which is interpreted as an uncertain reward. The set of all gambles on Ω is denoted by $\mathcal{L}(\Omega)$. Call now $\mathcal{K}$ an arbitrary subset of $\mathcal{L}(\Omega)$; a lower prevision $\underline{P}$ is a real-valued function defined on $\mathcal{K}$ such that, for all $K \in \mathcal{K}$, $\underline{P}(K)$ is the supremum price μ for which it is asserted that the gamble $X - \mu$ is desirable to the agent (Walley, 1991, Section 2.3.1). Consider now a generic event $A \in \mathcal{F}$, and call $\mathcal{A} := \{\mathbb{I}_A(\omega) : \omega \in \Omega,\, A \in \mathcal{F}\}$ the collection of indicator functions of events $A \in \mathcal{F}$. We can see how an indicator function is just a $0-1$ valued gamble, and so lower probabilities can be seen as lower previsions defined on $\mathcal{A} \subset \mathcal{L}(\Omega)$ (Walley, 1991, Section 2.7.2). In this work we focus on lower probabilities because they are more immediately related to regular (additive) probabilities, and because they are easier to derive from a set of probability measures. In the future, we will generalize DIPK to deal with lower previsions.

We start our analysis by specifying a set $\mathcal{P} \subset \Delta(\Omega, \mathcal{F})$ of probability measures on Ω. We then consider $\underline{P} \equiv \underline{P}_{\mathcal{E}_0}$, the lower probability associated with $\mathcal{P}$. The set representing the agent's initial beliefs is given by $\mathcal{P}^{\text{co}}_{\mathcal{E}_0} = \text{core}(\underline{P}_{\mathcal{E}_0})$, where superscript co denotes the fact that $\mathcal{P}^{\text{co}}_{\mathcal{E}_0}$ is convex and compact. The importance of this properties is explained in Remark 31. We also need to consider the set $\mathcal{P}_{\mathcal{E}_0} = \text{ex}\mathcal{P}^{\text{co}}_{\mathcal{E}_0}$ of extrema of $\mathcal{P}^{\text{co}}_{\mathcal{E}_0}$. Of course, $\mathcal{P}^{\text{co}}_{\mathcal{E}_0} = \text{Conv}(\mathcal{P}_{\mathcal{E}_0})$.

We then compute the DPK update of every element in $\mathcal{P}_{\mathcal{E}_0}$, and we obtain

$$\mathcal{P}_{\mathcal{E}_1} := \left\{ P_{\mathcal{E}_1} \in \Delta(\Omega, \mathcal{F}) : P_{\mathcal{E}_1}(A) = \sum_{E_j \in \mathcal{E}_1} P_{\mathcal{E}_0}(A \mid E_j) P_{\mathcal{E}_1}(E_j),\ \forall A \in \mathcal{F}, P_{\mathcal{E}_0} \in \mathcal{P}_{\mathcal{E}_0} \right\}.$$

After that, we compute $\mathcal{P}^{\text{co}}_{\mathcal{E}_1} = \text{Conv}(\mathcal{P}_{\mathcal{E}_1}) = \text{core}(\underline{P}_{\mathcal{E}_1})$, where $\underline{P}_{\mathcal{E}_1}$ is the updated lower probability, and the last equality holds by Theorem 6.

Repeating this procedure, we build two sequences, $(\mathcal{P}_{\mathcal{E}_t})$ and $(\mathcal{P}^{\text{co}}_{\mathcal{E}_t})$. Notice that for any $t \in \mathbb{N}$, the lower and upper probabilities associated with $\mathcal{P}_{\mathcal{E}_t}$ are equal to the lower and upper probabilities associated with $\mathcal{P}^{\text{co}}_{\mathcal{E}_t}$, respectively. Recall that d_{TV} denotes the total variation distance

$$d_{TV}(\pi, \gamma) = \sup_{A \in \mathcal{F}} |\pi(A) - \gamma(A)|,$$

for all $\pi, \gamma \in \Delta(\Omega, \mathcal{F})$. Suppose that the data points $x_1, \ldots, x_{n_t}$ that induce partition $\mathcal{E}_t$ are sampled independently, $\mathbb{E}(X) < \infty$, and $\beta(n_t) = o(1/n_t)$. Call

$$\mathcal{P}_{\tilde{\mathcal{E}}} := \left\{ P_{\tilde{\mathcal{E}}} \in \Delta(\Omega, \mathcal{F}) : d_{TV}(P_{\mathcal{E}_t}, P_{\tilde{\mathcal{E}}}) \xrightarrow[n_t \to \infty]{a.s.} 0,\ P_{\mathcal{E}_t} \in \mathcal{P}_{\mathcal{E}_t} \right\}.$$

That is, $\mathcal{P}_{\tilde{\mathcal{E}}}$ is the set of limits (as n_t goes to infinity with probability 1 in the total variation metric) of the elements $P_{\mathcal{E}_t}$ of set $\mathcal{P}_{\mathcal{E}_t}$ representing the (extrema of the) agent's updated beliefs. We are sure $\mathcal{P}_{\tilde{\mathcal{E}}}$ is not empty by Propositions 24 and 25. Then, by construction, we have that

$$d_H(\mathcal{P}_{\mathcal{E}_t}, \mathcal{P}_{\tilde{\mathcal{E}}}) = \max \left(\sup_{P \in \mathcal{P}_{\mathcal{E}_t}} d_{TV}(P, \mathcal{P}_{\tilde{\mathcal{E}}}), \sup_{P' \in \mathcal{P}_{\tilde{\mathcal{E}}}} d_{TV}(\mathcal{P}_{\mathcal{E}_t}, P') \right) \to 0 \tag{3.13}$$

as n_t goes to infinity with probability 1, where d_H denotes the Hausdorff metric, and, in general, $d_{TV}(\pi, \Gamma) := \inf_{\gamma \in \Gamma} d_{TV}(\pi, \gamma)$, for all $\pi \in \Delta(\Omega, \mathcal{F})$ and all $\Gamma \subset \Delta(\Omega, \mathcal{F})$. Such a convergence is true also for $\mathcal{P}^{\text{co}}_{\mathcal{E}_t}$ and $\mathcal{P}^{\text{co}}_{\tilde{\mathcal{E}}}$.

Proposition 30. If the data points are sampled independently, $\mathbb{E}(X) < \infty$, and $\beta(n_t) = o(1/n_t)$, then the following is true with probability 1

$$d_H(\mathcal{P}^{\text{co}}_{\mathcal{E}_t}, \mathcal{P}^{\text{co}}_{\tilde{\mathcal{E}}}) \to 0$$

as n_t go to infinity.

Remark 31. Let us discuss the importance of $\mathcal{P}^{\text{co}}_{\mathcal{E}_t}$ being convex and compact. Consider a generic set of probabilities Π on a measurable space $(\Omega, \mathcal{F})$. Suppose Π is finite, i.e. $\Pi = \{\pi_j\}_{j=1}^k$, for some $k \in \mathbb{N}$. Then, the lower probability associated with Π is equivalent to the one associated with its convex hull $\text{Conv}(\Pi)$. If instead Π is convex but open, then the lower probability associated with Π is equivalent to the one associated with its closure $\text{Cl}(\Pi)$. To this extent, lower probabilities are not able to detect "holes and dents" in their associated set of probabilities. This is why we need the sequence of convex and (weak*-)compact sets ($\mathcal{P}^{\text{co}}_{\mathcal{E}_t}$) to represent the agent's belief updating procedure.

Remark 32. A natural question the reader may ask is why do we need the core to represent the agent beliefs. Indeed, it would be easier to require the agent to specify a finite set of plausible probability measures, and then let the convex hull of such finite set represent their initial beliefs.[4] The answer is because the lower probability completely characterizes the core, but does not completely characterize the convex hull. This because in general the convex hull of a finite set of probabilities is a *proper subset* of the core of the lower probability associated with that set (Amarante and Maccheroni, 2006, Example 1) and (Amarante et al., 2006, Examples 6,7,8). This means that when studying the DIPK update from $\mathcal{P}^{\text{co}}_{\mathcal{E}_t}$ to $\mathcal{P}^{\text{co}}_{\mathcal{E}_{t+1}}$ we can just update the lower probability $\underline{P}_{\mathcal{E}_t}$ to $\underline{P}_{\mathcal{E}_{t+1}}$ to be able to specify the whole $\mathcal{P}^{\text{co}}_{\mathcal{E}_{t+1}}$. This would not be the case had we not represented the agent's beliefs via the core of a lower probability.

Remark 33. Notice how in the case of DIPK $\beta(n_t)$ captures the attitude of the agent towards ambiguity: if it is a constant, then ambiguity never fades, while if it converges to 0 as n_t goes to infinity, ambiguity is resolved as more and more data is

[4] Notice that the convex hull is both convex and weak*-compact. Compactness comes from it being the convex hull of a finite set in a Banach space (the normed vector space induced by $(\Delta(\Omega, \mathcal{F}), d_{TV})$ is complete because d_{TV} is a complete metric; notice also that $\|\cdot\|_{TV}$-compact implies weak*-compact by the definition of weak*-compactness).

collected. The speed of convergence is subjectively determined; if it is $o(1/n_t)$, we retrieve the convergence result in Proposition 30.

3.7 Procedures to obtain and bound upper and lower probabilities

As we have seen in Remark 32, the lower probability associated with $\mathcal{P}^{\text{co}}_{\mathcal{E}_t}$ encodes all the information contained in the set. It is natural, then, that we focus our attention on $\underline{P}_{\mathcal{E}_t}$. In this Section, given a generic $t \in \mathbb{N}$, we derive bounds for $\underline{P}_{\mathcal{E}_{t+1}}$ that can be computed without performing the DIPK updated of $\mathcal{P}^{\text{co}}_{\mathcal{E}_t}$. They are interesting in their own right, and will be put to use in Section 3.8 to study the behavior of set $\mathcal{P}^{\text{co}}_{\mathcal{E}_t}$ with respect to set $\mathcal{P}^{\text{co}}_{\mathcal{E}_{t-1}}$.

For any $A \in \mathcal{F}$, and any element E of a generic partition $\mathcal{E}$, define

$$\underline{P}^{B}_{\mathcal{E}_t}(A \mid E) := \inf_{P \in \mathcal{P}^{\text{co}}_{\mathcal{E}_t}} P(A \mid E) = \inf_{P \in \mathcal{P}^{\text{co}}_{\mathcal{E}_t}} \frac{P(A \cap E)}{P(E)}$$

and

$$\overline{P}^{B}_{\mathcal{E}_t}(A \mid E) := \sup_{P \in \mathcal{P}^{\text{co}}_{\mathcal{E}_t}} P(A \mid E) = \sup_{P \in \mathcal{P}^{\text{co}}_{\mathcal{E}_t}} \frac{P(A \cap E)}{P(E)}.$$

These are called the generalized Bayes conditional lower and upper probabilities (Wasserman and Kadane, 1990), respectively. We have the following.

Proposition 34. For any $A \in \mathcal{F}$ and any $t \in \mathbb{N}$,

$$\underline{P}_{\mathcal{E}_{t+1}}(A) \geqslant \sum_{E_j \in \mathcal{E}_{t+1}} \underline{P}^{B}_{\mathcal{E}_t}(A \mid E_j) \left[\beta(n_t)\underline{P}_{\mathcal{E}_t}(E_j) + (1-\beta(n_t))P^{emp}_{t+1}(E_j)\right] \tag{3.14}$$

and

$$\overline{P}_{\mathcal{E}_{t+1}}(A) \leqslant \sum_{E_j \in \mathcal{E}_{t+1}} \overline{P}^{B}_{\mathcal{E}_t}(A \mid E_j) \left[\beta(n_t)\overline{P}_{\mathcal{E}_t}(E_j) + (1-\beta(n_t))P^{emp}_{t+1}(E_j)\right]. \tag{3.15}$$

Corollary 35. For all $t \in \mathbb{N}$, all $P_{\mathcal{E}_{t+1}} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_{t+1}}$, and all $A \in \mathcal{F}$,

$$P_{\mathcal{E}_{t+1}}(A) \in \Bigg[\sum_{E_j \in \mathcal{E}_{t+1}} \underline{P}^B_{\mathcal{E}_t}(A \mid E_j) \left[\beta(n_t)\underline{P}_{\mathcal{E}_t}(E_j) + (1-\beta(n_t))P^{emp}_{t+1}(E_j)\right], \\ \sum_{E_j \in \mathcal{E}_{t+1}} \overline{P}^B_{\mathcal{E}_t}(A \mid E_j) \left[\beta(n_t)\overline{P}_{\mathcal{E}_t}(E_j) + (1-\beta(n_t))P^{emp}_{t+1}(E_j)\right] \Bigg].$$

There are two other ways to define lower and upper conditional probabilities. The first one, called geometric update, is such that for any $A \in \mathcal{F}$, and any element E of a generic partition $\mathcal{E}$,

$$\underline{P}^G_{\mathcal{E}_t}(A \mid E) := \frac{\inf_{P \in \mathcal{P}^{\text{co}}_{\mathcal{E}_t}} P(A \cap E)}{\inf_{P \in \mathcal{P}^{\text{co}}_{\mathcal{E}_t}} P(E)} = \frac{\underline{P}_{\mathcal{E}_t}(A \cap E)}{\underline{P}_{\mathcal{E}_t}(E)} \quad \text{and} \quad \overline{P}^G_{\mathcal{E}_t}(A \mid E) = \frac{\overline{P}_{\mathcal{E}_t}(A \cap E)}{\overline{P}_{\mathcal{E}_t}(E)}.$$

The other one, called Dempster's rule of conditioning, is the natural dual to the geometric procedure. It differs from this latter from the operational point of view (Gong and Meng, 2021, Section 2), but since mathematically they are the same, we are not going to cover Dempster's rule in the present work.

An interpretation of how generalized Bayes and geometric rules come about when a generic partition $\{E_j\}$ of Ω is available is the following. Let $\sqcup$ denote the union of disjoint sets, and $\underline{P}$ a generic lower probability. We know that lower probabilities are superadditive, so since given any $A \in \mathcal{F}$ we have that $A = \sqcup_j (A \cap E_j)$, it follows that

$$\underline{P}(A) \geqslant \sum_j \underline{P}(A \cap E_j). \tag{3.16}$$

Now, $\underline{P}(A \cap E_j)$ can be interpreted as the lowest possible probability attached to event $A \cap E_j$, in which case we retrieve generalized Bayes rule. It can also be rewritten as $\frac{\underline{P}(A \cap E_j)}{\underline{P}(E_j)} \underline{P}(E_j)$; in this latter case, we retrieve the geometric rule. It is worth noting that, for any lower probability $\underline{P}$, by (Gong and Meng, 2021, Lemma 5.3) we have that

$$\underline{P}^B(A \mid B) \leqslant \underline{P}^G(A \mid B) \leqslant \overline{P}^G(A \mid B) \leqslant \overline{P}^B(A \mid B), \tag{3.17}$$

for all $A, B \in \mathcal{F}$.

3.7.1 Geometric rule

As we have seen in Proposition 34, generalized Bayes comes naturally from our updating procedure. This because, as shown in Section 3.1.2, Jeffrey's rule is a generalization of Bayesian conditioning. Given the inequalities in (3.17), we can sharpen the bounds we found using generalized Bayes rule by using the geometric rule.

Notice that, by Remark 26, we have that, for all $A \in \mathcal{F}$ and all $t \in \mathbb{N}$,

$$\sum_{E_j \in \mathcal{E}_{t+1}} P_{\mathcal{E}_t}(A \mid E_j) P_{\mathcal{E}_{t+1}}(E_j) = \sum_{E_j \in \mathcal{E}_{t+1} : P_{\mathcal{E}_t}(E_j) \neq 0} P_{\mathcal{E}_t}(A \mid E_j) P_{\mathcal{E}_{t+1}}(E_j)$$

since we assumed that if $P_{\mathcal{E}_t}(E_j) = 0$ for some $E_j \in \mathcal{E}_{t+1}$, then $P_{\mathcal{E}_t}(A \mid E_j) = 0$.

Proposition 36. For any $A \in \mathcal{F}$ and any $t \in \mathbb{N}$,

$$\underline{P}_{\mathcal{E}_{t+1}}(A) \geqslant \sum_{E_j \in \mathcal{E}_{t+1} : P_{\mathcal{E}_t}(E_j) \neq 0} \underline{P}^G_{\mathcal{E}_t}(A \mid E_j) \left[\beta(n_t) \underline{P}_{\mathcal{E}_t}(E_j) + (1 - \beta(n_t)) P^{emp}_{t+1}(E_j) \right] \tag{3.18}$$

and

$$\overline{P}_{\mathcal{E}_{t+1}}(A) \leqslant \sum_{E_j \in \mathcal{E}_{t+1} : P_{\mathcal{E}_t}(E_j) \neq 0} \overline{P}^G_{\mathcal{E}_t}(A \mid E_j) \left[\beta(n_t) \overline{P}_{\mathcal{E}_t}(E_j) + (1 - \beta(n_t)) P^{emp}_{t+1}(E_j) \right]. \tag{3.19}$$

Corollary 37. For all $t \in \mathbb{N}$, all $P_{\mathcal{E}_{t+1}} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_{t+1}}$, and all $A \in \mathcal{F}$,

$$P_{\mathcal{E}_{t+1}}(A) \in \Bigg[\sum_{E_j \in \mathcal{E}_{t+1} : P_{\mathcal{E}_t}(E_j) \neq 0} \underline{P}^G_{\mathcal{E}_t}(A \mid E_j) \left[\beta(n_t) \underline{P}_{\mathcal{E}_t}(E_j) + (1 - \beta(n_t)) P^{emp}_{t+1}(E_j) \right], \sum_{E_j \in \mathcal{E}_{t+1} : P_{\mathcal{E}_t}(E_j) \neq 0} \overline{P}^G_{\mathcal{E}_t}(A \mid E_j) \left[\beta(n_t) \overline{P}_{\mathcal{E}_t}(E_j) + (1 - \beta(n_t)) P^{emp}_{t+1}(E_j) \right] \Bigg]. \tag{3.20}$$

In addition,

$$\sum_{E_j \in \mathcal{E}_{t+1}: P_{\mathcal{E}_t}(E_j) \neq 0} \underline{P}^G_{\mathcal{E}_t}(A \mid E_j) \left[\beta(n_t)\underline{P}_{\mathcal{E}_t}(E_j) + (1-\beta(n_t))P^{emp}_{t+1}(E_j)\right]$$

$$\geqslant \sum_{E_j \in \mathcal{E}_{t+1}} \underline{P}^B_{\mathcal{E}_t}(A \mid E_j) \left[\beta(n_t)\underline{P}_{\mathcal{E}_t}(E_j) + (1-\beta(n_t))P^{emp}_{t+1}(E_j)\right] \tag{3.21}$$

and

$$\sum_{E_j \in \mathcal{E}_{t+1}: P_{\mathcal{E}_t}(E_j) \neq 0} \overline{P}^G_{\mathcal{E}_t}(A \mid E_j) \left[\beta(n_t)\overline{P}_{\mathcal{E}_t}(E_j) + (1-\beta(n_t))P^{emp}_{t+1}(E_j)\right]$$

$$\leqslant \sum_{E_j \in \mathcal{E}_{t+1}} \overline{P}^B_{\mathcal{E}_t}(A \mid E_j) \left[\beta(n_t)\overline{P}_{\mathcal{E}_t}(E_j) + (1-\beta(n_t))P^{emp}_{t+1}(E_j)\right]. \tag{3.22}$$

We need to explicitly require $P_{\mathcal{E}_t}(E_j) \neq 0$ for a technical detail in the proof of Proposition 36. Corollary 37 implies that

$$\begin{aligned}
&\Bigg[\sum_{E_j \in \mathcal{E}_{t+1}} \underline{P}^G_{\mathcal{E}_t}(A \mid E_j) \left[\beta(n_t)\underline{P}_{\mathcal{E}_t}(E_j) + (1-\beta(n_t))P^{emp}_{t+1}(E_j)\right], \\
&\quad \sum_{E_j \in \mathcal{E}_{t+1}} \overline{P}^G_{\mathcal{E}_t}(A \mid E_j) \left[\beta(n_t)\underline{P}_{\mathcal{E}_t}(E_j) + (1-\beta(n_t))P^{emp}_{t+1}(E_j)\right] \Bigg] \\
\subset &\Bigg[\sum_{E_j \in \mathcal{E}_{t+1}} \underline{P}^B_{\mathcal{E}_t}(A \mid E_j) \left[\beta(n_t)\underline{P}_{\mathcal{E}_t}(E_j) + (1-\beta(n_t))P^{emp}_{t+1}(E_j)\right], \\
&\quad \sum_{E_j \in \mathcal{E}_{t+1}} \overline{P}^B_{\mathcal{E}_t}(A \mid E_j) \left[\beta(n_t)\underline{P}_{\mathcal{E}_t}(E_j) + (1-\beta(n_t))P^{emp}_{t+1}(E_j)\right] \Bigg],
\end{aligned}$$

so we retrieve tighter bounds for $\underline{P}_{\mathcal{E}_{t+1}}(A)$ and $\overline{P}_{\mathcal{E}_{t+1}}(A)$, and also obtain a tighter interval around $P_{\mathcal{E}_{t+1}}(A)$, for all $A \in \mathcal{F}$ and all $t \in \mathbb{N}$.

3.8 Behavior of updated sets of probabilities

In the imprecise probabilities literature, three concepts are crucial regarding the behavior of updated sets of probabilities. They are contraction, dilation, and sure loss. In this Section, building on the definitions in (Gong and Meng, 2021, Section 3), we introduce the concepts of DIPK-contraction, DIPK-dilation and DIPK-sure loss, and we give sufficient conditions for them to take place.

Fix some $t \in \mathbb{N}$. We say that $\mathcal{P}^{\text{co}}_{\mathcal{E}_t}$ *DIPK-contracts with respect to* $\mathcal{P}^{co}_{\mathcal{E}_{t-1}}$ *for some* $A \in \mathcal{F}$ if $\underline{P}_{\mathcal{E}_t}(A) \geqslant \underline{P}_{\mathcal{E}_{t-1}}(A)$ and $\overline{P}_{\mathcal{E}_t}(A) \leqslant \overline{P}_{\mathcal{E}_{t-1}}(A)$. It *strictly DIPK-contracts* if the inequalities are strict. In addition, we say that sequence $(\mathcal{P}^{\text{co}}_{\mathcal{E}_t})$ *DIPK-contracts for some* $A \in \mathcal{F}$ if for any $t \in \mathbb{N}$, we have that $\underline{P}_{\mathcal{E}_t}(A) \geqslant \underline{P}_{\mathcal{E}_{t-1}}(A)$, $\overline{P}_{\mathcal{E}_t}(A) \leqslant \overline{P}_{\mathcal{E}_{t-1}}(A)$, and the inequalities are strict for some t.

DIPK-dilation is defined analogously, by inverting the inequality signs.

Finally, we say that $\mathcal{P}^{\text{co}}_{\mathcal{E}_t}$ exhibits *DIPK-sure loss with respect to* $\mathcal{P}^{co}_{\mathcal{E}_{t-1}}$ *for some* $A \in \mathcal{F}$ if $\underline{P}_{\mathcal{E}_t}(A) > \overline{P}_{\mathcal{E}_{t-1}}(A)$ or $\overline{P}_{\mathcal{E}_t}(A) < \underline{P}_{\mathcal{E}_{t-1}}(A)$.

Proposition 38. For any $t \in \mathbb{N}$, sufficient conditions for $\mathcal{P}^{\text{co}}_{\mathcal{E}_t}$ to DIPK-contract with respect to $\mathcal{P}^{\text{co}}_{\mathcal{E}_{t-1}}$ for some $A \in \mathcal{F}$ are the following

$$\sum_{E_j \in \mathcal{E}_t} \underline{P}^B_{\mathcal{E}_{t-1}}(A \mid E_j) \left[\beta(n_t)\underline{P}_{\mathcal{E}_{t-1}}(E_j) + (1-\beta(n_t))P^{emp}_t(E_j)\right] \geqslant \underline{P}_{\mathcal{E}_{t-1}}(A)$$

and

$$\sum_{E_j \in \mathcal{E}_t} \overline{P}^B_{\mathcal{E}_{t-1}}(A \mid E_j) \left[\beta(n_t)\overline{P}_{\mathcal{E}_{t-1}}(E_j) + (1-\beta(n_t))P^{emp}_t(E_j)\right] \leqslant \overline{P}_{\mathcal{E}_{t-1}}(A).$$

Notice that we obtain strict DIPK-contraction if the inequalities are strict. We have the same results if we use geometric lower conditional probabilities instead of the generalized Bayes ones. We also have the following.

Proposition 39. For any $t \in \mathbb{N}$, sufficient conditions for $\mathcal{P}^{\text{co}}_{\mathcal{E}_t}$ to exhibit DIPK-sure loss with respect to $\mathcal{P}^{\text{co}}_{\mathcal{E}_{t-1}}$ for some $A \in \mathcal{F}$ are the following

$$\sum_{E_j \in \mathcal{E}_t} \underline{P}^B_{\mathcal{E}_{t-1}}(A \mid E_j) \left[\beta(n_t)\underline{P}_{\mathcal{E}_{t-1}}(E_j) + (1-\beta(n_t))P_t^{emp}(E_j)\right] > \overline{P}_{\mathcal{E}_{t-1}}(A)$$

or

$$\sum_{E_j \in \mathcal{E}_t} \overline{P}^B_{\mathcal{E}_{t-1}}(A \mid E_j) \left[\beta(n_t)\overline{P}_{\mathcal{E}_{t-1}}(E_j) + (1-\beta(n_t))P_t^{emp}(E_j)\right] < \underline{P}_{\mathcal{E}_{t-1}}(A).$$

Again, we obtain the same conditions if we use geometric lower conditional probabilities instead of the generalized Bayes ones.

Giving a sufficient condition for DIPK-dilation without directly computing lower and upper probabilities $\underline{P}_{\mathcal{E}_t}(A)$ and $\overline{P}_{\mathcal{E}_t}(A)$ is less straightforward. We have the following.

Proposition 40. For any $t \in \mathbb{N}$ and some $A \in \mathcal{F}$, if there exist $P_{s,\mathcal{E}_t}, P_{k,\mathcal{E}_t} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_t}$ such that $\underline{P}_{\mathcal{E}_{t-1}}(A) \geqslant P_{s,\mathcal{E}_t}(A)$ and $\overline{P}_{\mathcal{E}_{t-1}}(A) \leqslant P_{k,\mathcal{E}_t}(A)$, then $\mathcal{P}^{\text{co}}_{\mathcal{E}_t}$ DIPK-dilates with respect to $\mathcal{P}^{\text{co}}_{\mathcal{E}_{t-1}}$ for A.

We obtain strict DIPK-dilation if the inequalities in Proposition 40 are strict. As we can see, we do not need to directly compute $\underline{P}_{\mathcal{E}_t}(A)$ and $\overline{P}_{\mathcal{E}_t}(A)$. We only need to find $P_{s,\mathcal{E}_{t-1}}, P_{k,\mathcal{E}_{t-1}} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_{t-1}}$ such that their updates satisfy the assumptions in Proposition 40.

We can give a result, similar to Proposition 40 that provides sufficient conditions for $\mathcal{P}^{\text{co}}_{\mathcal{E}_t}$ to DIPK-contract with respect to $\mathcal{P}^{\text{co}}_{\mathcal{E}_{t-1}}$ for some $A \in \mathcal{F}$. This is interesting because, contrary to what we have in Proposition 38, we do not use the notions of lower and upper conditional probabilities. Its downside is that it requires the computation of both $\underline{P}_{\mathcal{E}_t}(A)$ and $\overline{P}_{\mathcal{E}_t}(A)$.

Proposition 41. For any $t \in \mathbb{N}$ and some $A \in \mathcal{F}$, if there exist $P_{s,\mathcal{E}_{t-1}}, P_{k,\mathcal{E}_{t-1}} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_{t-1}}$ such that $\underline{P}_{\mathcal{E}_t}(A) \geqslant P_{k,\mathcal{E}_{t-1}}(A)$ and $\overline{P}_{\mathcal{E}_t}(A) \leqslant P_{s,\mathcal{E}_{t-1}}(A)$, then $\mathcal{P}^{\text{co}}_{\mathcal{E}_t}$ DIPK-contracts with respect to $\mathcal{P}^{\text{co}}_{\mathcal{E}_{t-1}}$ for A.

We obtain strict DIPK-contraction if the inequalities in Proposition 41 are strict. Notice that we cannot give a result similar to Propositions 40 and 41 for DIPK-sure loss because we cannot require any assumption on any $P_{\mathcal{E}_{t-1}} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_{t-1}}$, $P_{\mathcal{E}_t} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_t}$ to make them "sit in between" $\underline{P}_{\mathcal{E}_t}(A)$ and $\overline{P}_{\mathcal{E}_{t-1}}(A)$, or $\underline{P}_{\mathcal{E}_{t-1}}(A)$ and $\overline{P}_{\mathcal{E}_t}(A)$.

3.9 Two simple examples of DPK and DIPK updating

In this Section, we present two examples on how to update subjective beliefs according to DPK and DIPK procedures.

3.9.1 Trials of a new surgical procedure

We continue Example 20, and show how to frame it within the DPK paradigm. The problem is the following: three trials of a new surgical procedure are to be conducted at a hospital. Let 1 denote a successful outcome, and 0 an unsuccessful one. The state space has the form $\Omega = \{000, 001, 010, 011, 100, 101, 110, 111\}$. A colleague informs us that another hospital performed this type of procedure with a success rate of 0.8. To update our beliefs according to the new piece of information, we consider random variable $X : \Omega \to \mathcal{X} = \{0, 1, 2, 3\}$ whose distribution is unknown and such that $X(\omega)$ represents the number of 1's in ω. As we can see,

$$X^{-1}(3) = \{111\}, \quad X^{-1}(2) = \{011, 101, 110\},$$
$$X^{-1}(1) = \{001, 010, 100\}, \quad X^{-1}(0) = \{000\}.$$

The finest partition of Ω according to DPK, then, is given by $\tilde{\mathcal{E}} = \{E_0, E_1, E_2, E_3, E_4\}$, where $E_j = X^{-1}(j)$, $j \in \{0, 1, 2, 3\}$, and $E_4 = \varnothing$. If we assume the trials are independent, the information that our colleague provided us is equivalent to observ-

ing 1000 data points $x_1, \ldots, x_{1000}$, out of which 512 are all 3's, 384 are all 2's, 96 are all 1's, and 8 are all 0's. This because the relative frequency Fr of the elements of $\mathcal{X}$ is $Fr(\{3\}) = 512/1000 = 1 \cdot 0.8^3$, $Fr(\{2\}) = 384/1000 = 3 \cdot 0.2 \cdot 0.8^2$, $Fr(\{1\}) = 96/1000 = 3 \cdot 0.2^2 \cdot 0.8$, and $Fr(\{0\}) = 8/1000 = 1 \cdot 0.2^3$. But why should they be derived in this way? We have that $Fr(\{3\}) = 1 \cdot 0.8^3$ because there is only 1 way of obtaining three successes, each of which has probability 0.8 in the procedures conducted at the hospital that our colleague informed us about. Instead, $Fr(\{2\}) = 3 \cdot 0.2 \cdot 0.8^2$ because there are 3 ways of obtaining two successes and one failure, where the probability of the latter is 0.2 according to our colleague. Finally, $Fr(\{1\}) = 3 \cdot 0.2^2 \cdot 0.8$ because there are 3 ways of obtaining one successes and two failures, and $Fr(\{0\}) = 1 \cdot 0.2^3$ because there is only 1 way of obtaining three failures.

Relative frequency Fr implies that

$$P_1^{emp}(E_0) = 0.008, \quad P_1^{emp}(E_1) = 0.096, \quad P_1^{emp}(E_2) = 0.384$$
$$P_1^{emp}(E_3) = 0.512, \quad P_1^{emp}(E_4) = 0.$$

This corresponds to collecting the following probabilistic evidence: three failures with probability 0.008, only one success with probability 0.096, two successes with probability 0.384, and three successes with probability 0.512. We are now ready to compute the DPK update of our initial P. Recall that we assumed it to be exchangeable, so we have

$$P(\{000\}) = p_0, \quad P(\{001\}) = P(\{100\}) = P(\{010\}) = p_1,$$
$$P(\{110\}) = P(\{101\}) = P(\{011\}) = p_2, \quad P(\{111\}) = p_3.$$

Suppose $\beta(n_t) = 1/n_t$; in turn we have

$$\begin{aligned}
P_{\mathcal{E}_1}(\{000\}) &= \frac{p_0}{P_{\mathcal{E}_0}(E_0)}\left(\frac{p_0}{1000} + \frac{999}{1000}P_1^{emp}(E_0)\right) \\
&= 1 \cdot \left(\frac{p_0}{1000} + \frac{7.992}{1000}\right) = \frac{p_0 + 7.992}{1000}, \\
P_{\mathcal{E}_1}(\{001\}) = P_{\mathcal{E}_1}(\{010\}) = P_{\mathcal{E}_1}(\{100\}) &= \frac{p_1}{P_{\mathcal{E}_0}(E_1)}\left(\frac{p_1}{1000} + \frac{999}{1000}P_1^{emp}(E_1)\right) \\
&= \frac{1}{3} \cdot \left(\frac{p_1}{1000} + \frac{95.904}{1000}\right) = \frac{p_1 + 95.904}{3000}, \\
P_{\mathcal{E}_1}(\{011\}) = P_{\mathcal{E}_1}(\{110\}) = P_{\mathcal{E}_1}(\{101\}) &= \frac{p_2}{P_{\mathcal{E}_0}(E_2)}\left(\frac{p_2}{1000} + \frac{999}{1000}P_1^{emp}(E_2)\right) \\
&= \frac{1}{3} \cdot \left(\frac{p_2}{1000} + \frac{383.616}{1000}\right) = \frac{p_2 + 383.616}{3000}, \\
P_{\mathcal{E}_1}(\{111\}) &= \frac{p_3}{P_{\mathcal{E}_0}(E_3)}\left(\frac{p_3}{1000} + \frac{999}{1000}P_1^{emp}(E_3)\right) \\
&= 1 \cdot \left(\frac{p_3}{1000} + \frac{511.488}{1000}\right) = \frac{p_3 + 511.488}{1000}.
\end{aligned}$$

We can see how, because we chose P to be exchangeable, in the case of only one successful outcome the updated probability $P_{\mathcal{E}_1}$ assigned to $\{001\}$, $\{010\}$, and $\{100\}$ is exactly $1/3$ of the mixture between the prior and the empirical probability of E_1. The same is true for the case of two successful outcomes.

To generalize the DPK updating presented here to a DIPK updating involving a set $\mathcal{P}$ of probability measures representing the initial beliefs of the agent one can follow the procedure explained in Section 3.9.2.

3.9.2 Soccer match results

This example is built on (Walley, 1991, Section 4.6.1). Let $\Omega = \{W, D, L\}$ represent the result of soccer match Juventus Turin vs Inter Milan, where W denotes a win for Juventus Turin, D a draw, and L a loss for Juventus Turin. Let then $X : \Omega \to \mathcal{X} =$

$\{0,1\}$, where 1 denotes a useful result (a victory or a draw) and 0 denotes a defeat, so X can be thought of as a Bernoulli random variable with unknown parameter. It is immediate to see how the finest partition of Ω according to DPK is given by $\tilde{\mathcal{E}} = \{E_1, E_2, E_3\}$, where $E_1 = \{W, D\}$, $E_2 = \{L\}$, and $E_3 = \varnothing$. Although $\tilde{\mathcal{E}}$ is attained almost immediately (it is enough to observe $x_j \neq x_k$, for some $j \neq k$), we maintain notation $P_{\mathcal{E}_t}$ to denote the t-th update of $P \equiv P_{\mathcal{E}_0}$. $P_{\tilde{\mathcal{E}}}$ will denote the limit of sequence $(P_{\mathcal{E}_t})$.

The data points $x_1, \ldots, x_n$ that we collect represent the outcomes of past matches. Because the two teams are well established and high-level, it is reasonable to assume that function X is fixed.

Let us describe how to perform a DIPK update of subjective beliefs in this context. Let the agent specify $\mathcal{P} \subset \Delta(\Omega, \mathcal{F})$, and suppose that the lower and upper probabilities $\underline{P} \equiv \underline{P}_{\mathcal{E}_0}$ and $\overline{P} \equiv \overline{P}_{\mathcal{E}_0}$ associated with $\mathcal{P}$ are such that $\underline{P}(W) = \underline{P}(D) = 0.27$, $\overline{P}(W) = \overline{P}(D) = 0.52$, $\underline{P}(L) = 0.21$, and $\overline{P}(L) = 0.31$.[5]

A simplex representation is given in Figure 3.1 where each assessment is represented by a line parallel to one side of the simplex.[6] The initial beliefs of the agent are encapsulated in $\mathcal{P}^{\text{co}}_{\mathcal{E}_0} = \text{core}(\underline{P})$. To update $\mathcal{P}^{\text{co}}_{\mathcal{E}_0}$ we need to find $\mathcal{P}_{\mathcal{E}_0} = \text{ex}\mathcal{P}^{\text{co}}_{\mathcal{E}_0}$. This is an easy job; it is sufficient to

1. equate $P(\omega)$ to either $\underline{P}(\omega)$ or $\overline{P}(\omega)$ for two of the three events. The probability of the third is then determined;

2. check which of the resulting P satisfies $\underline{P} \leqslant P \leqslant \overline{P}$.

[5] We write $\underline{P}(\omega)$ in place of $\underline{P}(\{\omega\})$ and $\overline{P}(\omega)$ in place of $\overline{P}(\{\omega\})$, $\omega \in \{W, D, L\}$, for notational convenience.

[6] Notice that the higher the values assigned by P to $\{\omega\} \subset \Omega$, the closer the line representing $P(\{\omega\})$ is to vertex $\omega \in \{W, D, L\}$.

This procedure gives us four extreme points $\mathcal{P}_{\mathcal{E}_0} = \{P^{ex}_{1,\mathcal{E}_0}, P^{ex}_{2,\mathcal{E}_0}, P^{ex}_{3,\mathcal{E}_0}, P^{ex}_{4,\mathcal{E}_0}\}$ such that

$$(P^{ex}_{1,\mathcal{E}_0}(W), P^{ex}_{1,\mathcal{E}_0}(D), P^{ex}_{1,\mathcal{E}_0}(L)) = (0.52, 0.27, 0.21),$$

$$(P^{ex}_{2,\mathcal{E}_0}(W), P^{ex}_{2,\mathcal{E}_0}(D), P^{ex}_{2,\mathcal{E}_0}(L)) = (0.27, 0.42, 0.31),$$

$$(P^{ex}_{3,\mathcal{E}_0}(W), P^{ex}_{3,\mathcal{E}_0}(D), P^{ex}_{3,\mathcal{E}_0}(L)) = (0.42, 0.27, 0.31),$$

$$(P^{ex}_{4,\mathcal{E}_0}(W), P^{ex}_{4,\mathcal{E}_0}(D), P^{ex}_{4,\mathcal{E}_0}(L)) = (0.27, 0.52, 0.21).$$

The extrema $\mathcal{P}_{\mathcal{E}_0}$ of $\mathcal{P}^{\text{co}}_{\mathcal{E}_0}$ are the vertices of the grey trapezoid in Figure 3.1.

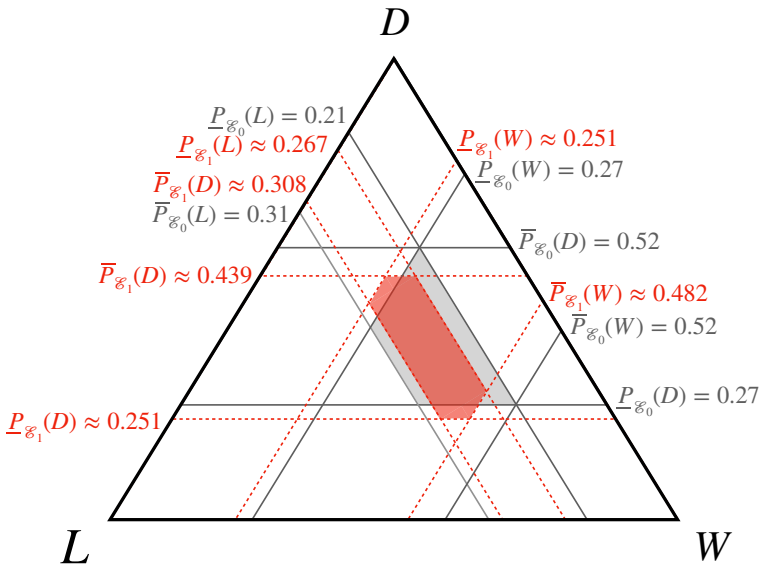

FIGURE 3.1: Visual representation of $\mathcal{P}^{\text{co}}_{\mathcal{E}_0}$ (the grey trapezoid) and of $\mathcal{P}^{\text{co}}_{\mathcal{E}_1}$ (the red hexagon) in our soccer example.

As of January 12, 2022, there have been 257 matches between the two teams, with 178 useful results for Juventus Turin and 79 wins for Inter Milan.[7] This is to say that we observe $x_1, \ldots, x_{257}$ such that 178 are 1's, and 79 are 0's. Then, to compute $\mathcal{P}^{\text{co}}_{\mathcal{E}_1}$ it is enough to update the extrema in $\mathcal{P}_{\mathcal{E}_0}$ so to obtain $\mathcal{P}_{\mathcal{E}_1}$, and then consider the convex hull of the latter. The partition induced by the collected data is $\mathcal{E}_1 = \{E_1, E_2, E_3\}$, and we have that $P^{emp}_1(E_1) = 178/257$, $P^{emp}_1(E_2) = 79/257$ and

[7] Data available here.

$P_1^{emp}(E_3) = 0$. This corresponds to collecting the following probabilistic evidence: Juventus Turin obtains a useful result with probability 178/257, and it loses with probability 79/257. Let us update $P^{ex}_{1,\mathcal{E}_0}$ to $P^{ex}_{1,\mathcal{E}_1}$. Suppose $\beta(n_t) = \frac{1}{\log(n_t+1)}$; we have

$$\begin{aligned}
P^{ex}_{1,\mathcal{E}_1}(W) &= \frac{P^{ex}_{1,\mathcal{E}_0}(W)}{P^{ex}_{1,\mathcal{E}_0}(E_1)} P^{ex}_{1,\mathcal{E}_1}(E_1) \\
&= \frac{0.52}{0.52+0.27}\left(\frac{0.52+0.27}{\log(258)} + \frac{\log(258)-1}{\log(258)} \cdot \frac{178}{257}\right) \approx 0.482, \\
P^{ex}_{1,\mathcal{E}_1}(D) &= \frac{P^{ex}_{1,\mathcal{E}_0}(D)}{P^{ex}_{1,\mathcal{E}_0}(E_1)} P^{ex}_{1,\mathcal{E}_1}(E_1) \\
&= \frac{0.27}{0.52+0.27}\left(\frac{0.52+0.27}{\log(258)} + \frac{\log(258)-1}{\log(258)} \cdot \frac{178}{257}\right) \approx 0.251, \\
P^{ex}_{1,\mathcal{E}_1}(L) &= \frac{P^{ex}_{1,\mathcal{E}_0}(L)}{P^{ex}_{1,\mathcal{E}_0}(E_2)} P^{ex}_{1,\mathcal{E}_1}(E_2) = 1 \cdot \left(\frac{0.21}{\log(258)} + \frac{\log(258)-1}{\log(258)} \cdot \frac{79}{257}\right) \approx 0.267,
\end{aligned}$$

so

$$(P^{ex}_{1,\mathcal{E}_1}(W), P^{ex}_{1,\mathcal{E}_1}(D), P^{ex}_{1,\mathcal{E}_1}(L)) \approx (0.482, 0.251, 0.267).$$

The other elements of $\mathcal{P}_{\mathcal{E}_0}$ are updated similarly. In particular,

$$(P^{ex}_{2,\mathcal{E}_1}(W), P^{ex}_{2,\mathcal{E}_1}(D), P^{ex}_{2,\mathcal{E}_1}(L)) \approx (0.271, 0.421, 0.308),$$

$$(P^{ex}_{3,\mathcal{E}_1}(W), P^{ex}_{3,\mathcal{E}_1}(D), P^{ex}_{3,\mathcal{E}_1}(L)) \approx (0.421, 0.271, 0.308),$$

$$(P^{ex}_{4,\mathcal{E}_1}(W), P^{ex}_{4,\mathcal{E}_1}(D), P^{ex}_{4,\mathcal{E}_1}(L)) \approx (0.251, 0.482, 0.267).$$

So we have that $\underline{P}_{\mathcal{E}_1}(W) \approx 0.251 \approx \underline{P}_{\mathcal{E}_1}(D)$, $\overline{P}_{\mathcal{E}_1}(W) \approx 0.482 \approx \overline{P}_{\mathcal{E}_1}(D)$, $\underline{P}_{\mathcal{E}_1}(L) \approx 0.267$, and $\overline{P}_{\mathcal{E}_1}(L) \approx 0.308$. As we can see from Figure 3.1, the graphical representation of $\mathcal{P}^{\text{co}}_{\mathcal{E}_1}$ is a hexagon (in red). Notice also that, since $0.267 \approx \underline{P}_{\mathcal{E}_1}(L) > \underline{P}_{\mathcal{E}_0}(L) = 0.21$ and $0.308 \approx \overline{P}_{\mathcal{E}_1}(L) < \overline{P}_{\mathcal{E}_0}(L) = 0.31$, we have that $\mathcal{P}^{\text{co}}_{\mathcal{E}_1}$ exhibits DIPK-contraction with respect to $\mathcal{P}^{\text{co}}_{\mathcal{E}_0}$ for $\{L\}$.

4

Ergodic Theorems in Dynamic Imprecise Probability Kinematics

4.1 Introduction

Given an initial set of probability measures $\mathcal{P}^{\text{co}}_{\mathcal{E}_0}$ representing the agent's initial beliefs about the state space Ω and the sequence of its successive DIPK updates $(\mathcal{P}^{\text{co}}_{\mathcal{E}_n})$, we are interested in studying its limit $\mathcal{P}^{\text{co}}_{\tilde{\mathcal{E}}}$. We use tools from dynamical systems theory to provide an ergodic theory for $\mathcal{P}^{\text{co}}_{\tilde{\mathcal{E}}}$.

4.1.1 Why ergodic theory?

Ergodic theory is a branch of mathematics that studies the long-term average behavior of complex dynamical systems. It was first introduced in (Boltzmann, 1887). Working with gases, the author suggested that the spatial average values giving rise to macroscopic features also arose as averages over time of observable quantities that could be calculated from microscopic states. Hence, what can be considered the "ergodic mantra": space average equals time average. The best-known ergodic theorem is arguably Birkhoff's one.

Theorem 42. (Birkhoff, cf. (Cornfeld et al., 1982)) If we have a probability space $(\Omega, \mathcal{F}, P)$, a measurable self-map $T : \Omega \to \Omega$ such that $P(A) = P(T^{-1}(A))$, for all $A \in \mathcal{F}$, and a measurable functional f on Ω, then the limit as $n \to \infty$ of the time average $\frac{1}{n}\sum_{j=1}^{n} f(T^{j-1}(\omega))$ exists and it is equal to the space average $\frac{1}{P(\Omega)}\int_\Omega f \mathrm{d}P = \int_\Omega f \mathrm{d}P$, P-almost surely.

This result is especially meaningful because it characterizes the behavior of the orbit of operator T over a large time period. In particular, the time average

$$\frac{1}{n}\sum_{j=1}^{n} f(T^{j-1}(\omega))$$

of f will almost surely converge to $\mathbb{E}_P(f) = \int_\Omega f \ \mathrm{d}P$ as the time horizon n receeds to infinity, where "almost surely" means that the probability that it does not happen is zero.

The importance of ergodic theory for computer scientists and statisticians is well documented in many works. For example, in (Berry et al., 2020), the authors combine ideas from the theory of dynamical systems with learning theory, providing an effective route to data-driven models of complex systems. They obtain refinable predictions as the amount of training data increases, and physical interpretability through discovery of coherent patterns around which the dynamics is organized. In (Zhang, 2018), the author proposes a generalization of ergodic measure preserving flow (EMPF), an optimisation-based inference method using ergodic results that overcomes the biasedness limitations of both Markov chain Monte Carlo (MCMC) and variational inference (VI). Such generalization, called ergodic inference, is necessary because of the lack of theoretical proof of the validity of EMPF. In (Arbabi and Mezić, 2017), the authors establish the convergence of a class of numerical algorithms, known as dynamic mode decomposition (DMD), for computation of the eigenvalues and eigenfunctions of the infinite dimensional Koopman operator. Koopman opera-

tor theory is an alternative formulation of dynamical systems theory which provides a versatile framework for data-driven study of high-dimensional nonlinear systems. Their work rely on the assumption that the underlying dynamical system is ergodic. In (Vose, 1999), the author points out that genetic algorithms are strongly related to dynamical systems. Ergodicity of such systems corresponds to an important property, called asymptotic correctness, roughly guaranteeing to eventually explore the whole solution space.

Ergodic theorems for lower probabilities have been studied in the context of imprecise Markov chains in (Bock and T'Joens, 2021; de Cooman et al., 2009), for capacity preserving $\mathbb{Z}_+^d$-actions in (Wu and Li, 2022), and more in general in (Cerreia-Vioglio et al., 2015). In the latter the authors work with an uncountable state space Ω. In section 4.2 we provide similar results in the context of dynamic imprecise probability kinematics.

Our main results is Theorem 45, where we show that $\frac{1}{k}\sum_{j=1}^{k} f(T^{j-1}(\omega))$, that is, our time average, may not converge now but – under some regularity assumptions – will almost surely be eventually contained in the interval

$$\left[\sum_{\omega\in\Omega} f^\star(\omega)\underline{P}_{\tilde{\mathcal{E}}}(\{\omega\}), \sum_{\omega\in\Omega} f^\star(\omega)\overline{P}_{\tilde{\mathcal{E}}}(\{\omega\})\right],$$

for a well defined function $f^\star$ on Ω. This means that the limit infimum and the limit supremum of $\frac{1}{k}\sum_{j=1}^{k} f(T^{j-1}(\omega))$ coincide and belong almost surely to the aforementioned interval. Their endpoints are given by the expected values of $f^\star$ with respect to the lower and upper probability measures associated with $\mathcal{P}^{\text{co}}_{\tilde{\mathcal{E}}}$, the limit set of the sequence $(\mathcal{P}^{\text{co}}_{\mathcal{E}_n})$ of DIPK updates of set $\mathcal{P}^{\text{co}}_{\mathcal{E}_0}$ representing the initial beliefs of the agent. Notice that here almost surely means that the lower probability $\underline{P}_{\tilde{\mathcal{E}}}$ that the event happens is 1.

4.2 Ergodic theory for the limit of $(\mathcal{P}^{\text{co}}_{\mathcal{E}_n})$

Recall that in the DIPK procedure described in chapter 3, Ω is assumed finite or countable. Let $T : \Omega \to \Omega$ be an $\mathcal{F}\backslash\mathcal{F}$-measurable transformation (this corresponds to the ergodic operator in classical ergodic theory) that explores all the state space, that is, for all $\omega \in \Omega$,

$$\bigcup_{j\in\mathbb{N}} T^{j-1}(\omega) = \Omega.$$

Let also $f : \Omega \to \mathbb{R}$ belong to $B(\Omega, \mathcal{F})$, the set of bounded and $\mathcal{F}$-measurable functionals on Ω. Then, the limit of the empirical average $\frac{1}{k}\sum_{j=1}^{k} f(T^{j-1}(\omega))$ as $k \to \infty$ exists $\underline{P}_{\tilde{\mathcal{E}}}$-almost surely, and it belongs to the interval generated by the space averages taken with respect to the boundary elements of $\mathcal{P}^{\text{co}}_{\tilde{\mathcal{E}}}$,

$$\mathcal{A}(\omega) := \left[\sum_{\omega\in\Omega} f^{\star}(\omega)\underline{P}_{\tilde{\mathcal{E}}}(\{\omega\}), \sum_{\omega\in\Omega} f^{\star}(\omega)\overline{P}_{\tilde{\mathcal{E}}}(\{\omega\})\right],$$

$\underline{P}_{\tilde{\mathcal{E}}}$-almost surely, for a well defined functional $f^{\star}$. That is,

$$\underline{P}_{\tilde{\mathcal{E}}}\left(\left\{\omega \in \Omega : \lim_{k\to\infty} \frac{1}{k}\sum_{j=1}^{k} f(T^{j-1}(\omega)) \in \mathcal{A}(\omega)\right\}\right) = 1.$$

Before giving the main result, we need to introduce three concepts. We say that a generic lower probability ν is (T-)invariant if, for all $A \in \mathcal{F}$,

$$\nu(A) = \nu(T^{-1}(A)).$$

We then call $\mathcal{I} \subset \Delta(\Omega, \mathcal{F})$ the set of (T-)invariant probability measures, that is,

$$\mathcal{I} := \left\{P \in \Delta(\Omega, \mathcal{F}) : P(A) = P(T^{-1}(A)),\ \forall A \in \mathcal{F}\right\}.$$

We call $\mathcal{G} \in \mathcal{F}$ the set of all (T-)invariant events of $\mathcal{F}$, that is,

$$\mathcal{G} := \left\{A \in \mathcal{F} : T^{-1}(A) = A\right\}.$$

Finally, we say that a generic lower probability ν is ergodic if and only if $\nu(\mathcal{G}) = \{0, 1\}$, that is, ν assigns value 0 or 1 to all the elements of $\mathcal{G}$.

Lemma 43. If there exist $T \in \mathbb{N}_0$ such that for all $t \geqslant T$ we can always find a collection $\{P^A_{\mathcal{E}_t}\}_{A\in\mathcal{F}} \subset \mathcal{P}_{\mathcal{E}_t}$ such that for $A' \in \mathcal{F}$

- $P^{A'}_{\mathcal{E}_t}(A') = P^{A'}_{\mathcal{E}_t}(T^{-1}(A'))$,
- $P^{A'}_{\mathcal{E}_t}(A') \leqslant P_{\mathcal{E}_t}(A')$, for all $P_{\mathcal{E}_t} \in \mathcal{P}_{\mathcal{E}_t}$,
- $P^{A'}_{\mathcal{E}_t}(T^{-1}(A')) \leqslant P_{\mathcal{E}_t}(T^{-1}(A'))$, for all $P_{\mathcal{E}_t} \in \mathcal{P}_{\mathcal{E}_t}$,

then $\underline{P}_{\tilde{\mathcal{E}}}$ is T-invariant.

Notice that we require collection $\{P^A_{\mathcal{E}_t}\}_{A\in\mathcal{F}}$ to be a subset of $\mathcal{P}_{\mathcal{E}_t} = ex(\mathcal{P}^{\text{co}}_{\mathcal{E}_t})$ because by Theorem 6 we have that $\underline{P}_{\mathcal{E}_t}(A) = \inf_{P\in\mathcal{P}_{\mathcal{E}_t}} P(A) = \inf_{P\in\mathcal{P}^{\text{co}}_{\mathcal{E}_t}} P(A)$, for all $A \in \mathcal{F}$.

Lemma 44. If there exist $T \in \mathbb{N}_0$ and $P'_{\mathcal{E}_T} \in \mathcal{P}_{\mathcal{E}_T}$ such that $P'_{\mathcal{E}_T}(A) = 0$ for all $A \in \mathcal{G}$, then $\underline{P}_{\tilde{\mathcal{E}}}$ is ergodic.

A result of this lemma is that if the agent selects $\mathcal{P}$ such that it contains an element P' that assigns probability 0 to all the elements in $\mathcal{G}$, then P' belongs to $\mathcal{P}^{\text{co}}_{\mathcal{E}_0}$. In turn this ensures that $\underline{P}_{\tilde{\mathcal{E}}}$ is ergodic. The following is our main result.

Theorem 45. If $\underline{P}_{\tilde{\mathcal{E}}}$ is invariant, then for all $f \in B(\Omega, \mathcal{F})$, there exists $f^\star \in B(\Omega, \mathcal{G})$ – that is, there exists a bounded and $\mathcal{G}$-measurable funtional $f^\star$ on Ω – such that

$$\lim_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} f(T^{j-1}(\omega)) = f^\star(\omega) \quad \underline{P}_{\tilde{\mathcal{E}}} - a.s. \tag{4.1}$$

If in addition $\underline{P}_{\tilde{\mathcal{E}}}$ is ergodic, then

$$\sum_{\omega\in\Omega} f^\star(\omega)\underline{P}_{\tilde{\mathcal{E}}}(\{\omega\}) \leqslant \lim_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} f(T^{j-1}(\omega)) \leqslant \sum_{\omega\in\Omega} f^\star(\omega)\overline{P}_{\tilde{\mathcal{E}}}(\{\omega\}) \tag{4.2}$$

$\underline{P}_{\tilde{\mathcal{E}}}$-almost surely.

We now give a subadditive ergodic theorem for $\underline{P}_{\tilde{\mathcal{E}}}$ and, as a result, we find a sharpening of Theorem 45 when some additional assumptions are met. Before giving these results, we need to introduce six concepts.[1] We say that a generic lower probability ν is

- convex if $\nu(A \cup B) + \nu(A \cap B) \geqslant \nu(A) + \nu(B)$, for all $A, B \in \mathcal{F}$;
- continuous at Ω if $\lim_{k\to\infty} \nu(A_k) = \nu(\Omega)$, when $A_k \uparrow \Omega$;
- strongly invariant if and only if for every $A \in \mathcal{F}$, the following holds

$$\nu(A \backslash T^{-1}(A)) = \overline{\nu}(T^{-1}(A) \backslash A) \quad \text{and} \quad \nu(T^{-1}(A) \backslash A) = \overline{\nu}(A \backslash T^{-1}(A)),$$

  where $\overline{\nu}$ is the upper probability associated with ν, that is, its conjugate;
- functionally invariant if and only if $\mathcal{M} \subset \mathcal{I}$, where $\mathcal{M} \subset \Delta(\Omega, \mathcal{F})$ is the set for which $\nu(A) = \inf_{P\in\mathcal{M}} P(A)$, for all $A \in \mathcal{F}$.

Lemma 46. If either of the following hold, then $\underline{P}_{\tilde{\mathcal{E}}}$ is convex

(i) There exists $T \in \mathbb{N}_0$ such that for all $t \geqslant T$ and all $A, B \in \mathcal{F}$ such that $A \subset B$, there exists $P'_{\mathcal{E}_t} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_t}$ such that

$$P'_{\mathcal{E}_t}(A \mid E) = \underline{P}^G_{\mathcal{E}_t}(A \mid E) \quad \text{and} \quad P'_{\mathcal{E}_t}(B \mid E) = \underline{P}^G_{\mathcal{E}_t}(B \mid E), \quad \forall E \in \mathcal{E}_{t+1}.$$

(ii) There exists $T \in \mathbb{N}_0$ such that for all $t \geqslant T$ and all finite chains $(A_i)_{i=1}^n \subset \mathcal{F}$, there exists $P'_{\mathcal{E}_t} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_t}$ such that

$$P'_{\mathcal{E}_t}(A_i \mid E) = \underline{P}^G_{\mathcal{E}_t}(A_i \mid E), \quad \forall i \in \{1, \ldots, n\}, \forall E \in \mathcal{E}_{t+1}.$$

(iii) There exists $T \in \mathbb{N}_0$ such that for all $t \geqslant T$ and all chains $(A_i)_{i\in I} \subset \mathcal{F}$, there exists $P'_{\mathcal{E}_t} \in \mathcal{P}_{\mathcal{E}_t}$ such that

$$P'_{\mathcal{E}_t}(A_i \mid E) = \underline{P}^G_{\mathcal{E}_t}(A_i \mid E), \quad \forall i \in I, \forall E \in \mathcal{E}_{t+1}.$$

[1] Two were already introduced in Definition 5.

Lemma 47. The following are true

(i) $\underline{P}_{\tilde{\mathcal{E}}}$ is always continuous at Ω.

(ii) If there exists $T \in \mathbb{N}_0$ such that for all $t \geqslant T$, $\underline{P}_{\mathcal{E}_t}$ is strongly invariant, then $\underline{P}_{\tilde{\mathcal{E}}}$ is strongly invariant.

(iii) If there exists $T \in \mathbb{N}_0$ such that for all $t \geqslant T$, $P_{\mathcal{E}_t}(A \cap E) = P_{\mathcal{E}_t}(T^{-1}(A) \cap E)$, for all $A \in \mathcal{F}$, all $E \in \mathcal{E}_{t+1}$, and all $P_{\mathcal{E}_t} \in \mathcal{P}_{\mathcal{E}_t}$, then $\underline{P}_{\tilde{\mathcal{E}}}$ is functionally invariant.

We call a sequence (S_k) of $\mathcal{F}$-measurable random variables superadditive if $S_{k+\ell} \geqslant S_k + S_\ell \circ T^k$, for all k and all ℓ. It is subadditive if the opposite inequality holds. It is additive if it is both super- and subadditive. A characterization of an additive sequence is the following: (S_k) is additive if and only if there exists an $\mathcal{F}$-measurable functional f on Ω such that

$$S_k = \sum_{j=1}^{k} f \circ T^{j-1}, \quad \forall k \in \mathbb{N}. \tag{4.3}$$

If we consider (S_k) as in (4.3) and we take its absolute value, that is, if we consider $(|S_k|)$, we obtain a subadditive sequence. Notice also that if f in our characterization belongs to $B(\Omega, \mathcal{F})$, we have that there exists $\lambda \in \mathbb{R}$ such that

$$-\lambda k \leqslant S_k(\omega) \leqslant \lambda k, \quad \forall k \in \mathbb{N}, \omega \in \Omega. \tag{4.4}$$

Similarly, $-\lambda k \leqslant |S_k(\omega)| \leqslant \lambda k$, for all $k \in \mathbb{N}$ and all $\omega \in \Omega$.

Lemma 48. Let (S_k) be a superadditive sequence satisfying (4.4), and suppose $\mathcal{P}_{\tilde{\mathcal{E}}} \subset \mathcal{I}$. Define then the sequence $(a_k) \in \mathbb{R}^{\mathbb{N}}$ as $a_k := -\sum_{\omega \in \Omega} \inf_{P \in \mathcal{P}_{\tilde{\mathcal{E}}}} S_k(\omega) P(\{\omega\})$, for all $k \in \mathbb{N}$. Then, (a_k) is subadditive, that is, $a_{k+\ell} \leqslant a_k + a_\ell$, for all $k, \ell \in \mathbb{N}$. If (S_k) is subadditive, we reach the same result by defining $a_k := \sum_{\omega \in \Omega} \sup_{P \in \mathcal{P}_{\tilde{\mathcal{E}}}} S_k(\omega) P(\{\omega\})$.

Theorem 49. If (S_k) is a super- or subadditive sequence satisfying (4.4) and $\underline{P}_{\tilde{\mathcal{E}}}$ is functionally invariant, then there is $f^\star \in B(\Omega, \mathcal{G})$ such that

$$\lim_{k\to\infty} \frac{1}{k} S_k(\omega) = f^\star(\omega) \quad \underline{P}_{\tilde{\mathcal{E}}} - a.s.$$

In addition,

1. If $\underline{P}_{\tilde{\mathcal{E}}}$ is convex and strongly invariant, and (S_k) is superadditive, then

$$\sum_{\omega\in\Omega} f^\star(\omega)\underline{P}_{\tilde{\mathcal{E}}}(\{\omega\}) = \sup_{k\in\mathbb{N}} \frac{1}{k} \sum_{\omega\in\Omega} S_k(\omega)\underline{P}_{\tilde{\mathcal{E}}}(\{\omega\}).$$

2. If $\underline{P}_{\tilde{\mathcal{E}}}$ is convex and strongly invariant, and (S_k) is subadditive, then

$$\sum_{\omega\in\Omega} f^\star(\omega)\overline{P}_{\tilde{\mathcal{E}}}(\{\omega\}) = \inf_{k\in\mathbb{N}} \frac{1}{k} \sum_{\omega\in\Omega} S_k(\omega)\overline{P}_{\tilde{\mathcal{E}}}(\{\omega\}).$$

3. If $\underline{P}_{\tilde{\mathcal{E}}}$ is ergodic and (S_k) is either super- or subadditive, then

$$\sum_{\omega\in\Omega} f^\star(\omega)\underline{P}_{\tilde{\mathcal{E}}}(\{\omega\}) \leqslant \lim_{k\to\infty} \frac{1}{k} S_k(\omega) \leqslant \sum_{\omega\in\Omega} f^\star(\omega)\overline{P}_{\tilde{\mathcal{E}}}(\{\omega\})$$

   $\underline{P}_{\tilde{\mathcal{E}}}$-almost surely.

Corollary 50. If $\underline{P}_{\tilde{\mathcal{E}}}$ is convex and strongly invariant, then for all $f \in B(\Omega, \mathcal{F})$ there exists $f^\star \in B(\Omega, \mathcal{G})$ such that

$$\lim_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} f(T^{j-1}(\omega)) = f^\star(\omega) \quad \underline{P}_{\tilde{\mathcal{E}}} - a.s. \tag{4.5}$$

In addition, the following are true

1. For every $P \in \mathcal{I}$, $f^\star$ is a version of the conditional expectation of f given $\mathcal{G}$.

2. $\sum_{\omega\in\Omega} f^\star(\omega)\underline{P}_{\tilde{\mathcal{E}}}(\{\omega\}) = \sum_{\omega\in\Omega} f(\omega)\underline{P}_{\tilde{\mathcal{E}}}(\{\omega\})$.

3. If $\underline{P}_{\tilde{\mathcal{E}}}$ is ergodic, then

$$\sum_{\omega\in\Omega} f(\omega)\underline{P}_{\tilde{\mathcal{E}}}(\{\omega\}) \leqslant \lim_{k\to\infty}\frac{1}{k}\sum_{j=1}^{k} f(T^{j-1}(\omega)) \leqslant \sum_{\omega\in\Omega} f(\omega)\overline{P}_{\tilde{\mathcal{E}}}(\{\omega\})$$

$\underline{P}_{\tilde{\mathcal{E}}}$-almost surely.

4.2.1 A strong law of large numbers

A consequence of Theorem 45 is a strong law of large numbers. Before stating it, we need to introduce two notions. We first generalize the concept of a stationary stochastic process by allowing the underlying probability measure to be a lower probability. We then present the shift map, a classic idea in dynamics and ergodic theory.

Denote by $\mathbf{f} \equiv (f_k)_{k\in\mathbb{N}} \in B(\Omega,\mathcal{F})^{\mathbb{N}}$ a sequence of bounded and $\mathcal{F}$-measurable functionals on Ω, and call $\mathcal{T} := \bigcap_{\ell\in\mathbb{N}} \sigma(f_\ell, f_{\ell+1}, \ldots)$ the tail sigma-algebra.

Given a generic lower probability ν on $(\Omega,\mathcal{F})$, $\mathbf{f}$ is stationary if and only if, for all $k\in\mathbb{N}$, all $\ell\in\mathbb{N}_0$, and all Borel subset $B\subset\mathbb{R}^{\ell+1}$,

$$\nu\left(\{\omega\in\Omega : (f_k(\omega),\ldots,f_{k+\ell}(\omega))\in B\}\right) = \nu\left(\{\omega\in\Omega : (f_{k+1}(\omega),\ldots,f_{k+\ell+1}(\omega))\in B\}\right).$$

Now, denote by $(\mathbb{R}^{\mathbb{N}},\sigma(\mathcal{C}))$ the measurable space of sequences endowed with the sigma-algebra generated by the algebra of cylinders. Also denote by $s:\mathbb{R}^{\mathbb{N}}\to\mathbb{R}^{\mathbb{N}}$ the shift transformation

$$s(x_1,x_2,x_3,\ldots) = (x_2,x_3,x_4,\ldots), \quad \forall x\in\mathbb{R}^{\mathbb{N}}.$$

The sequence $\mathbf{f}$ induces a (natural) measurable map between $(\Omega,\mathcal{F})$ and measurable space $(\mathbb{R}^{\mathbb{N}},\sigma(\mathcal{C}))$ defined by

$$\omega\mapsto\mathbf{f}(\omega) := (f_1(\omega),\ldots,f_k(\omega),\ldots).$$

Given any lower probability ν on $(\Omega, \mathcal{F})$, we can then define the map $\nu^{\mathbf{f}} : \sigma(\mathcal{C}) \to [0,1]$ as

$$C \mapsto \nu^{\mathbf{f}}(C) := \nu\left(\mathbf{f}^{-1}(C)\right).$$

We say that $\mathbf{f}$ is ergodic if and only if $\nu^{\mathbf{f}}$ is ergodic with respect to the shift transformation. The following is a direct consequence of (Cerreia-Vioglio et al., 2015, Lemma 1).

Lemma 51. If $\underline{P}_{\tilde{\mathcal{E}}}$ is convex and $\mathbf{f}$ is stationary, then $\underline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}}$ is convex, continuous at $\mathbb{R}^{\mathbb{N}}$, and shift invariant. In addition, $\underline{P}_{\tilde{\mathcal{E}}}(\mathcal{T}) = \{0,1\}$ implies that $\mathbf{f}$ is ergodic.

We are now ready for the strong law of large numbers.

Theorem 52. Let $\underline{P}_{\tilde{\mathcal{E}}}$ be convex. If $\mathbf{f} = (f_n)_{n\in\mathbb{N}}$ is stationary and ergodic, then

$$\sum_{\omega\in\Omega} f_1(\omega)\underline{P}_{\tilde{\mathcal{E}}}(\{\omega\}) \leqslant \lim_{k\to\infty} \frac{1}{k}\sum_{j=1}^{k} f_j(\omega) \leqslant \sum_{\omega\in\Omega} f_1(\omega)\overline{P}_{\tilde{\mathcal{E}}}(\{\omega\})$$

$\underline{P}_{\tilde{\mathcal{E}}}$-almost surely.

Notice that the assumption of stationarity gives us the fact that the limit for k growing to infinity of $\frac{1}{k}\sum_{j=1}^{k} f_j(\omega)$ exists $\underline{P}_{\tilde{\mathcal{E}}}$-almost surely. Then, to characterize this limit in terms of the expected values of f_1 taken with respect to $\underline{P}_{\tilde{\mathcal{E}}}$ and $\overline{P}_{\tilde{\mathcal{E}}}$, we need $\underline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}}$ to be ergodic.

5

Extended probabilities and their application to statistical inference

5.1 Introduction

Researchers use the terminology "extended probabilities" to refer to set functions whose codomain is either a superset of $[0, 1]$, or defined using entirely different number types, such as p-adic numbers (Khrennikov, 2009). They first came up in physics (noticed by (Dirac, 1930) and (Heisenberg, 1931)), where they are still studied today (Ferrie, 2011; Hartle, 2008; Kronz, 2007). They then became popular in other fields too, including economics and finance (Burgin and Meissner, 2011; Jarrow and Turnbull, 1995), machine learning (Lowe, 2007), stochastic processes (Wiewiora, 2008), and queuing theory (Tijms and Staats, 2007). Of course, they have been extensively studied in mathematics (see e.g. (Allen, 1976) and (Bartlett, 1945) for early works, and (Burgin, 2013) and (Khrennikov, 2009) for more recent ones). A complete account is given in (Burgin, 2012).

"Extended probabilities" have been given differing definitions and interpretations across – and even within – fields of study. In quantum theory, for example, (Benavoli

et al., 2019) point out that "negative probabilities" do not have intrinsic meaning beyond the fact that they constitute a probabilistic model compatible with quantum events. In (Gell-Mann and Hartle, 2012; Hartle, 2004, 2008), instead, the authors interpret them as being associated with unsettleable bets (see section 3.2).

In this chapter, we aim to give a foundational definition of extended probabilities. We give two properties that a set function must have in order to be called an extended probability, irrespective of the environment the scholar works in. This represents a great improvement with respect to previous works on the matter, in which definitions depend on the area of study.

We explain how extended probabilities are different from regular Kolmogorovian probabilities, and we characterize some of their more interesting properties. We also give a behavioral interpretation of extended probabilities taking on negative values that complements the frequentist one given in (Burgin, 2012) to "negative probabilities". Finally, we relate our definition and interpretation to the ones existing in the literature, and we illustrate how these latter can be reconciled with the ones we provide.

After that, we present what is to the best of our knowledge the first application of extended probabilities to an inference procedure. Consider a generic statistical experiment; let Ω be a finite or countable space, and endow it with the sigma-algebra $\mathcal{F} = 2^{\Omega}$. The space usually adopted to express uncertainty around the elements of $\mathcal{F}$ is the probability space $(\Omega, \mathcal{F}, P)$, for some probability measure $P : \mathcal{F} \to [0, 1]$. Now, suppose we want to express further uncertainty regarding either the composition of Ω, or which P to consider on $(\Omega, \mathcal{F})$; we can do so by using lower probabilities. In particular, in the first case, we consider the probability space $(\Omega, \mathcal{F}, P)$ and we follow the example in (Gong, 2018). There, as a consequence of survey nonresponse, the author is forced to consider $\check{\Omega} = 2^{\Omega}$. In this case, probabilities cannot be computed exactly: only lower and upper bounds to precise probabilities – lower and upper

probabilities, respectively – are available. In the second case, we consider the triple $(\Omega, \mathcal{F}, \mathcal{P})$, where $\mathcal{P}$ is a set of probability measures, and we proceed e.g. as in chapters 3 and 4.

To the best of our knowledge, there is no cogent way of expressing uncertainty on both the composition of Ω and which P to consider on $(\Omega, \mathcal{F})$. In this chapter, we aim to fill this gap by using extended probabilities. We describe an ex ante analysis, meaning one that takes place before the actual statistical analysis that requires the knowledge of the state space. In the most general case, we start from the number type we believe we are working with (naturals, wholes, integers, or rationals), and call it Ω. Then, at time 0 we divide it into an actual space Ω_0^+ that we deem a plausible state space for our experiment, and a latent one Ω_0^-, which we do not know about; notationally, $\Omega = \Omega_0^- \sqcup \Omega_0^+$, where $\sqcup$ denotes a disjoint union of sets. We assign negative extended probabilities to the subsets of Ω_0^-. To capture the uncertainty around which P to consider, we specify a set $\mathcal{P}^{ex}$ of extended probabilities supported on the whole Ω, instead of a single one. We then describe how we progressively discover the true composition of the state space associated with our experiment (which may be the whole Ω we initially specified, or a proper subset $\Omega' \subsetneq \Omega$), and how to update extended probabilities accordingly. We conclude our analysis by discovering the state space associated with our experiment.

In addition, we show that the limiting set of the sequence $(\mathcal{P}_t^{ex})$ of updates of the initially-specified set of extended probability measures must be one of the following. It is either a set of regular probability measures, if the state space associated with our experiment is the whole Ω, or a set of extended probability measures if the state space associated with our experiment is a proper subset Ω' of Ω. In this latter case, the limiting set induces a set of regular probability measures on Ω'.

We also develop the concept of lower and upper extended probabilities. They represent the "boundaries" of a generic set $\mathcal{P}^{ex}$ of extended probabilities, and more

in general allow for an imprecise elicitation of extended probabilities. They are extremely important because under a mild assumption, knowing lower probability $\underline{P}^{ex}$ is enough to be able to retrieve the whole set $\mathcal{P}^{ex}$. We provide bounds for the lower extended probability $\underline{P}^{ex}_t(A)$ of any element $A \in \mathcal{F}$, at any time t in our ex ante analysis. For the sake of completeness, we give the definition of an extended Choquet capacity, and we show how lower and upper extended probabilities are extended Choquet capacities.

In Example 71, we provide a simple application to the field of ecology. We illustrate how the analysis we describe in the chapter can be put to use in a species sampling problem, specifically to retrieve the number of birds that inhabit a certain region throughout the year.

We also provide an example – adapted from the one in (Allahverdyan and Galstyan, 2014) – in the field of opinion dynamics; in particular, we describe the boomerang effect. We have a persuading agent acting on a persuaded agent, but the latter perceives the former as having low credibility. This can be modeled so that the persuaded agent does not know the composition of the entire state space, while she suspects the persuading agent does: she thinks the persuading agent may be hiding something form her. As she discovers the true composition of the state space, the credibility of the persuading agent is restored.

This chapter is organized as follows: in section 5.2 we give the foundational definition of an extended probability, its properties, and its behavioral interpretation. In section 5.3 we use extended probabilities in an inference procedure. In section 5.4 we give the opinion dynamics example. Section 5.5 deals with lower extended probabilities.

Remark 53. To deal with uncertainty in the composition of Ω, one could proceed as in (Walley, 1991, section 4.3.3). The agent could begin the elicitation by spec-

ifying the state space Ω_0 and then, as they analyze the problem in greater detail, could realize that a refinement to a finer-grained Ω_1 is needed. This corresponds to specifying what (Dempster, 1967) calls a multivalued mapping from Ω_0 to Ω_1; that is,

$$A : \Omega_0 \rightrightarrows \Omega_1, \quad \omega_0 \mapsto A(\omega_0) \subset \Omega_1.$$

So Ω_0 corresponds to a partition of Ω_1, and each state $\omega_0 \in \Omega_0$ can be identified with the set $A(\omega_0)$ of "refined possibilities" in Ω_1. To illustrate the complication deriving from this approach, let Ω_0 and Ω_1 be finite or countable, and call $\mathcal{F}_0 = 2^{\Omega_0}$ and $\mathcal{F}_1 = 2^{\Omega_1}$. If the agent specifies a probability measure P_0 on $(\Omega_0, \mathcal{F}_0)$, they then need to come up with a probability measure P_1 on $(\Omega_1, \mathcal{F}_1)$ such that

$$P_0(\{\omega_0\}) = \sum_{\omega_1 \in A(\omega_0)} P_1(\{\omega_1\})$$

holds for all $\omega_0 \in \Omega_0$. This means that a new (subjective) probability elicitation must take place once the agent refines the state space to Ω_1. This can be avoided using extended probabilities, as we shall argue in section 5.3.

5.2 Extended probability measures

In this section, we first illustrate the philosophical reason to introduce extended probabilities, and then we dive into more technical details. We conclude with a thorough analysis on how our interpretation of extended probabilities relates to the existing literture.

5.2.1 Philosophical motivation for extended probabilities

We give a behavioral interpretation of extended probabilities. Consider a generic event A, and suppose we want to express our belief about the likelihood of it taking place. Suppose we can enter a bet about A that gives us \$1 if event A happens

and \$0 if it does not happen. Then, we say that the probability we attach to A is given by $p \geqslant 0$, the amount of money we deem fair to pay to enter the bet. Notice that if $p = 0$ we do not enter the bet, since we deem event A impossible. This interpretation is inspired by the classical subjective probability interpretation given in (de Finetti, 1974, 1975). As pointed out in (Nau, 2001), the other two fathers of subjective probability theory, (Ramsey, 1964) and (Savage, 1954) simultaneously introduced measurement schemes for utility. They tied their definitions of probability to bets in which the payoffs were effectively measured in utiles rather than dollars. In this way, they obtained probabilities that were interpretable as measures of pure belief, uncontaminated by marginal utilities for money. De Finetti later admitted that it might have been better to adopt the seemingly more general approach of Ramsey and Savage, since it leads to a theory of decision-making that does not rely on monetary values. Nevertheless, he found other reasons for preferring the money bet approach. In particular, he maintained that it would be extremely difficult to settle bets based on utiles, because the monetary sums needed to settle them would need to be adjusted to the complex variations in a unit of measure (utiles) that is unobservable.[1] This is why we retain the DeFinettian interpretation of (subjective) probability, and more in general why we adopt a betting scheme approach.

Suppose now that we are not given the possibility to enter a bet like the aforementioned one, so we cannot assess a probability for A as before. Instead, such a possibility is given to our doppelgänger, who tells us that their subjective assessment for the probability of A is some $q \in [0, 1]$. Here, it is assumed that the doppelgänger assigns the same probabilities as ourselves to all the events we both can enter a bet about. We can always find a person with such a preference pattern due to the axiom of dependent choice. This procedure of asking the doppelgänger is equivalent to setting the probability of A ourselves; we introduce the doppelgänger because, given

[1] The complete quotes from (de Finetti, 1974) can be found in Appendix C.

our interpretation of probability, it is impossible to elict the probability of event A if we cannot enter a bet similar to the one described before. We conclude that if we were given the opportunity to enter the bet about A, the amount of money we would deem fair to pay would be q dollars. Therefore we are prepared to lose $\$q$ in the case A^c happens. We express this by setting $p = -q$.

If after a while we are given the opportunity to enter a bet about an event A whose extended probability we deemed to be $-q < 0$, the price we consider fair to pay need not be $p = |-q|$. If $p \neq |-q|$, it means that we changed idea about how likely event A is once given the possibility of entering the bet. If instead $p = |-q|$, we are obeying to Allen's principle of conservation of knowledge (Allen, 1976). It states that, like the law of conservation of mass, information is not created nor destroyed, but just transformed; the total amount of knowledge possible is constant. By having us choose the probability equal in absolute value, we comply with conservation of knowledge. As we can see, this principle allows us to mechanically retrieve positive probabilities starting from negative ones. It also allows us to work backwards to negative probabilities starting from positive ones. Suppose we have an event $A \in \mathcal{F}$ to which we attach probability $p \geqslant 0$. Then, we know that if we were not given the opportunity to enter the bet that allowed us to indicate p as the probability that event A happens, then we would have expressed our uncertainty by assigning A probability $-p \leqslant 0$.

A Dutch book is, informally, the possibility of constructing a bet such that the bookmaker always profits, while the punter always loses money (see Remark 58 for a formal definition). In (de Finetti, 1937), a coherent subjective probability is defined as one that does not allow for a Dutch book to be made against the punter, however a bet is made. Necessary and sufficient conditions for coherence require that subjective probabilities satisfy the Kolmogorovian axioms of probability (with only finite additivity). As we shall see in section 5.2.2, this does not hold for extended

probabilities. This is not a problem though, in light of the interpretation we give to negative probabilities. The events whose attached probabilities are negative are events for which the punter cannot enter a bet; hence, they cannot be used to build a Dutch book. We give the definition of coherence in the context of extended probabilities and the formal statement that extended probabilities are always coherent in Remark 58. In addition, as we show in section 5.3, the inferential procedure we consider is such that, starting with extended probabilities, we recover – at the end of our analysis – regular probabilities. This also ensures that no Dutch books can be created.

5.2.2 Technical definition and properties

Consider a measurable space $(\Omega, \mathcal{F})$. An extended probability $P^{ex} : \mathcal{F} \to \mathbb{R}$ is a set function on $\mathcal{F}$ such that

(i*) $P^{ex}(A) \in [-1, 1]$, for all $A \in \mathcal{F}$;

(ii*) if $\{A_j\}_{j \in I}$ is a countable collection of disjoint events such that $\cup_{j \in I} A_j \in \mathcal{F}$, then

$$P^{ex}\left(\bigcup_{j \in I} A_j\right) = \sum_{j \in I} P^{ex}(A_j).$$

The Kolmogorovian axioms for any regular probability measure $P : \mathcal{F} \to \mathbb{R}$ are the following

- $P(A) \in [0, 1]$, for all $A \in \mathcal{F}$;
- $P(\Omega) = 1$;
- if $\{A_j\}_{j \in I}$ is a countable collection of disjoint events such that $\cup_{j \in I} A_j \in \mathcal{F}$, then $P(\cup_{j \in I} A_j) = \sum_{j \in I} P(A_j)$.

It is immediate to see, then, how conditions (i*) and (ii*) are more general than the axioms given by Kolmogorov. To this extent, extended probabilities are a generalization of the concept of regular probabilities. From its definition, we see that an extended probability is a finite signed measure. We call the triple $(\Omega, \mathcal{F}, P^{ex})$ an extended probability space.

Because we do not require $P^{ex}(\Omega) = 1$, the extended probability of the complement of an event A is given by

$$P^{ex}(A^c) = P^{ex}(\Omega \backslash A) = P^{ex}(\Omega) - P^{ex}(A). \tag{5.1}$$

Equation (5.1) comes from (ii*); indeed, consider the disjoint sets $\Omega \backslash A$ and $\Omega \cap A = A$. Then,

$$\begin{aligned} P^{ex}(\Omega) = P^{ex}\left([\Omega \backslash A] \sqcup A\right) = P^{ex}(\Omega \backslash A) + P^{ex}(A) \\ \iff P^{ex}(\Omega \backslash A) = P^{ex}(\Omega) - P^{ex}(A), \end{aligned}$$

where $\sqcup$ denotes the union of disjoint sets.

Equation (5.1) ensures us that $P^{ex}(\varnothing) = P^{ex}(\Omega^c) = P^{ex}(\Omega) - P^{ex}(\Omega) = 0$. We also have that $P^{ex}(\Omega) = \sum_{E_j \in \mathcal{E}} P^{ex}(E_j)$, where $\mathcal{E} = \{E_j\}$ is any finite or countable partition of Ω.

Proposition 54. Extended probabilities have the adequacy property: for all $A, B \in \mathcal{F}$ such that $A = B$, then $P^{ex}(A) = P^{ex}(B)$.

We also have that extended probabilities retain a version of the monotonic property of regular probabilities.

Proposition 55. For all $A, B \in \mathcal{F}$ such that $A \subset B$, if $P^{ex}(A) \geqslant 0$, $P^{ex}(B) \geqslant 0$, and $P^{ex}(B \cap A^c) \geqslant 0$, then $P^{ex}(A) \leqslant P^{ex}(B)$. If instead $P^{ex}(A) \leqslant 0$, $P^{ex}(B) \leqslant 0$, and $P^{ex}(B \cap A^c) \leqslant 0$, then $P^{ex}(A) \geqslant P^{ex}(B)$.

This version of the monotonic property implies a version of the continuity property enjoyed by regular probabilities.

Corollary 56. If we have a collection $\{A_j\}$ of elements of $\mathcal{F}$ such that

- it is nested (i.e. $A_1 \supset A_2 \supset \cdots \supset A_n \supset \cdots$),
- $\cap_j A_j = \varnothing$,
- $P^{ex}(A_j) \geqslant 0$, for all A_j,

then $\lim_{j\to\infty} P^{ex}(A_j) = P^{ex}(\cap_j A_j) = 0$.

Extended probabilities satisfy the inclusion-exclusion principle.

Proposition 57. The following is true:

$$P^{ex}(A \cup B) = P^{ex}(A) + P^{ex}(B) - P^{ex}(A \cap B),$$

for all $A, B \in \mathcal{F}$.

This implies immediately that, for any $A, B, C \in \mathcal{F}$,

$$\begin{aligned} P^{ex}((A \cup B) \cap C) &= P^{ex}((A \cap C) \cup (B \cap C)) \\ &= P^{ex}(A \cap C) + P^{ex}(B \cap C) - P^{ex}(A \cap B \cap C), \end{aligned}$$

by Proposition 57 and the De Morgan laws.

We define the extended conditional probability of A given B to be the extended counterpart of regular conditional probabilities,

$$P^{ex}(A \mid B) := \frac{P^{ex}(A \cap B)}{P^{ex}(B)} \in [-1, 1], \tag{5.2}$$

for all $A, B \in \mathcal{F}$ such that $P^{ex}(B) \neq 0$.

To avoid confusion arising from the sign, we say that $A, B \in \mathcal{F}$ are independent if and only if

$$|P^{ex}(A \cap B)| = |P^{ex}(A) \times P^{ex}(B)|.$$

This way of describing independence corresponds to the one given in (Allen, 1976, section 2).

Remark 58. Consider any extended probability space $(\Omega, \mathcal{F}, P^{ex})$. Call $\mathscr{B}$ the sigma-algebra generated by the events B in Ω that we can enter a bet about, that is, the sigma-algebra generated by the collection $\{B \subset \Omega : P^{ex}(B) \geqslant 0\}$. Then, a Dutch book is a finite collection $\{B_j\}_{j=1}^n \subset \mathscr{B}$ along with numbers $\{s_j\}_{j=1}^n \subset \mathbb{R}$ such that, for all $\omega \in \Omega$,

$$\mathfrak{f}(\omega) := \sum_{j=1}^{n} s_j \left[\mathbb{1}_{B_j}(\omega) - P^{ex}(B_j)\right] < 0. \tag{5.3}$$

Notice that the s_j's are the payout for a winning bet on B_j, and $s_j P^{ex}(B_j)$ is the bet's fair buy-in. The inequality in (5.3) suggests that a bet can be built such that the bookmaker always gets a profit, while the punter always loses money.

Definition 59. Extended probability P^{ex} is coherent if no Dutch books can be made against the punter.

Then, we have the following important result.

Theorem 60. P^{ex} is always coherent.

In our definition of Dutch book, we do not take into account the events $B' \in \mathcal{F}$ for which $P^{ex}(B') < 0$ since those are events for which the punter cannot enter a bet. Hence, they cannot be used to build a Dutch book.

5.2.3 Related literature

Let us now inspect how the definition and the interpretation of extended probabilities we have given so far relates to the existing literature.

Our framework can be seen as an extension to the one studied in (Burgin, 2012). He studies a static environment (as opposed to the dynamic one we inspect in this chapter in the next section) where the state space Ω can be divided in two irreducible parts Ω^+ and Ω^- such that $\#\Omega^+ = \#\Omega^-$, where $\#$ denotes the cardinality operator.

The elements of Ω^- are called anti-events, and they are usually connected to negative objects: encountering a negative object is a negative event. The author then states that an example of a negative object is given by antiparticles, the antimatter counterpart of quantum particles. Anti-events are given negative probabilities. If we require that the following conditions hold, we obtain Burgin's framework:

- extended probabilities are finitely additive (instead of countably additive);
- our state space can be divided in two irreducible parts Ω^+ and Ω^-;
- $P^{ex}(A) \geqslant 0$ if and only if $A \subset \Omega^+$;
- there exists a function $\alpha : \Omega \to \Omega$ such that $\alpha(\omega) = -\omega$ and $\alpha^2(\omega) = \omega$;
- $P^{ex}(\Omega^+) = 1$;
- $\{v_i, \omega, -\omega : v_i, \omega \in \Omega,\, i \in I\} = \{v_i, : v_i \in \Omega,\, i \in I\}$, for all $\omega \in \Omega$ and all set of indices I.

To this extent, our setup can be seen as a generalization of Burgin's one. Our interpretation of negative probabilities reconciles with Burgin's one if we consider anti-events as events for which we cannot enter a bet, which seems a reasonable assumption. He also gives a frequentist interpretation of negative probabilities, that complements our interpretation of negative extended probabilities.

In a subsequent paper, (Burgin, 2013) generalizes his own setup, mainly by not requiring $\#\Omega^+ = \#\Omega^-$ and by allowing any event to have either positive or negative probability, depending on external conditions. If we require that the following conditions hold, we obtain Burgin's generalized framework:

- extended probabilities are finitely additive;
- $A \in \mathcal{F}$ implies $-A := \{-\omega : \omega \in A\} \in \mathcal{F}$;

- for all $A \in \mathcal{F}$, $P^{ex}(-A) = -P^{ex}(A)$.

Notice that this last condition is very similar to Allen's principle of conservation of knowledge. The interpretation he gives for negative probabilities is similar to the one given in his previous work, with the peculiarity that now negative probabilities are not assigned exclusively to anti-events. Our interpretation can be seen as being a step behind Burgin's one: we give a specific reason for why an event $\tilde{A}$ is assigned a negative probability, namely that we cannot enter a bet on it. Then, Burgin states that the anti-event of $\tilde{A}$, $-\tilde{A}$, will have a positive (regular) probability. Of course, the vice versa holds: if an event is assigned a positive probability (because we can enter the bet), then according to Burgin its anti-event will have a negative probability.

Another interesting interpretation is given in (Székely, 2005). The author proves that we can encounter a negative probability if we work with a random variable having a signed distribution. In addition, if X has a signed distribution, then there exist two random variables Y, Z having an ordinary (not signed) distribution such that $X + Y = Z$ in distribution. Therefore X can be seen as a "difference" of two ordinary random variables Z and Y. We reconcile our interpretation and Székely's one as follows. A signed distribution P^s is simply an extended probability on the space of outcomes of a random variable. A pullback argument can then be used to define an extended probability P^{ex} on Ω: $P^{ex}(X^{-1}(I)) = P^s(I)$, for all subsets I of the outcome space of random variable X. So there are going to be events in Ω having negative probabilities: those are events we cannot enter a bet about.

The interpretation (Kronz, 2007) gives of negative probabilities is the following: he calls negative probabilities inferred probabilities, that can only be obtained indirectly by inference from operational (regular) probabilities. The associated events (that is, events having negative probabilities) are called virtual events in that they are non-operational, and so do not give rise to directly accessible relative frequencies.

Actual events have non-negligible effect on them, and the reverse is also true. As it appears clear, virtual events are equivalent to latent events e.g. in the psychology (Bollen, 2002), economics (Hu, 2017) and medicine (Rabe-Hesketh and Skrondal, 2008) literatures: Kronz assigns negative probabilities to latent events. These latter are events we do not observe, so it is fair to think we cannot enter a bet involving them. In this respect, our interpretation can be reconciled with Kronz's one.

In (Benavoli et al., 2021), the authors show how negative probabilities arise when an agent would like to enter a given bet involving an event, but computational limitations related to the event prevent them from doing so. This interpretation coincides with the one we give in the present chapter: they give a reason for which an agent is denied the possibility of entering a bet.

Finally, the interpretation of negative probabilities in (Gell-Mann and Hartle, 2012) and (Hartle, 2004, 2008) is very similar – although not identical – to the one in the present chapter. We give the summary of their way of interpreting negative probabilities as reported in (Feintzeig and Fletcher, 2017). First, they point out that the probabilities dictated by a physical theory instruct (rational) agents on how to bet on the outcomes of phenomena. Standard arguments from subjective Bayesian probability theory – the Dutch book arguments – demand that the Kolmogorovian axioms for classical probability theory must hold for the degrees of belief of any rational agent, which determine which bets the agent regard as fair. However, they point out that these arguments only apply to bets which are settleable, that is, bets about events the agent will certainly know at some point. They then argue that only bets on sufficiently coarse-grained alternative histories will be settleable, where this coarse-graining guarantees that the alternatives' probabilities will lie in the unit interval. The main difference between their interpretation and the one presented in this chapter is that for Gell-Mann and Hartle the bets on events that have negative probabilities cannot be settled, whereas we deem those bets to be settlable. The agent

simply cannot enter them at the moment, but they may be given the opportunity in the future.

5.3 Extended probabilities in statistical inference

In this section, we are going to consider an ex ante analysis in which we progressively learn the composition of the state space. It is ex ante in that it takes place before the actual statistical analysis. As time goes by, we collect new observations via a learning procedure specified in advance, e.g. an urn with or without replacement.

Consider a measurable space $(\Omega, \mathcal{F})$ with Ω at most countable and $\mathcal{F} = 2^{\Omega}$. At any time t, we have $\Omega = \Omega_t^- \sqcup \Omega_t^+$. For all $t \in \mathbb{N}_0 := \mathbb{N} \cup \{0\}$, Ω_t^- represents the "latent" part of Ω at time t, that is, the part that we have not yet observed. Ω_t^+ represents the "actual" part of Ω, that is, the portion of Ω that we have observed at time t (at time $t = 0$, Ω_0^+ is the portion of Ω that we know ex ante, which is assumed nonempty). This approach is similar to the "ex ante humility" one introduced in (Allen, 1976, section 1), where latent and actual portions of the state space are first introduced, and a dynamic is described. A graphical representation of $\Omega = \Omega_t^- \sqcup \Omega_t^+$ is given in Figure 5.1. At time $t = 0$, we specify the finest possible partition of Ω,

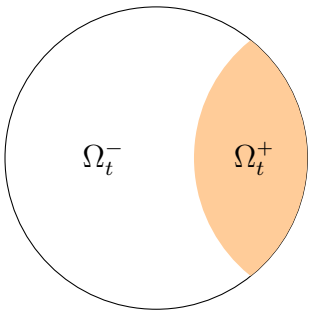

FIGURE 5.1: Graphical representation of $\Omega = \Omega_t^- \sqcup \Omega_t^+$.

$\mathcal{E} = \{\{\omega\}\}_{\omega \in \Omega}$. It is such that $\mathcal{E} = \mathcal{E}_0^- \sqcup \mathcal{E}_0^+$; in this notation, $\mathcal{E}_0^-$ partitions Ω_0^-, and $\mathcal{E}_0^+$ partitions Ω_0^+. The most general way of proceeding is by specifying the number type (natural numbers $\mathbb{N}$, whole numbers $\mathbb{N}_0$, integers $\mathbb{Z}$, or rationals $\mathbb{Q}$) we

subjectively believe we are working with, and setting it to be Ω. To this extent, we give to Ω the apparently possible interpretation of (Walley, 1991, section 2.1.2): Ω is the space of apparently possible states if it contains all the states ω that we believe are logically consistent with our available information. To give an example, a space of apparently possible states associated with a coin toss is

$\Omega = \{$heads, tails, coin landing on its edge, coin braking into pieces on landing,
coin disappearing down a crack in the floor$\}$.

Cosidering the space of apparently possible states, then, amounts to considering the most general state space associated with the statistical experiment of interest.

After that, we subjectively specify what we initially think the state space is, and we set it to be Ω_0^+; for example, we may have pregressed information coming from similar (but not equal) experiments.

Given the sequences (Ω_t^+) and (Ω_t^-), we require that, for all t,

$$\Omega_t^+ \subset \Omega_{t+1}^+ \quad \text{and} \quad \Omega_t^- \supset \Omega_{t+1}^-.$$

This means that (Ω_t^+) is monotone nondecreasing, and (Ω_t^-) is monotone nonincreasing, which implies that the limits of both exist and are well defined. In particular, we have that

$$\lim_{t\to\infty} \Omega_t^+ = \bigcup_{t\in\mathbb{N}_0} \Omega_t^+ \quad \text{and} \quad \lim_{t\to\infty} \Omega_t^- = \bigcap_{t\in\mathbb{N}_0} \Omega_t^-.$$

We then have two possible scenarios. In the first one, $\bigcup_{t\in\mathbb{N}_0} \Omega_t^+ = \Omega$, which implies that $\bigcap_{t\in\mathbb{N}_0} \Omega_t^- = \varnothing$. In the second one, $\bigcup_{t\in\mathbb{N}_0} \Omega_t^+ \equiv \Omega' \subsetneq \Omega$, which implies that $\bigcap_{t\in\mathbb{N}_0} \Omega_t^- \equiv \Omega'' := \Omega\backslash\Omega' \neq \varnothing$.

As we collect more and more observations, we progressively discover the composition of our sample space. In the first scenario, in the limit we discover that the sample space associated with our experiment corresponds to the whole set Ω we initially specified. In this scenario, at time $t+1$ when an element $\tilde{\omega}$ is observed

that belongs to Ω_t^-, the element $E_{t,\tilde{\omega}}^- = \{\tilde{\omega}\}$ of the partition $\mathcal{E}_t^-$ consistent with $\tilde{\omega}$ becomes an element of $\mathcal{E}_{t+1}^+$. We use the following notation: $E_{t+1,\tilde{\omega}}^+ = E_{t,\tilde{\omega}}^-$. So the partition of the latent space loses an element, while the partition of the actual space gains one. This means that, for all t, $\mathcal{E}_t^+ \subset \mathcal{E}_{t+1}^+$ and $\mathcal{E}_t^- \supset \mathcal{E}_{t+1}^-$. We abuse notation: we write $\mathcal{E}_t^+ \subset \mathcal{E}_{t+1}^+$ to indicate that $\#\mathcal{E}_t^+ \leqslant \#\mathcal{E}_{t+1}^+$, where $\#\mathcal{E}_t^+$ denotes the number of elements of the partition $\mathcal{E}_t^+$, for all t, and we write $\mathcal{E}_t^- \supset \mathcal{E}_{t+1}^-$ to indicate that $\#\mathcal{E}_t^- \geqslant \#\mathcal{E}_{t+1}^-$. Hence, in the limit, the partition $\mathcal{E}_t^+$ of the "actual" space coincides with the partition $\mathcal{E}$ of the whole set Ω. This mirrors the behavior of Ω_t^+, that converges to Ω.

In the second scenario Ω' can be finite or countable. In the former case, there exists a $T \in \mathbb{N}$ after which the observations we collect at time $T + i$ already belong to Ω_T^+, for all $i \in \mathbb{N}$. If that is the case, we write

$$\Omega' \equiv \Omega_T^+. \tag{5.4}$$

This corresponds to discovering that the actual sample space is finite and smaller than the whole set Ω that we specified at the beginning of our analysis. If Ω' is countable, we can only say that it is a proper subset of Ω, and that we discover its composition in the limit. This may happen, for example, if we begin our analysis by setting $\Omega = \mathbb{N}_0$, but then realize that the state space associated with our experiment is actually $\Omega' = \mathbb{N}$.

In general, when the second scenario takes place, we have $\Omega_t^+ \uparrow \Omega'$, and $\Omega_t^- \downarrow \Omega'' := \Omega \backslash \Omega'$. We also have that, in the limit, the partition $\mathcal{E}_t^+$ of the "actual" space coincides with the (finest possible) partition $\mathcal{E}'$ of set Ω'. We call $\mathcal{E}''$ the partition whose elements belong to $\mathcal{E}$ but not to $\mathcal{E}'$; again abusing notation, we write $\mathcal{E}'' := \mathcal{E} \backslash \mathcal{E}'$. We call $\mathcal{F}' = 2^{\Omega'}$ and $\mathcal{F}'' = 2^{\Omega''}$.

Remark 61. In this chapter, our ex ante analysis is concluded by finding the true composition of the state space. We can interpret this using the concept of a benev-

olent bookmaker. While our doppelgänger is always able to enter a bet on all the events in the space Ω of apparently possible states, we end up only able to bet on the events of Ω that are crucial to the statistical analysis taking place after our ex ante analysis. This is akin to a benevolent bookmaker preventing us from entering bets on irrelevant events, that, if we were to bet on, would certainly make us lose money.

However, there may be cases in which the state space does not get fully discovered. For example, this may happen if we have an urn with a timer attached; once the time runs out, the urn is sealed, so that we do not discover its entire composition. The following statistical analysis, which would require the use of extended probabilities, is explored in a future work. This case corresponds to the existence of a malevolent bookmaker that does not allow us to enter bets that are crucial to the statistical analysis. This setting is especially important when studying events about which we will never be able to enter a bet, for example latent events.

Remark 62. If at any time t we collect an observation $\breve{\omega}$ that does not belong to Ω, this means that the space Ω we initially specified is not rich enough. We have then to specify a richer, larger set $\breve{\Omega} \supset \Omega$, and start our analysis over. We can either consider $\breve{\Omega} = \Omega \cup \{\breve{\omega}\}$, or define $\breve{\Omega}$ as a larger number type. Let us give an example. Suppose we begin our analysis by setting $\Omega = \mathbb{N}$, and after a while we observe $\breve{\omega} = 1/2$. Then, we have to restart our analysis and we can either let $\breve{\Omega}$ be $\mathbb{N} \cup \{1/2\}$ or $\breve{\Omega} = \mathbb{Q}$.

It may also happen that our true sample space is $\breve{\Omega} \subsetneq \Omega_0^+$. In this case, equation (5.4) holds with $T = 0$, and our analysis would still be valid. The drawback is that, if that happens, we are not respecting one of the conditions listed in (Tsitsiklis, 2018) for a sample space to be valid. In particular, that the sample space Ω must have the right granularity depending on what we are interested in. This means that we must remove irrelevant information from the sample space. In other words, we must

choose the right abstraction and forget irrelevant information. This issue can be avoided by initially specifying Ω_0^+ so that it has the fewest possible elements.

On top of dealing with uncertainty on the composition of the sample space, we also address the problem of not being able to specify a unique (extended) probability measure on Ω. That is why we are going to work with sets of extended probability measures. We call this approach extended sensitivity analysis, since it corresponds to the extended probabilities counterpart of Bayesian sensitivity analysis (Berger, 1984). We begin by considering a set $\mathcal{P}^{ex} \equiv \mathcal{P}_0^{ex}$ of extended probabilities that represent the agent's initial beliefs, and we update it as described in section 5.3.2. We denote the sequence of successive updates of $\mathcal{P}_0^{ex}$ as $(\mathcal{P}_t^{ex})_{t\in\mathbb{N}}$. By working with sets of extended probability measures, we represent the condition of a researcher facing a decision under ambiguity. As we shall see in section 5.3.2, sequence $(\mathcal{P}_t^{ex})_{t\in\mathbb{N}_0}$ converges in the Hausdorff metric.

Now, fix any $t \in \mathbb{N}_0$, and consider any $P_t^{ex} \in \mathcal{P}_t^{ex}$. We require the following

(i) $P_t^{ex}(A) \in [0,1]$ if $A \subset \Omega_t^+$;

(ii) $P_t^{ex}(A) \in [-1,0]$ if $A \subset \Omega_t^-$;

(iii) $P_t^{ex}(A) \in [-1,1]$ if $A \cap \Omega_t^+ \neq \varnothing \neq A \cap \Omega_t^-$.

In particular, we compute the latter as follows.

Proposition 63. The following is true.

$$P_t^{ex}(A) = \sum_{E_{t,j}^+ \in \mathcal{E}_t^+ : P_t^{ex}(E_{t,j}^+) \neq 0} P_t^{ex}(A \mid E_{t,j}^+) P_t^{ex}(E_{t,j}^+) \tag{5.5}$$

$$+ \sum_{E_{t,j}^- \in \mathcal{E}_t^- : P_t^{ex}(E_{t,j}^-) \neq 0} P_t^{ex}(A \mid E_{t,j}^-) P_t^{ex}(E_{t,j}^-). \tag{5.6}$$

Notice that we need to consider the elements of $\mathcal{E}_t^+$ and $\mathcal{E}_t^-$ whose extended probabilities are not 0 otherwise the conditional extended probabilities $P_t^{ex}(A \mid E_{t,j}^+)$

and $P_t^{ex}(A \mid E_{t,j}^-)$ may result in an indeterminate form of the $\frac{0}{0}$ kind. This requirement yields no loss of generality since if an element of a partition is assigned extended probability 0, then it does not convey any information around event A. Notice also that, for the elements of $\mathcal{E}_t^+$ and $\mathcal{E}_t^-$ whose extended probabilities are not 0, $P_t^{ex}(A \mid E_{t,j}^+) = \frac{P_t^{ex}(A \cap E_{t,j}^+)}{P_t^{ex}(E_{t,j}^+)} \geqslant 0$ by (i), and $P_t^{ex}(A \mid E_{t,j}^-) = \frac{P_t^{ex}(A \cap E_{t,j}^-)}{P_t^{ex}(E_{t,j}^-)} \geqslant 0$ because it is the ratio of two negative quantities; once multiplied by $P_t^{ex}(E_{t,j}^-)$, which is negative by (ii), it gives us a negative value. So the sign of $P_t^{ex}(A)$ is not predetermined when $A \cap \Omega_t^+ \neq \varnothing \neq A \cap \Omega_t^-$. The interpretation of these conditions is straightforward: we assign negative extended probabilities to events that belong to the latent space at time t (meaning that at time t we cannot enter a bet about them), while we assign positive extended probabilities to events that are in the actual, observed space (meaning that at time t we can enter a bet about them). If a given event is only partially known, its extended probability has not a predetermined sign: it will depend on whether we know enough about it (then the probability will be positive), or not (vice versa). This means that we can enter a bet about "sub-event" $A \cap \Omega_t^+$, but not about $A \cap \Omega_t^-$. For example, let $A = \{\text{tomorrow there will be a thunderstorm}\}$. Then, for some t, suppose that

$$A \cap \Omega_t^+ = \{\text{tomorrow will rain}\},$$

$$A \cap \Omega_t^- = \{\text{tomorrow there will be a dry thunderstorm}\}.$$

Then we can place a bet on $A \cap \Omega_t^+$ but not on $A \cap \Omega_t^-$, so the extended probability we assign to A does not have a predetermined sign.

Notice also that we can define a set of events $\mathscr{C}_{P_t^{ex}} := \{A \in \mathcal{F} : P_t^{ex}(A \cap \Omega_t^-) = -P_t^{ex}(A \cap \Omega_t^+)\}$, which we call critical events according to P_t^{ex}, with the property that $P_t^{ex}(A) = 0$, for all $A \in \mathscr{C}_{P_t^{ex}}$ (immediate from the definition). We assign the sub-event we can enter a bet about the same probability that our doppelgänger assigns to the sub-event that we are not allowed to bet on. This means that we deem their

"actual" portion (the one we know/we have observed so far) to be just as likely than their "latent" portion (the one we do not know/we have not yet observed). The set of critical events according to the whole set of extended probability measures $\mathcal{P}_t^{ex}$ is given by

$$\mathscr{C}_t := \bigcap_{P_t^{ex} \in \mathcal{P}_t^{ex}} \mathscr{C}_{P_t^{ex}}.$$

5.3.1 Properties of this environment

We now give some results concerning the environment we depicted so far. Let $\mathcal{F}_t^+ = 2^{\Omega_t^+}$ and $\mathcal{F}_t^- = 2^{\Omega_t^-}$.

Proposition 64. For all t, we have that $\mathcal{F}_t^+ \subset \mathcal{F}$, $\mathcal{F}_t^- \subset \mathcal{F}$, and $\mathcal{F}_t^+ \cup \mathcal{F}_t^- = \mathcal{F}$.

We go on claming the following.

Proposition 65. Let $A_t^+ := A \cap \Omega_t^+$, $A_t^- := A \cap \Omega_t^-$, $B_t^+ := B \cap \Omega_t^+$, and $B_t^- := B \cap \Omega_t^-$. Then, $A \cup B = (A_t^+ \cup B_t^+) \cup (A_t^- \cup B_t^-)$ and $A \cap B = (A_t^+ \cap B_t^+) \cup (A_t^- \cap B_t^-)$, for all t. Also, $A \backslash B = (A_t^+ \backslash B_t^+) \cup (A_t^- \backslash B_t^-)$, for all t.

Recall now that a set ring is a system of sets $\mathbb{B}$ such that $A, B \in \mathbb{B}$ implies $A \cap B \in \mathbb{B}$ and $(A \backslash B) \cup (B \backslash A) =: A \triangle B \in \mathbb{B}$. A set ring $\mathbb{B}$ with a unit element, i.e. $E \in \mathbb{B}$ such that for all $A \in \mathbb{B}$, $A \cap E = A$, is called a set algebra.

Proposition 66. $\mathcal{F}_t^+$ and $\mathcal{F}_t^-$ are set algebras, for all t.

We also point out that if $A = \{\omega_1, \ldots, \omega_k\}$ and $\omega_1, \ldots, \omega_k \in \Omega$, then $P^{ex}(A) = \sum_{j=1}^k P^{ex}(\{\omega_j\})$. This is immediate from the countable additivity of extended probabilities.

Another property is the following. Fix any t and let $A, B \in \mathcal{F}_t^+$ such that $A \subset B$. Then, $P_t^{ex}(A) \leqslant P_t^{ex}(B)$. Let then $C, D \in \mathcal{F}_t^-$ such that $C \subset D$. Then, $P_t^{ex}(C) \geqslant P_t^{ex}(D)$. Both these results come from Proposition 55.

The following is also interesting.

Proposition 67. Consider $A \subset \cup_{j\in\mathbb{N}_0} A_j$, where $A_j \in \mathcal{F}$ for all j, and also $A \in \mathcal{F}$. Then $P_t^{ex}(A) \leqslant \sum_{j\in\mathbb{N}_0} P_t^{ex}(A_j)$ if $\cup_{j\in\mathbb{N}_0} A_j \in \mathcal{F}_t^+$, and $P_t^{ex}(A) \geqslant \sum_{j\in\mathbb{N}_0} P_t^{ex}(A_j)$ if $\cup_{j\in\mathbb{N}_0} A_j \in \mathcal{F}_t^-$. If instead some of the A_j's are in $\mathcal{F}_t^+$ and some are in $\mathcal{F}_t^-$, then $P_t^{ex}(A) \gtreqless \sum_{j\in\mathbb{N}_0} P^{ex}(A_j)$.

In the setting we have outlined so far, there is a way of operationalizing equation (5.1). Pick any $A \in \mathcal{F}$; we have

$$P_t^{ex}(A^c) = P_t^{ex}\left(\Omega_t^+ \backslash [\Omega_t^+ \cap A]\right) + P_t^{ex}\left(\Omega_t^- \backslash [\Omega_t^- \cap A]\right) \tag{5.7}$$

We retain the fact that $P_t^{ex}(\varnothing) = 0$; indeed

$$\begin{aligned} P_t^{ex}(\Omega^c) = P_t^{ex}(\varnothing) &= P_t^{ex}\left(\Omega_t^+ \backslash [\Omega_t^+ \cap \Omega]\right) + P_t^{ex}\left(\Omega_t^- \backslash [\Omega_t^- \cap \Omega]\right) \\ &= P_t^{ex}(\varnothing) + P_t^{ex}(\varnothing) \\ &\iff P_t^{ex}(\varnothing) = 0. \end{aligned}$$

We can also write $P_t^{ex}(A^c) = P_t^{ex}(A^c \cap \Omega_t^+) + P_t^{ex}(A^c \cap \Omega_t^-)$, because, as we know, $A^c = (A^c \cap \Omega_t^+) \sqcup (A^c \cap \Omega_t^-)$.

5.3.2 Interpretation and updating procedure

Let us now discuss the probability assigned to the whole sample space Ω. From (ii*), we know that since, for all t, $\Omega_t^+ \sqcup \Omega_t^- = \Omega$, then $P_t^{ex}(\Omega) = P_t^{ex}(\Omega_t^+) + P_t^{ex}(\Omega_t^-)$. Also, from (i) we know that $P_t^{ex}(\Omega_t^+) \geqslant 0$ and $P_t^{ex}(\Omega_t^-) \leqslant 0$. So

$$P_t^{ex}(\Omega) = \begin{cases} p > 0 & \text{if } P_t^{ex}(\Omega_t^+) > |P_t^{ex}(\Omega_t^-)| \\ p = 0 & \text{if } P_t^{ex}(\Omega_t^+) = |P_t^{ex}(\Omega_t^-)| \\ p < 0 & \text{if } P_t^{ex}(\Omega_t^+) < |P_t^{ex}(\Omega_t^-)| \end{cases}$$

What does it mean, then, for $P_t^{ex}(\Omega)$ to be equal to 0? And to be negative? And to be positive, but not 1? Given the benevolent bookmaker interpretation, we have the following. If $P_t^{ex}(\Omega) = 0$, it means that we have no sufficient information to say whether the sample space associated with our experiment is in fact Ω. If $P_t^{ex}(\Omega) < 0$,

it means that, for the time being, the sample space associated with our experiment appears to be some $\breve{\Omega} \subsetneq \Omega$. If $P_t^{ex}(\Omega) \in (0,1)$, it means that there is evidence that the sample space associated with our experiment could be in fact Ω, but we cannot state it with certainty.

The natural question that one might ask now is, for some $t \in \mathbb{N}_0$, how do we come up with negative numbers to assign to events that belong to the latent space Ω_t^-. Or, for that matter, how do we come up with positive numbers to assign to events that belong to the actual space Ω_t^+. Call $\Delta(\Omega, \mathcal{F})$ the set of probability measures on $(\Omega, \mathcal{F})$, and $\Delta^{ex}(\Omega, \mathcal{F})$ the set of extended probability measures on $(\Omega, \mathcal{F})$. The latter is a linear space, as shown in (Rao and Rao, 1983). In a future work, we will argue that it is a Dedekind complete Banach lattice with respect to the norm induced by the total variation of an element P^{ex} of $\Delta^{ex}(\Omega, \mathcal{F})$. We will also show that if Ω is a compact separable space, then the subset $\Delta_{Baire}^{ex}(\Omega, \mathcal{F}) \subset \Delta^{ex}(\Omega, \mathcal{F})$ of Baire extended probability measures is the dual of the real Banach space of all continuous real-valued functions on Ω.

Consider any $P \in \Delta(\Omega, \mathcal{F})$ such that, for all $\tilde{A} \in \mathcal{F}_0^+$, $P(\tilde{A}) = p \in [0,1]$ is the amount we deem fair to pay to enter a bet about $\tilde{A}$. Then, for a generic $A \in \mathcal{F}$ we have that

$$P_0^{ex}(A \cap \Omega_0^+) = P(A \cap \Omega_0^+) \tag{5.8}$$

and

$$P_0^{ex}(A \cap \Omega_0^-) = -P(A \cap \Omega_0^-). \tag{5.9}$$

It is immediate to see that P_0^{ex} satisfies (i*) and (ii*), so it is a properly defined extended probability measure. Notice also that

$$\sum_{E \in \mathcal{E}} |P_0^{ex}(E)| = \sum_{E_0^+ \in \mathcal{E}_0^+} P_0^{ex}(E_0^+) + \sum_{E_0^- \in \mathcal{E}_0^-} \left|P_0^{ex}(E_0^-)\right| = 1. \tag{5.10}$$

Clearly, (5.10) holds for all $t \in \mathbb{N}_0$, not just for $t = 0$.

This way of assessing initial extended probabilites well reconciles with the interpretation we gave in general for extended probabilities: we cannot enter bets about events $A \notin \mathcal{F}_0^+$ (in this case, because we cannot observe them for the time being). So we assess the probabilites of the events $A \in \mathcal{F}_0^+$ as specified in section 5.2, and then we ask our doppelgänger the probabilities $P_{0,D}(B) \geqslant 0$ they assign to the elements $B \in \mathcal{F}_0^-$.[2] Then, we flip the sign to those, that is, $P_0^{ex}(B) = -P_{0,D}(B)$, for all $B \in \mathcal{F}_0^-$. In this notation, $P_{0,D}(B)$ is the probability the doppelgänger assigns to event B at time $t = 0$.

Because the gent faces ambiguity, they need to specify a set $\mathcal{P} \subset \Delta(\Omega, \mathcal{F})$ of probability measures. Every element $P \in \mathcal{P}$ induces an extended probability P_0^{ex} as we just described. In this way, we build the set $\mathcal{P}_0^{ex}$.

Let us now discuss how to update extended probabilities, that is, how to update $P_t^{ex}(A)$ to $P_{t+1}^{ex}(A)$, for all $A \in \mathcal{F}$, for all $P_t^{ex} \in \mathcal{P}_t^{ex}$, for all $t \in \mathbb{N}_0$. We first consider a procedure to discover the components of the sample space that is equivalent to an urn without replacement. That is, after specifying Ω and Ω_0^+, we start our analysis with an urn whose content is unknown and possibly countable; it represents the true sample space associated with our experiment. At any time point t, we extract a ball (an element ω of the sample space). Once we learn about that element, our knowledge about the composition of the urn increases. We do not put the ball back into the urn; we discuss the case in which the discover procedure is equivalent to an urn with replacement later in the chapter.

Let us begin with the updating procedure from $P_0^{ex} \in \mathcal{P}_0^{ex}$ to $P_1^{ex} \in \mathcal{P}_1^{ex}$.

We collect a new observation $\omega \in \Omega$. If $\omega \in \Omega_0^-$, this means that, at time $t = 1$, we learn a new element of the true sample space. Notice that ω is consistent with an element $E_{0,\omega}^-$ of $\mathcal{E}_0^-$ (that is, $E_{0,\omega}^- = \{\omega\}$), which – given that we observe such ω – at time $t = 1$ becomes an element of $\mathcal{E}_1^+$; in formulas, $E_{0,\omega}^- = E_{1,\omega}^+$. Then, $\mathcal{E}_1^+ \supset \mathcal{E}_0^+$,

[2] Subscript D stands for "doppelgänger".

$\mathcal{E}_1^- \subset \mathcal{E}_0^-$, and we update the extended probability assigned to $E_{0,\omega}^- = E_{1,\omega}^+$ as follows

$$P_1^{ex}(E_{1,\omega}^+) = \left|P_0^{ex}(E_{0,\omega}^-)\right|. \tag{5.11}$$

The extended probabilities assigned to the other elements of $\mathcal{E}$ (the finest possible partition of the whole Ω) are held constant. From (5.5), the updated extended probability associated with event A is given by

$$\begin{aligned} P_1^{ex}(A) = & \sum_{E_{1,j}^+ \in \mathcal{E}_1^+ : P_1^{ex}(E_{1,j}^+) \neq 0} P_1^{ex}(A \mid E_{1,j}^+) P_1^{ex}(E_{1,j}^+) \\ & + \sum_{E_{1,j}^- \in \mathcal{E}_1^- : P_1^{ex}(E_{1,j}^-) \neq 0} P_1^{ex}(A \mid E_{1,j}^-) P_1^{ex}(E_{1,j}^-), \end{aligned} \tag{5.12}$$

where

$$P_1^{ex}(A \mid E_{1,j}^+) = \frac{P_1^{ex}(A \cap E_{1,j}^+)}{P_1^{ex}(E_{1,j}^+)}$$

and

$$P_1^{ex}(A \mid E_{1,j}^-) = \frac{P_1^{ex}(A \cap E_{1,j}^-)}{P_1^{ex}(E_{1,j}^-)}.$$

Notice that we are working with the finest possible partition of Ω, that is, $\mathcal{E} = \{\{\omega\}\}_{\omega \in \Omega}$. Then, for any $A \in \mathcal{F}$ and any $E \in \mathcal{E}$, $A \cap E = A \cap \{\omega\}$, which is equal to the empty set if $\omega \notin A$, and it is equal to $\{\omega\} = E$ if $\omega \in A$. Hence, for all $E \in \mathcal{E}$ such that $P_1^{ex}(E) \neq 0$,

$$P_1^{ex}(A \mid E) = \frac{P_1^{ex}(A \cap E)}{P_1^{ex}(E)} = \begin{cases} \frac{P_1^{ex}(\varnothing)}{P_1^{ex}(E)} = 0 & \text{if } A \cap E = \varnothing \\ \frac{P_1^{ex}(E)}{P_1^{ex}(E)} = 1 & \text{if } A \cap E = \{\omega\} \end{cases},$$

for all $E \in \mathcal{E}$.

So, equation (5.12) can be rewritten as

$$\begin{aligned} P_1^{ex}(A) = & \sum_{E_{1,j}^+ \in \mathcal{E}_1^+ : A \cap E_{1,j}^+ \neq \varnothing,\ P_1^{ex}(E_{1,j}^+) \neq 0} P_1^{ex}(E_{1,j}^+) \\ & + \sum_{E_{1,j}^- \in \mathcal{E}_1^- : A \cap E_{1,j}^- \neq \varnothing,\ P_1^{ex}(E_{1,j}^-) \neq 0} P_1^{ex}(E_{1,j}^-). \end{aligned} \tag{5.13}$$

Of course this holds for all $P_1^{ex} \in \mathcal{P}_1^{ex}$. Notice that we are implicitly assuming Allen's principle of conservation of knowledge. Indeed, suppose at time $\mathbf{t}$ the event A is entirely in the latent portion of Ω, that is, $A \cap \Omega_{\mathbf{t}}^- = A$, and at time $\mathbf{t}+k$ it is entirely in the actual portion of Ω, that is, $A \cap \Omega_{\mathbf{t}+k}^+ = A$. Then, by (5.11) and (5.13), we have that $P_{\mathbf{t}+k}^{ex}(A) = |P_{\mathbf{t}}^{ex}(A)|$. This has the practical advantage of making the updating procedure mechanical (we do not need to reassess any subjective extended probability when new observations become available), and of preserving the initial opinion of the researcher. The importance of this last point is discussed in Remark 70.

If instead we observe $\omega \in \Omega_0^+$, this means that we made a good job in specifying Ω_0^+, and so we keep the extended probability constant

$$P_1^{ex}(E_{1,\omega}^+) = P_0^{ex}(E_{0,\omega}^+) \geqslant 0. \tag{5.14}$$

Of course, the extended probabilities assigned to the other elements of $\mathcal{E}$ remain constant as well.

We follow the procedures in (5.11) and (5.14) to update P_t^{ex} to P_{t+1}^{ex} for all $t \in \mathbb{N}_0$, not just from $t = 0$ to $t = 1$. We call (P_t^{ex}) the sequence of updates of initial extended probability P_0^{ex}. Now consider the extended probability $P_\infty^{ex} \in \Delta^{ex}(\Omega, \mathcal{F})$ such that

$$P_\infty^{ex}\left(A \cap \bigcup_{t\in\mathbb{N}_0} \Omega_t^+\right) = P\left(A \cap \bigcup_{t\in\mathbb{N}_0} \Omega_t^+\right)$$

and

$$P_\infty^{ex}\left(A \cap \bigcap_{t\in\mathbb{N}_0} \Omega_t^-\right) = -P\left(A \cap \bigcap_{t\in\mathbb{N}_0} \Omega_t^-\right),$$

for all $A \in \mathcal{F}$, where P is the regular probability measure we used in (5.8) and (5.9) to specify P_0^{ex}. Let us denote by d_{ETV} the extended total variation distance,

$$d_{ETV}(P^{ex}, Q^{ex}) := \sup_{A\in\mathcal{F}} |P^{ex}(A) - Q^{ex}(A)|.$$

It is routine to check that d_{ETV} is a metric: the proof goes along the lines of showing that the total variation distance is a metric. Then, the following holds.

Proposition 68. $P_t^{ex} \to P_\infty^{ex}$ as $t \to \infty$ in the extended total variation metric.

The following claim is especially important.

Proposition 69. If $\Omega_t^+ \uparrow \Omega$, then $P_\infty^{ex}(\Omega) = 1$.

This result implies that if $\Omega_t^+ \uparrow \Omega$, then P_∞^{ex} is a regular probability measure. Indeed, it is easy to see that it satisfies the Kolmogorovian axioms for regular probability measures.

Now call $\mathcal{P}_\infty^{ex} \subset \Delta^{ex}(\Omega, \mathcal{F})$ the following set

$$\mathcal{P}_\infty^{ex} := \left\{ P_\infty^{ex} \in \Delta^{ex}(\Omega, \mathcal{F}) : d_{ETV}(P_t^{ex}, P_\infty^{ex}) \xrightarrow[t\to\infty]{} 0,\ P_t^{ex} \in \mathcal{P}_t^{ex} \right\}.$$

That is, $\mathcal{P}_\infty^{ex}$ is the set of limits (in the extended total variation metric) of the sequences (P_t^{ex}) whose elements P_t^{ex} belong to $\mathcal{P}_t^{ex}$.

The Hausdorff distance between an element of $(\mathcal{P}_t^{ex})_{t\in\mathbb{N}_0}$ and $\mathcal{P}_\infty^{ex}$ is given by

$$\begin{aligned} d_H\left(\mathcal{P}_t^{ex}, \mathcal{P}_\infty^{ex}\right) = \max \Bigg\{ & \sup_{P_t^{ex} \in \mathcal{P}_t^{ex}} \inf_{P_\infty^{ex} \in \mathcal{P}_\infty^{ex}} d_{ETV}\left(P_t^{ex}, P_\infty^{ex}\right), \\ & \sup_{P_\infty^{ex} \in \mathcal{P}_\infty^{ex}} \inf_{P_t^{ex} \in \mathcal{P}_t^{ex}} d_{ETV}\left(P_t^{ex}, P_\infty^{ex}\right) \Bigg\}. \end{aligned} \tag{5.15}$$

Then, the sequence $(\mathcal{P}_t^{ex})_{t\in\mathbb{N}_0}$ of successive updates of set $\mathcal{P}_0^{ex}$ representing the initial beliefs of the agent facing ambiguity converges in the Hausdorff distance to $\mathcal{P}_\infty^{ex}$. This result is an immediate consequence of Proposition 68: every element of $\mathcal{P}_t^{ex}$ converges to an element of $\mathcal{P}_\infty^{ex}$, so the distance between the "borders" of these two sets – measured by the Hausdorff metric – converges to 0.

Now, there are two possible scenarios: one in which we continue discovering the elements of the state space until we retrieve the full Ω we specified ex ante (this

corresponds to $\Omega_t^+ \uparrow \Omega$), and another one in which we discover that the actual sample space associated with our experiment is $\Omega' \subsetneq \Omega$. In the first scenario, any $P_\infty^{ex} \in \mathcal{P}_\infty^{ex}$ is a regular probability measure, a consequence of Proposition 69. So, after discovering the composition of the state space, we have that $(\Omega, \mathcal{F}, P_\infty^{ex})$ is a regular probability space, for all $P_\infty^{ex} \in \mathcal{P}_\infty^{ex}$.

In the second scenario, Ω' can be finite or countable. In the former case, we have that, for some $T \in \mathbb{N}$, $\Omega_T^+ \equiv \Omega' \subsetneq \Omega$, so $\sum_{E_T^+ \in \mathcal{E}_T^+} P_T^{ex}(E_T^+) = q < 1$ because, by (5.10) and (5.11),

$$\sum_{E_T^+ \in \mathcal{E}_T^+} P_T^{ex}(E_T^+) + \sum_{E_T^- \in \mathcal{E}_T^-} \left| P_T^{ex}(E_T^-) \right| = 1.$$

This may seem problematic: $\mathcal{P}_t^{ex}$ converges to $\mathcal{P}_T^{ex}$ (that is, $\mathcal{P}_\infty^{ex}$ coincides with $\mathcal{P}_T^{ex}$), which is not a set of regular probability measures. To solve this issue, we need to describe the regular probability measure induced by every $P_T^{ex} \in \mathcal{P}_T^{ex}$. It would be desirable to find $\tilde{P}$ such that $\tilde{P}(A) = cP_T^{ex}(A)$, for all $A \in \mathcal{F}_T^+$. This because such a regular probability measure preserves the ratios between extended probabilities of the elements of $\mathcal{F}_T^+$, that is,

$$\frac{\tilde{P}(A)}{\tilde{P}(B)} = \frac{cP_T^{ex}(A)}{cP_T^{ex}(B)} = \frac{P_T^{ex}(A)}{P_T^{ex}(B)},$$

for all $A, B \in \mathcal{F}_T^+$ such that $P_T^{ex}(B) \neq 0$. To find such a c, the following needs to hold

$$\sum_{E_T^+ \in \mathcal{E}_T^+} cP_T^{ex}(E_T^+) = 1,$$

which happens if and only if

$$c = \frac{1}{\sum_{E_T^+ \in \mathcal{E}_T^+} P_T^{ex}(E_T^+)} = \frac{1}{P_T^{ex}(\Omega_T^+)}.$$

Hence, $\tilde{P} = cP_T^{ex}$ is the regular probability measure induced by P_T^{ex} that preserves

the ratios between extended probabilities of elements of $\mathcal{F}_T^+$. Clearly, this holds for all $P_T^{ex} \in \mathcal{P}_T^{ex}$, so $(\Omega_T^+, \mathcal{F}_T^+, \tilde{P})$ is a regular probability space, for all $\tilde{P} \in \tilde{\mathcal{P}}$, where $\tilde{\mathcal{P}}$ is the set of regular probabilities induced by the elements of $\mathcal{P}_T^{ex}$.

If Ω' is countable, we proceed in a similar way. We still want to preserve the ratios between extended probabilities of the elements of $\mathcal{F}'$, so we have to find c such that $\sum_{E \in \mathcal{E}'} cP_\infty^{ex}(E) = 1$. This happens when

$$c = \frac{1}{\sum_{E \in \mathcal{E}'} P_\infty^{ex}(E)} = \frac{1}{P_\infty^{ex}(\Omega')},$$

so $\tilde{P} = cP_\infty^{ex}$ is the regular probability measure we were looking for. This holds for all $P_\infty^{ex} \in \mathcal{P}_\infty^{ex}$, so $(\Omega', \mathcal{F}', \tilde{P})$ is a regular probability space, for all $\tilde{P} \in \tilde{\mathcal{P}}$.

Remark 70. Notice that, in the first scenario ($\Omega_t^+ \uparrow \Omega$), every element $P_\infty^{ex} \in \mathcal{P}_\infty^{ex}$ coincides with its corresponding probability measure $P \in \mathcal{P}$ we expressed ex ante on the whole Ω. Indeed, we have that $P_\infty^{ex}(E) = |P_\infty^{ex}(E)|$, for all $E \in \mathcal{E}$, which implies $P_\infty^{ex}(A) = |P_\infty^{ex}(A)|$, for all $A \in \mathcal{F}$. But then, by (5.10), we know that $\sum_{E \in \mathcal{E}} P_\infty^{ex}(E) = 1$, and that $P_\infty^{ex}(A) = P(A)$, for all $A \in \mathcal{F}$. This should not surprise: the updating procedure we described earlier is not based on collecting new data like the Bayesian one, nor on repeating the experiment many times like the frequentist one. Rather, it is based on discovering the true composition of the state space associated with our inference procedure. Then, it is natural to retrieve the opinion we expressed ex ante on the state space Ω once we get the confirmation that the true state space is indeed Ω. This also reconciles well with the interpretation we gave to negative extended probabilities: once we are given the possibility to enter the bets we were denied before, we tend to agree with our doppelgänger who had the opportunity to bet on those events in the first place (Allen's principle of conservation of knowledge holds).

Notice also that $\tilde{P}$, albeit not equal, is proportional to P_∞^{ex}, $\tilde{P} = cP_\infty^{ex}$, for every

$\tilde{P} \in \tilde{\mathcal{P}}$ and its corresponding $P_\infty^{ex} \in \mathcal{P}_\infty^{ex}$. The interpretation is immediate: we maintain our opinion on the elements of the sample space $\Omega' \subsetneq \Omega$ that pertains to our experiment.

If the procedure for discovering the composition of the state space is equivalent to an urn with replacement, everything we discussed so far still holds with just two differences:

- if we extract twice or more times the same element, its extended probability flips sign the first time, and then stays constant;
- the convergence may be slower, because we may need more extractions from the urn to learn the true composition (because we can extract twice or more times the same element).

Let us give a simple example that illustrates one of the possible situations in which the analysis depicted so far can be put to use.

Example 71. We consider a species sampling problem in the field of ecology. Suppose we want to know the number of bird species that inhabit a certain region throughout the year. What we can do is to start our analysis by letting $\Omega = \mathbb{N}$ and $\Omega_0^+ = \{1, \ldots, n\}$, where n is the number of species that inhabit a region similar to the one of interest throughout the year. Then, we specify a set $\mathcal{P}$ of probability measures on $\mathbb{N}$, e.g. a collection $\{\text{Geom}(p)\}_{p\in[0,1]}$ of geometric distributions having parameter $p \in [0, 1]$. We specify a set of probabilities because we are not able to express our initial opinion via a unique probability measure (we face ambiguity). We specify the probability measures on the whole number field $\mathbb{N}$ because we do not know exactly the composition of our state space. Then, after eliciting the set $\mathcal{P}_0^{ex}$ of extended probability measures induced by (the elements of) $\mathcal{P}$, we begin the ex ante analysis described in the present chapter. After collecting observations for an entire

year, we end up discovering that the state space associated with our experiment is $\Omega' = \{1, \ldots, m\} \subsetneq \mathbb{N}$, where $m \geqslant n$ is the number of species we counted during the year. We recover also the set $\tilde{\mathcal{P}}$ of regular probability measures induced by $\mathcal{P}_\infty^{ex}$. Now, the "real" statistical analysis can take place. Indeed, we know that m is the maximum number of species that live in the region during the year, but it does not take into account migrations to or from the region itself. Hence, the number of bird species that inhabit the region throughout the year may well be smaller than m. So, every $\tilde{P} \in \tilde{\mathcal{P}}$ will be such that $\tilde{P}(\{\omega_k\}) \in [0, 1]$, and $\sum_{k=1}^{m} \tilde{P}(\{\omega_k\}) = 1$, where $\omega_k = k$, for all $k \in \{1, \ldots, m\}$. Gathering data $y_1, \ldots, y_\ell$ and updating these $\tilde{P}$'s via Bayesian conditioning, we obtain the set $\{\tilde{P}(\cdot \mid y_1, \ldots, y_\ell)\}_{\tilde{P} \in \tilde{\mathcal{P}}}$ of posterior (regular) probability measures. This gives a robust analysis: for all $\omega_k \in \Omega'$, the posterior probability of ω_k being the correct number of species belongs to the interval

$$\left[\underline{\tilde{P}}(\{\omega_k\} \mid y_1, \ldots, y_\ell), \overline{\tilde{P}}(\{\omega_k\} \mid y_1, \ldots, y_\ell)\right],$$

where the lower bound is a lower (regular) probability, and the upper bound is an upper (regular) probability. The narrower the interval, the less imprecise our beliefs resulting from the analysis. $\triangle$

A criticism that can be made of our ex ante analysis is the following: why should the scholar be concerned with the exact composition of the state space? It should be enough to specify it as richly as possible, and then proceed to a regular statistical analysis. There are two responses to such a critique. First, as pointed out in Remark 62, in doing so the scholar would not be respecting the condition listed in (Tsitsiklis, 2018) that the state space must have the right granularity depending on the statistical experiment they are interested in. Second, in the case that the state space associated with our statistical experiment is $\Omega' \subsetneq \Omega$, working with probability measures supported on the whole space Ω of apparently possible states may

be computationally costly. Our ex ante analysis allows the scholar to focus on the "minimal" state space – the one containing only the necessary states – that is more meaningful to the analysis.

5.4 Application to opinion dynamics

This example comes from the model in (Allahverdyan and Galstyan, 2014). There, opinion dynamics between a persuaded and a persuading agents is studied, in particular when the persuaded agent evaluates new information in a way that is consistent with her own preexisting belief. In this example we are going to adoperate extended probabilities to model the boomerang effect. This phenomenon corresponds to the empirical observation that sometimes persuasion yields the opposite effect: the persuaded agents moves her opinion away from the opinion of the persuading agent. That is, she enforces her old opinion.

In (Allahverdyan and Galstyan, 2014), the authors assume that the state of the world does not change, that the agents are aware of this fact, and that the persuaded agent changes her opinion only under the influence of the opinion of the persuading agent. They model this dynamic in a linear fashion. The first iteration is the following

$$\hat{P}_1(\{\omega_k\}) = \epsilon P_0(\{\omega_k\}) + (1-\epsilon)Q(\{\omega_k\}), \quad \epsilon \in [0,1], \tag{5.16}$$

where $\omega_k \in \Omega$, a finite sample space, and P_0 and Q are regular probability measures that represent the initial opinions of the persuaded and the persuading agents, respectively. Q is not indicized to time because they make the simplifying assumption that the persuading agent does not change his mind: he tries to persuade the other agent of the same thing at every iteration. ϵ is a weight, and several qualitative factors contribute to its subjective assessment: egocentric attitude of the persuaded agent, the fact that the persuaded agent has access to internal reasons for choosing her opinion, while she is not aware of the internal reasons of the persuading agent,

and many more. The authors relate ϵ to the credibility of the persuading agent: the higher ϵ, the less credible he is. The successive iterations are modeled as follows

$$\hat{P}_{t+1}(\{\omega_k\}) = \epsilon \hat{P}_t(\{\omega_k\}) + (1-\epsilon)Q(\{\omega_k\}), \tag{5.17}$$

for $t \geqslant 1$. This means that at every iteration, the persuading agent tries to shift the persuaded agent's opinion closer to his own. This continues until either the persuaded agent is content with her opinion (and hence does not further change her beliefs), or the persuading agent completely convinces the other agent.

Now, one way to model the boomerang effect is to consider $\epsilon > 1$. This conveys the idea that the persuading agent has an extremely low credibility. This results in the updated opinion of the persuaded agent to be an extended probability measure. Indeed, for $\epsilon > 1$, $\hat{P}_{t+1}(\{\omega_k\})$ is negative whenever $\epsilon \hat{P}_t(\{\omega_k\}) < |1-\epsilon| Q(\{\omega_k\})$. (Allahverdyan and Galstyan, 2014) consider the induced regular probability measure to avoid working with extended probabilities.

We modify slightly the linear model in (Allahverdyan and Galstyan, 2014). The main differences are three: we allow the use of extended probabilities, we describe a state space that is divided in latent and known, and whose composition is gradually discovered as the time passes by, and we do not let ϵ be a free parameter. It depends on both the element ω_k of the state space we examine and on the iteration t we are considering.

Mathematically, we can describe the low credibility of the persuading agent through hidden states: the persuaded agent may think that she does not know the state space well enough, that is, that there are some hidden portions of Ω she is not (yet) aware of.

We assume that the state space $\Omega = \{\omega_1, \ldots, \omega_N\}$ is a finite set (a common simplifying assumption in opinion dynamics). Notice that the finest possible partition of Ω is $\mathcal{E} = \{\{\omega_j\}\}_{j=1}^N$. At the beginning of the interaction between persuading and

persuaded agents, the former is fully aware of the composition of the state space, while the latter is only aware of the composition of $\Omega_0^+ \subsetneq \Omega$, but suspects that the actual state space is larger. This corresponds to having suspects on the persuading agent hiding some pieces of information. In particular, she correctly guesses that the true state space is Ω. This correct guess is without loss of generality for our analysis: we are in the $\Omega_t^+ \uparrow \Omega$ case; we also assume that the discovering procedure is equivalent to an urn without replacement. She then defines an extended probability on Ω the way we explained in section 5.3. That is, she specifies a probability distribution P on the whole Ω and then flips the sign to the probabilities of the latent events: $P_0^{ex}(\{\omega_k\}) = P(\{\omega_k\})$ if $\omega_k \in \Omega_0^+$, and $P_0^{ex}(\{\omega_k\}) = -P(\{\omega_k\})$ if $\omega_k \notin \Omega_0^+$.

To obtain the influenced extended probability at any time $t \geqslant 0$, we modify slightly equation (5.17) to get

$$\hat{P}_t^{ex}(\{\omega_k\}) = \epsilon_{k,t} P_t^{ex}(\{\omega_k\}) + (1 - \epsilon_{k,t}) Q(\{\omega_k\}). \tag{5.18}$$

The nonnegative $\epsilon_{k,t}$'s have to be chosen such that $\hat{P}_t^{ex}$ is an extended probability measure, that is, $\hat{P}_t^{ex}(\{\omega_k\}) \in [-1, 1]$ for all k, for all t, and

$$\hat{P}_t^{ex}(\Omega) = \sum_{k=1}^{N} \hat{P}_t^{ex}(\{\omega_k\}) \leqslant 1.$$

Here, $\hat{P}_t^{ex}$ denotes the influenced extended probability at time t. Notice that $\hat{P}_t^{ex}$ is analytically similar to an ϵ-contaminated probability measure. There are of course two major differences: for some ω_k, $\epsilon_{k,t}$ is greater than 1, and also one of the elements of the mixture is an extended probability measure (rather than a regular one). Notice also that Q is not indicized to time because we too make the simplifying assumption that the persuading agent does not change his mind. Another characteristic worth noting is that in our model, at every iteration t, the persuaded agent combines her updated belief (expressed via the extended probability P_t^{ex}) with the other agent's

belief to obtain $\hat{P}_t^{ex}$. This is different from the model in (Allahverdyan and Galstyan, 2014) where at every iteration t the persuaded agent combines her influenced belief at iteration $t-1$ (expressed through $\hat{P}_{t-1}$) with the other agent's belief to obtain $\hat{P}_t$. This because in their model the world does not change, so the only way of describing an opinion dynamics is the one the authors illustrate.

The persuaded agent updates P_t^{ex} as specified in section 5.3. That is, when at time t she observes ω_k that used to belong to the latent space, $P_t^{ex}(\{\omega_k\}) = |P_{t-1}^{ex}(\{\omega_k\})|$, while the extended probabilities for $\omega_s \neq \omega_k$ are kept constant. Let us be more precise about the differences with equation (5.16); $\epsilon_{k,t}$ depends on both k and t. The persuaded agent has a different perception of the opinion of the persuading agent depending on whether she is not sure the topic they are debating about belongs to the state space Ω, so for ω_k in the latent space, $\epsilon_{k,t}$ is greater than 1. In addition, as time passes by, the hidden elements of the state space become known, so (part of) the credibility of the persuading agent is restored. The $\epsilon_{k,t}$ associated with ω_k observed at time t becomes smaller than 1, for all ω_k.

Notice that, for $t \geqslant N$, P_t^{ex} is a regular probability measure, because the persuaded agent discovers the composition of the whole state space, so the latent space shrinks to the empty set.

Throughout this section we made the tacit assumption that $\mathcal{P}$, the set of probability measures on Ω that induces the set of extended probabilities $\mathcal{P}_0^{ex}$ at time $t = 0$, is the singleton $\{P\}$, so that $\mathcal{P}_0^{ex} = \{P_0^{ex}\}$. This simplifying assumption can be dropped, and the analysis stays the same. The only difference is that we have to repeat it for all the elements of $\mathcal{P}_t^{ex}$, for all $t \in \mathbb{N}_0$. Every set $\mathcal{P}_t^{ex}$ of extended probabilities induces a set $\hat{\mathcal{P}}_t^{ex}$ of influenced extended probabilities. For all $t \geqslant N$, $\mathcal{P}_t^{ex}$ is a set of regular probability measures, and $\hat{\mathcal{P}}_t^{ex}$ is a set of influenced regular probability measures.

5.5 Upper and lower extended probabilities

Fix any $t \in \mathbb{N}_0$, and consider the "boundary elements" of the set $\mathcal{P}_t^{ex}$,

$$\underline{P}_t^{ex}(A) = \inf_{P^{ex} \in \mathcal{P}_t^{ex}} P^{ex}(A) \tag{5.19}$$

and

$$\overline{P}_t^{ex}(A) = \sup_{P^{ex} \in \mathcal{P}_t^{ex}} P^{ex}(\Omega) - \underline{P}_t^{ex}(A^c) = \sup_{P^{ex} \in \mathcal{P}_t^{ex}} P^{ex}(A), \tag{5.20}$$

for all $A \in \mathcal{F}$. They are not extended probabilities; we call $\underline{P}_t^{ex}(A)$ a lower extended probability measure, and $\overline{P}_t^{ex}(A)$ an upper extended probability measure. Notice that upper extended probability measures differ from upper regular probability measures. These latter are defined as 1 minus the lower regular probability of the complement of the event we are interested in. In (5.20) we give a similar conjugate type of definition, but we cannot write 1, because we do not require that the lower extended probability of Ω is 1.

Given a generic lower extended probabilities $\underline{P}^{ex}$ and a generic event $A \in \mathcal{F}$, we interpret $\underline{P}^{ex}(A)$ as follows. If $\underline{P}^{ex}(A) = p > 0$, then p represents the supremum price that we are willing to pay to enter a bet on A that gives us \$1 if A takes place and \$0 otherwise. If $\underline{P}^{ex}(A) = 0$, then in the most conservative of our mental states we do not enter a bet on A, since we deem it impossible to take place. Finally, if $\underline{P}^{ex}(A) = q < 0$, then $|q|$ represents the infimum betting rate at which our doppelgänger takes bets on A that pay \$1 if A takes place and \$0 otherwise. As we can see, this interpretation captures the ideas of worst case scenario and of prudent behavior. It can be seen as a betting scheme analogous to the one in (Walley, 1991, section 2.3.1), but with monetary instead of utiles outcomes, and extended to probabilities that can take on negative values as well.

Both lower and upper extended probabilities are extended Choquet capacities. A generic extended Choquet capacity is defined as a set function $\nu^{ex} : \mathcal{F} \to \mathbb{R}$ such

that

(EC1) $\nu^{ex}(\varnothing) = 0$,

(EC2) $\nu^{ex}(A) \in [-1, 1]$, for all $A \in \mathcal{F}$,

(EC3) for any $A, B \in \mathcal{F}$ such that $A \subset B$, if $\nu^{ex}(A) \geqslant 0$, $\nu^{ex}(B) \geqslant 0$, and $\nu^{ex}(B \cap A^c) \geqslant 0$, then $\nu^{ex}(A) \leqslant \nu^{ex}(B)$. If instead $\nu^{ex}(A) \leqslant 0$, $\nu^{ex}(B) \leqslant 0$, and $\nu^{ex}(B \cap A^c) \leqslant 0$, then $\nu^{ex}(A) \geqslant \nu^{ex}(B)$.

As we can see, we do not require countable additivity (ii*) to hold for extended capacities. An extended probability measure is an additive extended capacity.

It is immediate to see that upper and lower extended probability measures satisfy (EC1)-(EC3); in addition, lower extended probabilities are superadditive, while upper extended probabilities are subadditive. That is, for all $A, B \in \mathcal{F}$,

$$\underline{P}_t^{ex}(A \sqcup B) \geqslant \underline{P}_t^{ex}(A) + \underline{P}_t^{ex}(B) \tag{5.21}$$

and

$$\overline{P}_t^{ex}(A \sqcup B) \leqslant \overline{P}_t^{ex}(A) + \overline{P}_t^{ex}(B). \tag{5.22}$$

These inequalities come immediately from the properties of the infimum and supremum operators.

Remark 72. Notice that the behavioral interpreatation that we gave to lower extended probabilities entails that they can be specified even without eliciting a set of extended probability measures first. To this extent, our behavioral interpretation can be called minimal, similarly to (Walley, 1991, section 2.3.1). An immediate question the reader may ask is: "If we were to specify a lower extended probability without resorting to a set of extended probabilities, are we sure it is subadditive?". The answer is yes, under a mild assumption: Example 75 – based on the example in (Walley, 1991, section 1.6.4) – shows that coherence of lower extended probabilities

implies their superadditivity, and Theorem 74 shows that if $\underline{P}^{ex}$ can be obtained as the infimum of a set of extended probabilities, then it is coherent.

We first define coherence for lower extended probabilities. It is the immediate lower counterpart of Definition 59.

Definition 73. Lower extended probability $\underline{P}^{ex}$ is coherent if no Dutch books can be made against the punter, that is, if we cannot find a finite collection $\{B_j\}_{j=1}^n \subset \mathscr{B}'$ along with numbers $\{s_j\}_{j=1}^n \subset \mathbb{R}$ such that, for all $\omega \in \Omega$,

$$\sum_{j=1}^{n} s_j \left[\mathbb{1}_{B_j}(\omega) - \underline{P}^{ex}(B_j)\right] < 0, \tag{5.23}$$

where $\mathscr{B}'$ is the sigma-algebra generated by the collection $\{B \subset \Omega : \underline{P}^{ex}(B) \geqslant 0\}$.

We have the following interesting result, that is a version of the lower envelope theorem in (Walley, 1991, Corollary 2.8.6).

Theorem 74. Consider a generic lower extended probability $\underline{P}^{ex}$ defined on a measurable space $(\Omega, \mathcal{F})$. If there exists a nonempty set $\mathcal{P}^{ex}$ of extended probabilities on $(\Omega, \mathcal{F})$ such that $\underline{P}^{ex}(A) = \inf_{P^{ex} \in \mathcal{P}^{ex}} P^{ex}(A)$, for all $A \in \mathcal{F}$, then $\underline{P}^{ex}$ is coherent.

The following example shows that if $\underline{P}^{ex}$ is coherent, then it is superadditive.

Example 75. Pick two mutually exclusive events A and B. By our behavioral interpretation, the highest amount we are willing to pay to get \$1 if A occurs (or the highest amount we would deem reasonable to lose if we were given the possibility to enter the bet) is $\underline{P}^{ex}(A)$. The same holds for $\underline{P}^{ex}(B)$. By the coherence of our beliefs, the net outcome is equivalent to paying $\underline{P}^{ex}(A) + \underline{P}^{ex}(B)$ to get \$1 if $A \sqcup B$ occurs. Now, $\underline{P}^{ex}(A \sqcup B)$ is the highest price we are willing to pay to obtain \$1 if $A \sqcup B$ occurs. So we recover the superadditivity constraint

$$\underline{P}^{ex}(A \sqcup B) \geqslant \underline{P}^{ex}(A) + \underline{P}^{ex}(B).$$

Similarly, an upper lower probability $\overline{P}^{ex}$ should satisfy the subadditivity constraint $\overline{P}^{ex}(A \sqcup B) \leqslant \overline{P}^{ex}(A) + \overline{P}^{ex}(B)$. $\triangle$

An important concept worth introducing is the core of a lower extended probability measure.

Definition 76. Given a generic lower extended probability $\underline{P}^{ex}$, we call core of $\underline{P}^{ex}$ the set

$$\text{core}(\underline{P}^{ex}) := \{P^{ex} \in \Delta^{ex}(\Omega, \mathcal{F}) : P^{ex}(A) \geqslant \underline{P}^{ex}(A),\ \forall A \in \mathcal{F} \text{ and } P^{ex}(\Omega) = \underline{P}^{ex}(\Omega)\}.$$

This definition tells us that the core of $\underline{P}^{ex}$ is the set of all suitably normalized extended probabilities that setwise dominate $\underline{P}^{ex}$. Notice that in general it may be empty, and that

$$\begin{aligned}\text{core}(\underline{P}^{ex}) &= \left\{P^{ex} \in \Delta^{ex}(\Omega, \mathcal{F}) : \underline{P}^{ex} \leqslant P^{ex} \leqslant \overline{P}^{ex}\right\} \\ &= \left\{P^{ex} \in \Delta^{ex}(\Omega, \mathcal{F}) : P^{ex}(A) \leqslant \overline{P}^{ex}(A),\ \forall A \in \mathcal{F} \text{ and } P^{ex}(\Omega) = \underline{P}^{ex}(\Omega)\right\},\end{aligned}$$

so the core can be seen as the set of extended probabilities "sandwiched" between lower extended probability $\underline{P}^{ex}$ and upper extended probability $\overline{P}^{ex}$, as well as the set of extended probabilities setwise dominated by $\overline{P}^{ex}$. The following is a corollary to Theorem 74.

Corollary 77. If $\text{core}(\underline{P}^{ex}) \neq \varnothing$, then $\underline{P}^{ex}$ is coherent.

A crucial property of the core is the following.

Proposition 78. Given a generic lower extended probability $\underline{P}^{ex}$, its core is convex and weak*-compact.

Being compact and convex, the core of $\underline{P}^{ex}$ is completely characterized by $\underline{P}^{ex}$. This means that it is enough to know $\underline{P}^{ex}$ to be able to retrieve every element in its core. So in our analysis we can focus on updating $\underline{P}^{ex}_t$ to $\underline{P}^{ex}_{t+1}$, and then require

that $\mathcal{P}^{ex}_{t+1} = \text{core}(\underline{P}^{ex}_{t+1})$, instead of updating $\mathcal{P}^{ex}_t$ to $\mathcal{P}^{ex}_{t+1}$ elementwise. This justifies our focus in the remainder of this section on studying how to update lower extended probabilities. The procedure to update upper extended probability measures is going to be similar (their relation is described by equation (5.20)).

Recall that the conditions to perform the update in the additive case are given by (5.11) and (5.14).

Proposition 79. The sublinear counterpart of (5.11) is the following,

$$\underline{P}^{ex}_{t+1}(E^+_{t+1,\omega}) = \left|\overline{P}^{ex}_t(E^-_{t,\omega})\right| \quad \text{and} \quad \overline{P}^{ex}_{t+1}(E^+_{t+1,\omega}) = \left|\underline{P}^{ex}_t(E^-_{t,\omega})\right|. \tag{5.24}$$

It holds when we learn a new element ω of our sample space, that is, when the new observation ω belongs to the latent space at time t, but "moves" to the actual space at time $t+1$. In formulas, $\omega \in \Omega^-_t$, but $\omega \in \Omega^+_{t+1}$. The lower extended probabilities assigned to the other elements of $\mathcal{E}$ are held constant.

Notice that we are working with the finest possible partition of Ω, so, as before, the intersection $A \cap E$ between any $A \in \mathcal{F}$ and any $E = \{\omega\} \in \mathcal{E}$ is either the empty set, $A \cap E = \varnothing$, if $\omega \notin A$, or the element ω itself, $A \cap E = \{\omega\} = E$, if $\omega \in A$. Hence, for any $t \in \mathbb{N}_0$, for any $E = \{\omega\} \in \mathcal{E}$, and for any $A \in \mathcal{F}$, we have that

$$\underline{P}^{ex}_t(A \cap E) = \begin{cases} \underline{P}^{ex}_t(E) & \text{if } \omega \in A \\ \underline{P}^{ex}_t(\varnothing) = 0 & \text{if } \omega \notin A \end{cases}. \tag{5.25}$$

Given any $A, B \in \mathcal{F}$, if $\underline{P}^{ex}_t(B) \neq 0$, we define

$$\underline{P}^{ex}_t(A \mid B) := \frac{\inf_{P^{ex} \in \mathcal{P}^{ex}_t} P^{ex}(A \cap B)}{\inf_{P^{ex} \in \mathcal{P}^{ex}_t} P^{ex}(B)} = \frac{\underline{P}^{ex}_t(A \cap B)}{\underline{P}^{ex}_t(B)}, \tag{5.26}$$

for all $t \in \mathbb{N}_0$. We call it the extended Geometric rule. We notice immediately that, combining (5.25) and (5.26), we get

$$\underline{P}^{ex}_t(A \mid E) = \frac{\underline{P}^{ex}_t(A \cap E)}{\underline{P}^{ex}_t(E)} = \begin{cases} \frac{\underline{P}^{ex}_t(E)}{\underline{P}^{ex}_t(E)} = 1 & \text{if } \omega \in A \\ \frac{\underline{P}^{ex}_t(\varnothing)}{\underline{P}^{ex}_t(E)} = 0 & \text{if } \omega \notin A \end{cases}, \tag{5.27}$$

for all $E = \{\omega\} \in \mathcal{E}$ such that $\underline{P}_t^{ex}(E) \neq 0$ and all $A \in \mathcal{F}$. So, we have that $\underline{P}_t^{ex}(A \mid E) = P_t^{ex}(A \mid E)$, for all A, all E, and all $P_t^{ex} \in \mathcal{P}_t^{ex}$. This is true only because we are working with the finest possible partition of Ω, and because we use the extended Geometric rule.

The sublinear counterpart of (5.14) is the following

$$\underline{P}_{t+1}^{ex}(E_{t+1,\omega}^{+}) = \underline{P}_t^{ex}(E_{t,\omega}^{+}) \geqslant 0, \tag{5.28}$$

for all $t \in \mathbb{N}_0$. This comes from the fact that we draw an element already belonging to the actual space, so we do not need to update its lower extended probability (similarly to what is described in equation (5.14)). The lower extended probabilities assigned to the other elements of $\mathcal{E}$ are held constant.

At this point, a natural question one may ask is how to compute $\underline{P}_t^{ex}(A)$, for any t, for any $A \in \mathcal{F}$. It would be tempting to write that

$$\begin{aligned}\underline{P}_t^{ex}(A) = &\sum_{E_{t,j}^{+} \in \mathcal{E}_t^{+} : \underline{P}_t^{ex}(E_{t,j}^{+}) \neq 0} \underline{P}_t^{ex}(A \mid E_{t,j}^{+}) \underline{P}_t^{ex}(E_{t,j}^{+}) \\ &+ \sum_{E_{t,j}^{-} \in \mathcal{E}_t^{-} : \underline{P}_t^{ex}(E_{t,j}^{-}) \neq 0} \underline{P}_t^{ex}(A \mid E_{t,j}^{-}) \underline{P}_t^{ex}(E_{t,j}^{-}).\end{aligned}$$

This would mimic exactly (5.12), with lower extended probabilities in place of additive extended probabilities. Alas, that would not be true, since lower extended probabilities are not additive. Instead, we have the following.

Fix any $t \in \mathbb{N}_0$. Call $\mathfrak{E}_t \equiv \{E_{t,A}\}$ the collection of elements of $\mathcal{E}$ such that $A \cap E_{t,A} \neq \varnothing$. We index $\mathfrak{E}_t$ to time t to highlight the fact that although its elements stay the same, some of them may "move" from the latent to the actual space as time goes by and we collect more obsevations. Since we are working with the finest possible partition of Ω, we can write

$$A = \bigsqcup_{E_{t,A} \in \mathfrak{E}_t} E_{t,A}. \tag{5.29}$$

In the most general case, some of these $E_{t,A}$'s belong to the actual space, and some to the latent space. Let us denote the former by $E^+_{t,A}$'s and the latter by $E^-_{t,A}$'s. Formally, we have that $E^+_{t,A} \cap \Omega^+_t \neq \varnothing$ and $E^+_{t,A} \cap \Omega^-_t = \varnothing$, and vice versa for the $E^-_{t,A}$'s. Hence, (5.29) can be rewritten as

$$A = \bigsqcup_{E^+_{t,A} \in \mathfrak{E}_t} E^+_{t,A} \sqcup \bigsqcup_{E^-_{t,A} \in \mathfrak{E}_t} E^-_{t,A}. \tag{5.30}$$

Now, from (5.24) and (5.28), we know how to update the lower extended probabilities of all the elements of the patition $\mathcal{E}$ of Ω, which implies that we know the value of $\underline{P}^{ex}_{t+1}(E_{t+1,A})$, for all $E_{t+1,A} \in \mathfrak{E}_{t+1}$. Then, consider now the summation

$$\sum_{E_{t+1,A} \in \mathfrak{E}_{t+1}} \underline{P}^{ex}_{t+1}(E_{t+1,A}).$$

From equation (5.21), we have that

$$\begin{aligned}
\underline{P}^{ex}_{t+1}(A) &\geqslant \sum_{E_{t+1,A} \in \mathfrak{E}_{t+1}} \underline{P}^{ex}_{t+1}(E_{t+1,A}) \\
&= \sum_{E^+_{t+1,A} \in \mathfrak{E}_{t+1}} \underline{P}^{ex}_{t+1}(E^+_{t+1,A}) + \sum_{E^-_{t+1,A} \in \mathfrak{E}_{t+1}} \underline{P}^{ex}_{t+1}(E^-_{t+1,A}).
\end{aligned}$$

From equation (5.22), we can give an upper bound for the upper extended probability of A,

$$\begin{aligned}
\overline{P}^{ex}_{t+1}(A) &\leqslant \sum_{E_{t+1,A} \in \mathfrak{E}_{t+1}} \overline{P}^{ex}_{t+1}(E_{t+1,A}) \\
&= \sum_{E^+_{t+1,A} \in \mathfrak{E}_{t+1}} \overline{P}^{ex}_{t+1}(E^+_{t+1,A}) + \sum_{E^-_{t+1,A} \in \mathfrak{E}_{t+1}} \overline{P}^{ex}_{t+1}(E^-_{t+1,A}).
\end{aligned}$$

6

Conclusion

In this work we give novel interesting results in Choquet theory and in the theory of capacities.

In chapter 2 there are two key ideas. The first one is that we can use techniques from stochastic convex geometry on the growth rate of the expected number of extrema of random polytopes to provide insight into the asymptotic growth rate of the expected number of mixture components in a finite admixture model. We prove that the expected number of identifiable mixture components increases at the rate of $(\log n)^{J-1}$ where J is the dimension of the Euclidean space we work with, and n is the amount of data points we collect. We also provide a central limit theorem for the distributions of the number of extrema. The other key concept is that we can retrieve admixture weights using techniques from Choquet theory. In particular, we show that if the convex hull $\mathcal{K}_M$ generated by the identifiable elements of the finite mixture model is a simplex, a Pólya tree posterior always recovers the Choquet measure for π_i, for any $\pi_i \in \mathcal{K}_M$. It does so with the proper minimax rate. We also give an algorithm to find the richest cheap admixture model. An interesting open question is whether there are other instances in Bayesian inference where coupling

results from stochastic geometry on extremal sets with results in Choquet theory allows to develop novel analyses, insights, models, or algorithms.

In chapter 3 we present dynamic probability kinematics (DPK) and dynamic imprecise probability kinematics (DIPK). These methods dynamically update subjective beliefs stated in terms of precise and imprecise probabilities, in the presence of partial information. In the case of DIPK, we provide bounds for the upper and lower probabilities associated with the updated sets, and study their set-specific behavior including contraction, dilation, and sure loss. Two examples are provided to illustrate the procedures. The results in this chapter are just the first step towards a fully developed DIPK theory. In the future, we plan to relax the assumption that Ω needs to be at most countable. We also plan to find sufficient conditions for the inequalities in Section 3.7 to hold with equality. For example, in Wasserman and Kadane (1990), the authors study a Bayes theorem for lower probabilities. They first find a lower bound for the lower posterior $\underline{P}_y$ coming from a generalization of Bayes rule combining lower prior $\underline{P}$ with likelihood $f(y \mid \theta)$. They then show that if lower prior $\underline{P}$ is convex, that is, if $\underline{P}(A \cup B) + \underline{P}(A \cap B) \geqslant \underline{P}(A) + \underline{P}(B)$, then the lower bound for lower posterior $\underline{P}_y$ holds with equality. We conjecture that convexity, possibly together with additional requirements, will allow us to reach our goal. Furthermore, we aim to generalize DIPK by allowing the agent to gather inconsistent evidence as in Marchetti and Antonucci (2018) and by letting partial information be modeled via a set of probability distributions on $\mathcal{X}$, as empirical probabilities usually need very large sample sizes to estimate probabilities which are very close to zero or one to a good standard of relative accuracy. After that, we intend to propose a way of performing statistical analysis based on DIPK updating. Our last goal is to generalize DIPK to work with lower previsions in place of lower probabilities.

In chapter 4 we give an ergodic theory for the limit $\mathcal{P}^{\text{co}}_{\tilde{\mathcal{E}}}$ of the sequence $(\mathcal{P}^{\text{co}}_{\mathcal{E}_n})$ of successive dynamic imprecise probability kinematics update of a set $\mathcal{P}^{\text{co}}_{\mathcal{E}_0}$ of probabil-

ities representing the initial beliefs of an agent on the state space Ω. A consequence of this ergodic theory is a strong law of large numbers. In the future, we plan to find sufficient conditions that are easier to verify with respect to the ones we have given for $\underline{P}_{\tilde{\mathcal{E}}}$ to be T-invariant, ergodic, convex, strictly invariant, and functionally invariant. We also aim to show that if for the generalizations to DPK and DIPK discussed earlier we are able to prove that sequence $(\mathcal{P}^{\text{co}}_{\mathcal{E}_n})$ converges, then a version of the results presented in this chapter will continue to hold (for example, if Ω is uncountable, we will need to consider Choquet integrals as in (Cerreia-Vioglio et al., 2015)).

In chapter 5 we give a definition of extended probability measures that does not depend on the environment a scholar works in. We give some of their more interesting properties, and a behavioral interpretation to positive and negative values of extended probabilities. We then apply extended probabilities to statistical inference. Given the probability space $(\Omega, \mathcal{F}, P)$ associated with the experiment we want to conduct, we use extended probabilities to express uncertainty about both the composition of the state space Ω, and which probability measure P to select. We develop an ex-ante analysis; our method describes how the researcher progressively discovers the true composition of the state space, so that, at the end of the process, a regular statistical analysis (that requires the knowledge of Ω) can take place. We introduce the concept of extended Choquet capacities, and in particular of upper and lower extended probabilities that represent the "borders" of sets of extended probabilities, and we give bounds for the lower extended probability of any element $A \in \mathcal{F}$. We also apply our model to the fields of opinion dynamics and species sampling models. This chapter is important because it gives a foundational definition of extended probability measures and makes these latter relevant to the field of statistics. It also provides a very interesting way of using these tools in social sciences, in particular in the study of opinion dynamics. In the future, we will deal with the possibility

that in our ex-ante analysis we are not able to discover the whole composition of the state space associated with our experiment. In that case, the statistical analysis itself will have to be carried out using extended probabilities. We also plan to relax the assumption we made about working with a finite or countable state space. Furthermore, we would like to deepen the study of lower extended probabilities.

6.1 More open problems

In this section, we provide some of the open projects – related to geometric and statistical Choquet theories – we are currently working on.

We plan to address an inference problem on a convex model using a Bayesian nonparametric approach when the prior on the distribution cannot be exactly specified; in particular, we will consider the case in which the researcher can only address a set of such priors. This corresponds to modeling the ambiguity the scientist faces when expressing their uncertainty around events not via a set of probabilities, but rather via a set of random measures. In addition, we will provide a result on convergence stating that, under some finitness assumptions, our way of proceeding remarkably simplifies.

We recently started working on framing the dirty Bayes results given in (Meng, 2021) in an imprecise probabilities framework, a project that seems very promising since it will allow to utilize the Bayes paradigm when the prior is not unique and the likelihood is incomplete, in the spirit of what we did in chapter 3 for Jeffrey's updating.

Finally, in the long term we would like to find new exciting results regarding Fubini's theorem for lower probabilities, and to use the idea of probability filters (collections of sets of probabilities) to model opinion dynamics. In particular, if agents in a group face ambiguity, they can specify sets of probabilities to represent their beliefs, and the cumulative beliefs in the group can be represented by probabil-

ity filters. This may be used to combine opinions of experts to build richer, sounder statistical models. We also started a project that tries to reconcile subjective probability with conformal inference.

Appendix A

Appendix to Chapter 2

A.1 Distribution of our sequence of random points

In this Appendix, we assume that the number L of admixture components is known, but the parameters $\{f_1, \ldots, f_L\}$ of the multinomial components are not. We assume they are identically distributed random variables, but we do not require independence. After realizing that collection $\{f_1, \ldots, f_L\}$ can be seen as a finite exchangeble sequence, we inspect how to approximate its joint distribution applying de Finetti's theorem and a result by Diaconis and Freedman.

As pointed out in (Aldous, 2010), we can state de Finetti's result from a functional analytic viewpoint as follows. Let $\mathcal{S} \equiv \Delta^{J-1} \subset \mathbb{R}^J$, and recall that a sequence of random variables X_i's is exchangeable if

$$(X_i)_{i \geqslant 1} \stackrel{\mathrm{d}}{=} (X_{perm(i)})_{i \geqslant 1},$$

for any finite permutation $perm$, where $\stackrel{\mathrm{d}}{=}$ denotes equality in distribution. We can assume that the elements $f_1, \ldots, f_L$ form a finite exchangeable sequence because the order in which they appear provides no additional information about the finite

mixture model.

Let $\Delta(\mathcal{S}) \equiv \Delta(\mathcal{S}, \mathcal{B}(\mathcal{S}))$ be the set of probability measures on $(\mathcal{S}, \mathcal{B}(\mathcal{S}))$, where $\mathcal{B}(\mathcal{S})$ is the Borel sigma-algebra for $\mathcal{S}$. Let then $\Delta(\Delta(\mathcal{S}))$ be the set of probability measures on $\Delta(\mathcal{S})$. When we define an infinite exchangeable sequence of $\mathcal{S}$-valued random variables, we are actually defining an exchangeable measure, say Θ, on $\Delta(\mathcal{S}^\infty)$, where Θ is the distribution of the sequence.

Consider the set $\mathfrak{M} := \{\mu^\infty := \mu \times \mu \times \cdots \text{ s.t. } \mu \in \Delta(\mathcal{S})\} \subset \Delta(\mathcal{S}^\infty)$, that is the set of extrema of the convex set of exchangeable elements of $\Delta(\mathcal{S}^\infty)$. Then, we have

$$\Theta(A) = \int_{\Delta(\mathcal{S})} \mu^\infty(A)\ \Lambda(\mathrm{d}\mu), \qquad \forall A \subset \mathcal{S}^\infty.$$

Hence, there is a bijection between $\Lambda \in \Delta(\Delta(\mathcal{S}))$ and $\Theta \in \Delta(\mathcal{S})$. Notice that a consequence of Proposition 16 is that for all $x \in \Delta^{J-1}$, there exists a unique Choquet measure $\tilde{\nu}_x$, whose support are the extrema of Δ^{J-1}, denoted as $ex(\Delta^{J-1})$, that allows to represent x. Then, there always exists a probability measure $\breve{\nu}_x$ supported on the whole simplex Δ^{J-1}, whose restriction to $ex(\Delta^{J-1})$ is given by $\tilde{\nu}_x$. To this extent, $\breve{\nu}_x \times \breve{\nu}_x \times \cdots =: \breve{\nu}_x^\infty$ belongs to $\mathfrak{M}$.

As we pointed out before, we can assume $(f_1, \ldots,\ f_L)$ to be a finite exchangeable sequence. Suppose, without loss of generality, that it is part of a much longer sequence of m components

$$(f_1, \ldots, f_L, \ldots, f_m)\,.$$

Then, we can use (Diaconis and Freedman, 1980, Theorem 13) to compute an approximation of Θ_L, the distribution of our finite sequence. Let us denote by Θ_m the distribution of $(f_1, \ldots, f_L, \ldots, f_m)$; it is an exchangeable probability on $\mathcal{S}^m$. Then, Θ_L, $L \leqslant m$, is the projection of Θ_m onto $\mathcal{S}^L$. Define the value $\beta(m, L)$ as

$$\beta(m, L) := 1 - \frac{m^{-L} m!}{(m - L)!},$$

and notice that $\beta(m, L) \leqslant \frac{1}{2}\frac{L(L-1)}{m}$.

The theorem states that there exists $\tilde{\Lambda} \in \Delta(\Delta(\mathcal{S}))$ such that the probability $\Theta_{\mu L}$ defined on $\mathcal{S}^L$ as

$$\Theta_{\mu L}(A) = \int_{\Delta(\mathcal{S})} \mu^L(A)\ \tilde{\Lambda}(\mathrm{d}\mu), \quad \forall A \subset \mathcal{S}^L$$

is such that $d_{TV}(\Theta_L, \Theta_{\mu L}) \leqslant 2\beta(m, L)$, for all $L \leqslant m$. We denoted by μ^L the distribution of L independent picks from μ, that is, $\mu^L((s_1, \ldots, s_L)) = \prod_{j=1}^{L} \mu(s_j)$, and by d_{TV} the total variation distance

$$d_{TV}(\Theta_L, \Theta_{\mu L}) := \sup_{A \subset S^L} \left|\Theta_L(A) - \Theta_{\mu L}(A)\right|.$$

Notice that $\tilde{\Lambda}$ depends on m and Θ_m, but not on L, and its analytical form is given in (Diaconis and Freedman, 1980, Proof of Theorem 13).

A.2 Number of extrema of the convex hull having the least amount of vertices

The following is an interesting result dealing with the number of extrema of the convex hull in Δ^{J-1} having the least amount of vertices.

Proposition 80. Let $\tilde{e}$ be the number of extrema of the convex hull (polytope) in our unit simplex with the least amount of vertices. Then, $\tilde{e} = J$.

Appendix B

Proofs

Proof of Theorem 8. In (Reitzner, 2005b, Theorem 6) and (Bárány and Buchta, 1993, Theorem 5), the authors show that, given a convex polytope P in $\mathbb{R}^d$, if we call P_n the convex hull of n points sampled iid from a uniform on P, then

$$\mathbb{E}\left[F_0(P_n)\right] = \frac{1}{(d+1)^{d-1}(d-1)!}T(P)(\log n)^{d-1} + O\left((\log n)^{(d-2)}\log\log n\right).$$

Then, since Δ^{J-1} is a convex polytope in $\mathbb{R}^J$, and given the way we defined K_n, equation (2.2) follows immediately. □

Proof of Theorem 9. Let us denote by $\mathbb{K}_+^2$ the set of compact convex sets in $\mathbb{R}^d$, $d \geqslant 2$, having nonempty interior, boundary of differentiability class $\mathcal{C}^2$, and positive Gaussian curvature. Pick any $K \in \mathbb{K}_+^2$, and sample n points iid from the uniform on K. Call their convex hull P_n. Then, in (Reitzner, 2005a, Theorem 6), the author shows that there are numbers d_n bounded between two positive constants depending on K, and a constant $c(K)$, such that

$$\left|\mathbb{P}\left(\frac{F_i(P_n) - \mathbb{E}[F_i(P_n)]}{\sqrt{d_n n^{1-\frac{2}{d+1}}}} \leqslant t\right) - \Phi(t)\right| \leqslant c(K)n^{-\frac{1}{2(d+1)}}(\log n)^{2+3i+\frac{2}{d+1}}. \tag{B.1}$$

The denominator $\sqrt{d_n n^{1-\frac{2}{d+1}}}$ is of the same asymptotic order as the standard deviation of $F_i(P_n)$, so the inequality in (B.1) implies a central limit theorem for $F_i(P_n)$.

Notice then that $\Delta^{J-1} \in \mathbb{K}^2_+$ for any $\mathbb{R}^J$, $J \geqslant 2$. Hence, given the way we defined K_n, equation (2.3) follows immediately. As we can see, the rate of convergence of the distribution of $F_0(K_n)$ to Φ is given by $n^{-\frac{1}{2(J+1)}}(\log n)^{2+\frac{2}{J+1}}$. □

Proof of Theorem 10. Call E_n the extremal set of K_n, that is, $E_n = \text{ex}(K_n)$. Let $F_0(K_n) \to \infty$, and call $\tilde{E} \neq \varnothing$ the set that E_n tends to in the Hausdorff distance

$$d_H(E_n, \tilde{E}) := \max\left\{ \sup_{f\in E_n} \inf_{g\in\tilde{E}} d(f,g) \ , \ \sup_{g\in\tilde{E}} \inf_{f\in E_n} d(f,g) \right\}$$

$$= \sup_{s\in E_n\cup\tilde{E}} \left| \inf_{f\in E_n} d(s,f) - \inf_{g\in\tilde{E}} d(s,g) \right|,$$

as the cardinality of E_n approaches infinity. Here d denotes the usual Euclidean distance, and the second equality is an equivalent way of writing the Hausdorff distance. Let $\tilde{K}$ be the convex hull of $\tilde{E}$.

Step 1: We first show that $\tilde{K}$ is well defined. By construction, we know that $\tilde{E} \neq \varnothing$; $\tilde{K}$ is then the convex hull of the points in $\tilde{E}$, which is well defined as we can always construct the convex hull of any given (sub)set of a vector space.

Step 2: Now, we show that $\tilde{K}$ has infinitely many sides. Suppose for the sake of contradiction that $\tilde{K}$ has finitely many sides. Then, it has a finite number of ℓ-faces, for some ℓ, which implies a finite number of vertices. But $\tilde{K}$ is the convex hull of the elements in $\tilde{E}$, that are infinite, a contradiction.

Step 3: $\tilde{K}$ is convex: this is immediate from it being the convex hull of $\tilde{E}$.

Step 4: We are left to show that $\tilde{K}$ is the limit of K_n. We have seen that $E_n \to \tilde{E}$ in the Hausdorff metric as n goes to infinity; we also know that $K_n = \text{Conv}(E_n)$, for all n (there is a small abuse of notation here: K_n is the convex hull of the elements of E_n; since no confusion arises and since we save some notation, we leave it as it is).

But then

$$K_n = \text{Conv}\,(E_n) \xrightarrow[n\to\infty]{d_H} \text{Conv}(\tilde{E}) = \tilde{K},$$

which concludes our proof. □

Proof of Theorem 11. The proof consists of showing that if condition $\mathbb{E}[T(n)] = \gamma \cdot \mathbb{E}[M(n)]$ holds, then the growth rate of the extrema stated in Theorem 8 will hold for a more general procedure. We already know from Theorem 8 that if $S_1, \ldots, S_n \sim \text{Uniform}(\Delta^{J-1})$ iid, then $\lim_{n\to\infty}(\log n)^{-(J-1)}\mathbb{E}[F_0(K_n)] = c(J)$, a value depending on the dimension of the Euclidean space $\mathbb{R}^J$ we work in. Recall that the number of extrema $F_0(K_n)$ of the convex body associated with our mixture model in the uniform case corresponds to $M(n)$.

We now relax the assumption in equation (2.1).

Fix any $n \in \mathbb{N}$. Let then $\mathbb{E}[T(n)]$ be the expected number of extrema of $\breve{K}_n$. Assume that $\mathbb{E}[T(n)] = \gamma \cdot \mathbb{E}[M(n)]$, for some $\gamma \in \mathbb{Q}$. Then, by Theorem 8, we have that $\lim_{n\to\infty}(\log n)^{-(J-1)}\mathbb{E}[T(n)] = \gamma \cdot c(J)$. Equation (2.4) then follows by putting $c'(J, \gamma) = \gamma \cdot c(J)$. □

Proof of Theorem 13. First notice that sequence (γ_n) exists because we can always write a natural number as a linear function of another natural number, using a rational coefficient. By hypothesis, then, we have that $\mathbb{E}[T(n)] = \gamma_n\mathbb{E}[M(n)]$ and $\lim_{n\to\infty} \gamma_n/r_{\gamma_n} = 1$. In addition, by Theorem 8 we have that

$$\lim_{n\to\infty} \frac{1}{(\log n)^{J-1}}\mathbb{E}[M(n)] = c(J).$$

Hence we obtain that

$$\lim_{n\to\infty} \frac{\mathbb{E}[T(n)]}{r_{\gamma_n}(\log n)^{J-1}} = \lim_{n\to\infty} \frac{\gamma_n\mathbb{E}[M(n)]}{r_{\gamma_n}(\log n)^{J-1}} = \lim_{n\to\infty} \frac{\gamma_n}{r_{\gamma_n}} \lim_{n\to\infty} \frac{\mathbb{E}[M(n)]}{(\log n)^{J-1}} = c(J),$$

concluding the proof. □

Proof of Proposition 16. The proposition is an immediate consequence of Definition 2 and Theorem 3. □

Proof of Theorem 19. The need for $\mathcal{K}_M$ to be a simplex comes from Proposition 16. The weak convergence statement comes from (Ghosal and van der Vaart, 2017, Corollary 4.17), while the rate of convergence comes from (Ghosal and van der Vaart, 2017, Example 8.5). In this latter, the authors give it relative to the semimetric $d(P, Q) = |P(A) - Q(A)|$, for a fixed A. But because this holds for all measurable sets A, then the rate of convergence holds in terms of the total variation distance as well. □

Proof of Proposition 21. We begin by showing that $P_{\mathcal{E}}$ is a probability measure. We verify the Kolmogorovian axioms for a probability measure. First, we have that $P_{\mathcal{E}}(A) \geqslant 0$, for all $A \in \mathcal{F}$. This comes by its definition, since it is defined as the summation of products of nonnegative quantities. Second, we have that $P_{\mathcal{E}}(\Omega) = 1$. This comes from the following

$$P_{\mathcal{E}}(\Omega) = \sum_{E_j \in \mathcal{E}} P(\Omega \mid E_j) P_{\mathcal{E}}(E_j) = \sum_{E_j \in \mathcal{E}} P_{\mathcal{E}}(E_j) = 1.$$

Finally, we have that if $\{A_i\}_{i \in I}$ is a countable, pairwise disjoint collection of events,

then $P_{\mathcal{E}}(\cup_{i\in I} A_i) = \sum_{i\in I} P_{\mathcal{E}}(A_i)$. This because

$$\begin{aligned}
P_{\mathcal{E}}\left(\cup_{i\in I} A_i\right) &= \sum_{E_j\in\mathcal{E}} P\left(\cup_{i\in I} A_i \mid E_j\right) P_{\mathcal{E}}(E_j) \\
&= \sum_{E_j\in\mathcal{E}} \frac{P\left(\left[\cup_{i\in I} A_i\right] \cap E_j\right)}{P(E_j)} P_{\mathcal{E}}(E_j) \\
&= \sum_{E_j\in\mathcal{E}} \frac{P\left(\cup_{i\in I}\left[A_i \cap E_j\right]\right)}{P(E_j)} P_{\mathcal{E}}(E_j) \\
&= \sum_{E_j\in\mathcal{E}} \frac{\sum_{i\in I} P\left(A_i \cap E_j\right)}{P(E_j)} P_{\mathcal{E}}(E_j) \\
&= \sum_{i\in I}\sum_{E_j\in\mathcal{E}} \frac{P\left(A_i \cap E_j\right)}{P(E_j)} P_{\mathcal{E}}(E_j) = \sum_{i\in I} P_{\mathcal{E}}\left(A_i\right).
\end{aligned}$$

We now show that $P_{\mathcal{E}}$ is a Jeffrey's posterior for P. We use (Diaconis and Zabell, 1982, Theorem 2.1): it states that $P^\star$ is a Jeffrey's posterior for P if and only if there exists a constant $B \geqslant 1$ such that $P^\star(\{\omega\}) \leqslant BP(\{\omega\})$, for all $\omega \in \Omega$. Fix any $\omega \in \Omega$. We have that $P_{\mathcal{E}}(\{\omega\}) = \sum_{E_j\in\mathcal{E}} P(\{\omega\} \mid E_j) P_{\mathcal{E}}(E_j)$. Call E_ω the element in $\mathcal{E}$ such that $\{\omega\} \cap E_\omega \neq \varnothing$. Then, we have that

$$P_{\mathcal{E}}(\{\omega\}) = \sum_{E_j\in\mathcal{E}} P(\{\omega\} \mid E_j) P_{\mathcal{E}}(E_j) = \frac{P_{\mathcal{E}}(E_\omega)}{P(E_\omega)} P(\{\omega\}). \tag{B.2}$$

Now, let $B_\omega := \lceil \frac{P_{\mathcal{E}}(E_\omega)}{P(E_\omega)} + 1 \rceil$. We have that $P_{\mathcal{E}}(\{\omega\}) < B_\omega P(\{\omega\})$. Consider then the well-ordered collection $\{B_\omega\}_{\omega\in\Omega}$. If we let $B := \sup_{B'_\omega \in \{B_\omega\}} B'_\omega$, we conclude that $P_{\mathcal{E}}(\{\omega\}) < BP(\{\omega\})$, for all $\omega \in \Omega$. □

Proof of Proposition 24. We have two cases. If $\cup_{i\in\mathbb{N}} x_i = \mathcal{X}$, then, since we observed all the elements of $\mathcal{X}$, and given the procedure in Sections 3.3 and 3.4 to refine the partition, it is immediate to see that the partition $\tilde{\mathcal{E}}$ induced by $\{x_i\}_{i\in\mathbb{N}}$ cannot be further refined. If instead $\cup_{i\in\mathbb{N}} x_i = \mathcal{X}_{reduced} \subsetneq \mathcal{X}$, then the elements of partition $\tilde{\mathcal{E}}$

will be the unique elements of the collection $\{X^{-1}(x_i)\}_{x_i \in \mathcal{X}_{reduced}}$, plus an extra one given by $(\cup_{x_i \in \mathcal{X}_{reduced}} X^{-1}(x_i))^c$. □

Proof of Theorem 25. Pick any $t \in \mathbb{N}_0$ and consider an element $P_{\tilde{\mathcal{E}}}$ of $\mathscr{Q}$. By the law of total probability we can write, for all $A \in \mathcal{F}$,

$$P_{\tilde{\mathcal{E}}}(A) = \sum_{E \in \tilde{\mathcal{E}}} P_{\tilde{\mathcal{E}}}(A \mid E) P_{\tilde{\mathcal{E}}}(E). \tag{B.3}$$

Notice that, since $P_{\tilde{\mathcal{E}}}$ is an extension of $Q_{\tilde{\mathcal{E}}}$ to $\mathcal{F}$, we have that $P_{\tilde{\mathcal{E}}}(E) = Q_{\tilde{\mathcal{E}}}(E) = Q(E)$, for all $E \in \tilde{\mathcal{E}}$. So we can rewrite (B.3) as

$$P_{\tilde{\mathcal{E}}}(A) = \sum_{E \in \tilde{\mathcal{E}}} P_{\tilde{\mathcal{E}}}(A \mid E) Q(E). \tag{B.4}$$

Notice also that, by Proposition 24, any $E' \in \mathcal{E}_t$ can be written as the union of elements of $\tilde{\mathcal{E}}$, that is, for all $E' \in \mathcal{E}_t$, there exists a collection $\tilde{\mathcal{E}}_{E'} \subset \tilde{\mathcal{E}}$ such that $E' = \cup_{E \in \tilde{\mathcal{E}}_{E'}} E$. We then have that, for all $E' \in \mathcal{E}_t$, $P_{\tilde{\mathcal{E}}}(E') = P_{\tilde{\mathcal{E}}}(\cup_{E \in \tilde{\mathcal{E}}_{E'}} E) = \sum_{E \in \tilde{\mathcal{E}}_{E'}} P_{\tilde{\mathcal{E}}}(E) = \sum_{E \in \tilde{\mathcal{E}}_{E'}} Q(E) = Q(E')$. So we can rewrite (B.4) as

$$P_{\tilde{\mathcal{E}}}(A) = \sum_{E' \in \mathcal{E}_t} P_{\tilde{\mathcal{E}}}(A \mid E') Q(E'). \tag{B.5}$$

Suppose without loss of generality that $\mathcal{E}_t$ has $\ell + 1$ elements, and $E_{\ell+1}^{\mathcal{E}_t} \neq \varnothing$. We

have the following

$$
\begin{aligned}
d_{TV}(P_{\mathcal{E}_t}, P_{\tilde{\mathcal{E}}}) &:= \sup_{A\in\mathcal{F}} |P_{\mathcal{E}_t}(A) - P_{\tilde{\mathcal{E}}}(A)| \\
&\equiv |P_{\mathcal{E}_t}(\mathbf{A}) - P_{\tilde{\mathcal{E}}}(\mathbf{A})| \\
&= \left| \sum_{E'\in\mathcal{E}_t} P_{\mathcal{E}_{t-1}}(\mathbf{A} \mid E') P_{\mathcal{E}_t}(E') - P_{\tilde{\mathcal{E}}}(\mathbf{A} \mid E') P_{\tilde{\mathcal{E}}}(E') \right| \\
&= \left| \sum_{E'\in\mathcal{E}_t} P_{\mathcal{E}_{t-1}}(\mathbf{A} \mid E') \left[\beta(n_t) P_{\mathcal{E}_{t-1}}(E') + (1-\beta(n_t)) P_t^{emp}(E') \right] \right. \\
&\quad \left. - P_{\tilde{\mathcal{E}}}(\mathbf{A} \mid E') Q(E') \right| \qquad \text{(B.6)} \\
&= \left| \sum_{E'\in\mathcal{E}_t} P_{\mathcal{E}_{t-1}}(\mathbf{A} \mid E') \left[\beta(n_t) P_{\mathcal{E}_{t-1}}(E') + \frac{1-\beta(n_t)}{n_t+1} \sum_{i=1}^{n_t} \mathbb{I}(E' = X^{-1}(x_i)) \right] \right. \\
&\quad \left. - P_{\tilde{\mathcal{E}}}(\mathbf{A} \mid E') Q(E') \right| \qquad \text{(B.7)} \\
&\leqslant \sum_{E'\in\mathcal{E}_t} \left| \beta(n_t) P_{\mathcal{E}_{t-1}}(\mathbf{A} \cap E') + P_{\mathcal{E}_{t-1}}(\mathbf{A} \mid E') \frac{1-\beta(n_t)}{n_t+1} \sum_{i=1}^{n_t} \mathbb{I}(E' = X^{-1}(x_i)) \right. \\
&\quad \left. - P_{\tilde{\mathcal{E}}}(\mathbf{A} \mid E') Q(E') \right| \qquad \text{(B.8)} \\
&\leqslant \sum_{E'\in\mathcal{E}_t} \beta(n_t) P_{\mathcal{E}_{t-1}}(\mathbf{A} \cap E') + \left| P_{\mathcal{E}_{t-1}}(\mathbf{A} \mid E') \frac{1-\beta(n_t)}{n_t+1} \sum_{i=1}^{n_t} \mathbb{I}(E' = X^{-1}(x_i)) \right. \\
&\quad \left. - P_{\tilde{\mathcal{E}}}(\mathbf{A} \mid E') Q(E') \right| \qquad \text{(B.9)} \\
&\leqslant \sum_{E'\in\mathcal{E}_t} \beta(n_t) P_{\mathcal{E}_{t-1}}(\mathbf{A} \cap E') + \zeta \left| \frac{1-\beta(n_t)}{n_t+1} \sum_{i=1}^{n_t} \mathbb{I}(E' = X^{-1}(x_i)) - Q(E') \right| \qquad \text{(B.10)} \\
&= \beta(n_t) P_{\mathcal{E}_{t-1}}(\mathbf{A}) + \zeta \sum_{E'\in\mathcal{E}_t} \left| \frac{1-\beta(n_t)}{n_t+1} \sum_{i=1}^{n_t} \mathbb{I}(E' = X^{-1}(x_i)) - Q(E') \right|, \qquad \text{(B.11)}
\end{aligned}
$$

where ζ is an element of $\mathbb{R}_+$. Notice that (B.6) comes from (3.7) and (B.5), (B.7)

comes from (3.3), (B.8) comes from the triangular inequality and the fact that

$$P_{\mathcal{E}_{t-1}}(\mathbf{A} \mid E') = \frac{P_{\mathcal{E}_{t-1}}(\mathbf{A} \cap E')}{P_{\mathcal{E}_{t-1}}(E')},$$

and (B.9) comes again from the triangular inequality and the fact that both $\beta(n_t)$ and $P_{\mathcal{E}_{t-1}}(\mathbf{A} \cap E')$ are positive. We also have that (B.10) holds because, since $P_{\mathcal{E}_{t-1}}(\mathbf{A} \mid E')$ and $P_{\tilde{\mathcal{E}}}(\mathbf{A} \mid E')$ belong to $(0, 1)$, we can always find $\zeta \in \mathbb{R}_+$ such that

$$\begin{aligned}
&\zeta \left| \frac{1 - \beta(n_t)}{n_t + 1} \sum_{i=1}^{n_t} \mathbb{I}(E' = X^{-1}(x_i)) - Q(E') \right| \\
&\qquad \geqslant \left| P_{\mathcal{E}_{t-1}}(\mathbf{A} \mid E') \frac{1 - \beta(n_t)}{n_t + 1} \sum_{i=1}^{n_t} \mathbb{I}(E' = X^{-1}(x_i)) - P_{\tilde{\mathcal{E}}}(\mathbf{A} \mid E') Q(E') \right|,
\end{aligned}$$

for all $E' \in \mathcal{E}_t$. Finally, (B.11) holds because $\sum_{E' \in \mathcal{E}_t} P_{\mathcal{E}_{t-1}}(\mathbf{A} \cap E') = P_{\mathcal{E}_{t-1}}(\mathbf{A})$.

Now we are ready to consider the limit as $n_t \to \infty$

$$\begin{aligned}
&\lim_{n_t \to \infty} d_{TV}(P_{\mathcal{E}_t}, P_{\tilde{\mathcal{E}}}) \\
&\quad \leqslant \lim_{n_t \to \infty} \left\{ \beta(n_t) P_{\mathcal{E}_{t-1}}(\mathbf{A}) + \zeta \sum_{E' \in \mathcal{E}_t} \left| \frac{1 - \beta(n_t)}{n_t + 1} \sum_{i=1}^{n_t} \mathbb{I}(E' = X^{-1}(x_i)) - Q(E') \right| \right\} \\
&\quad = P_{\mathcal{E}_{t-1}}(\mathbf{A}) \lim_{n_t \to \infty} \beta(n_t) + \zeta \lim_{n_t \to \infty} \sum_{E' \in \mathcal{E}_t} \left| \frac{1 - \beta(n_t)}{n_t + 1} \sum_{i=1}^{n_t} \mathbb{I}(E' = X^{-1}(x_i)) - Q(E') \right| \qquad \text{(B.12)} \\
&\quad = 0 + \zeta \sum_{E' \in \tilde{\mathcal{E}}} \lim_{n_t \to \infty} \left| \frac{1 - \beta(n_t)}{n_t + 1} \sum_{i=1}^{n_t} \mathbb{I}(E' = X^{-1}(x_i)) - Q(E') \right| = 0 + 0 = 0 \qquad \text{(B.13)}
\end{aligned}$$

Q-almost surely. Notice that (B.12) comes from the fact that $P_{\mathcal{E}_{t-1}}$ does not depend on n_t. Then, in (B.13), $\beta(n_t)$ goes to zero from our assumption that $\beta(n_t) = o(1/n_t)$. In addition, as n_t grows to infinity, $\mathcal{E}_t$ becomes more and more similar to $\tilde{\mathcal{E}}$, so in the limit we sum over the elements of $\tilde{\mathcal{E}}$ and we can move the limit inside the sum. The fact that $\lim_{n_t \to \infty} |\frac{1-\beta(n_t)}{n_t+1} \sum_{i=1}^{n_t} \mathbb{I}(E' = X^{-1}(x_i)) - Q(E')| = 0$ a.s. for all $E' \in \tilde{\mathcal{E}}$ comes

from our assumptions that the data points are sampled independently, $\mathbb{E}(X) < \infty$, and $\beta(n_t) = o(1/n_t)$ as a result of the Glivenko-Cantelli Theorem. This concludes the proof. □

Proof of Proposition 28. We first point out that $\tilde{\mathcal{E}} = \tilde{\mathcal{E}}'$. This because, no matter the order in which we collect data points $x_i \in \mathcal{X}$, in the limit we either end up observing all the elements of $\mathcal{X}$, or all the elements of $\mathcal{X}_{reduced}$ in the case $\cup_{i\in\mathbb{N}} x_i = \mathcal{X}_{reduced} \subsetneq \mathcal{X}$. So if $\tilde{\mathcal{E}}$ is finer than $\tilde{\mathcal{E}}'$, this means that there exists an ω that is mapped by X into two different values, a contradiction. If instead $\tilde{\mathcal{E}}$ is coarser than $\tilde{\mathcal{E}}'$, this means that $\tilde{\mathcal{E}}$ can be further refined, which contradicts Proposition 24. Then, the claim follows by the uniqueness of the limit of a sequence. □

Proof of Proposition 30. Fix any $t \in \mathbb{N}$, and let $\mathcal{P}_{\mathcal{E}_t} = \{\check{P}_{k,\mathcal{E}_t}\}$. Pick any $P_{\mathcal{E}_t} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_t}$. Then, by the convexity of $\mathcal{P}^{\text{co}}_{\mathcal{E}_t}$, there exists a collection $\{\alpha_k\} \subset \mathbb{R}$ such that $\#\{\alpha_k\} = \#\mathcal{P}_{\mathcal{E}_t}$, $\sum_k \alpha_k = 1$, and $P_{\mathcal{E}_t}(A) = \sum_k \alpha_k \check{P}_{k,\mathcal{E}_t}(A)$, for all $A \in \mathcal{F}$. By construction and Theorem 25, given our assumptions we know that for all k,

$$d_{TV}(\check{P}_{k,\mathcal{E}_t}, \check{P}_{k,\tilde{\mathcal{E}}}) \to 0$$

as n_t goes to infinity with probability 1, where $\mathcal{P}_{\tilde{\mathcal{E}}} = \{\check{P}_{k,\tilde{\mathcal{E}}}\}$. So we can conclude that there is $P_{\tilde{\mathcal{E}}} \in \mathcal{P}^{\text{co}}_{\tilde{\mathcal{E}}}$ such that $P_{\tilde{\mathcal{E}}}(A) = \sum_k \alpha_k \check{P}_{k,\tilde{\mathcal{E}}}(A)$, for all $A \in \mathcal{F}$, and $d_{TV}(P_{\mathcal{E}_t}, P_{\tilde{\mathcal{E}}}) \to 0$ as n_t goes to infinity with probability 1.

That is to say that for every element $P_{\mathcal{E}_t}$ of $\mathcal{P}^{\text{co}}_{\mathcal{E}_t}$, there is an element $P_{\tilde{\mathcal{E}}}$ of $\mathcal{P}^{\text{co}}_{\tilde{\mathcal{E}}}$ that $P_{\mathcal{E}_t}$ converges to (with probability 1 in the total variation metric). This immediately implies that the Hausdorff distance between $\mathcal{P}^{\text{co}}_{\mathcal{E}_t}$ and $\mathcal{P}^{\text{co}}_{\tilde{\mathcal{E}}}$ goes to 0 as n_t goes to infinity with probability 1. □

Proof of Proposition 34. Fix any $A \in \mathcal{F}$ and any $t \in \mathbb{N}$. Notice that

$$\underline{P}_{\mathcal{E}_{t+1}}(A) := \inf_{P_{\mathcal{E}_{t+1}} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_{t+1}}} P_{\mathcal{E}_{t+1}}(A) = \inf_{\substack{P_{\mathcal{E}_{t+1}} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_{t+1}} \\ P_{\mathcal{E}_t} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_t}}} P_{\mathcal{E}_{t+1}}(A).$$

Then, we have that

$$
\begin{aligned}
\underline{P}_{\mathcal{E}_{t+1}}(A) &= \inf_{\substack{P_{\mathcal{E}_{t+1}} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_{t+1}} \\ P_{\mathcal{E}_t} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_t}}} \sum_{E_j \in \mathcal{E}_{t+1}} P_{\mathcal{E}_t}(A \mid E_j) P_{\mathcal{E}_{t+1}}(E_j) \\
&\geqslant \sum_{E_j \in \mathcal{E}_{t+1}} \inf_{\substack{P_{\mathcal{E}_{t+1}} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_{t+1}} \\ P_{\mathcal{E}_t} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_t}}} \left[P_{\mathcal{E}_t}(A \mid E_j) P_{\mathcal{E}_{t+1}}(E_j) \right] \qquad \text{(B.14)} \\
&\geqslant \sum_{E_j \in \mathcal{E}_{t+1}} \inf_{\substack{P_{\mathcal{E}_{t+1}} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_{t+1}} \\ P_{\mathcal{E}_t} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_t}}} P_{\mathcal{E}_t}(A \mid E_j) \inf_{\substack{P_{\mathcal{E}_{t+1}} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_{t+1}} \\ P_{\mathcal{E}_t} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_t}}} P_{\mathcal{E}_{t+1}}(E_j) \qquad \text{(B.15)} \\
&= \sum_{E_j \in \mathcal{E}_{t+1}} \inf_{\substack{P_{\mathcal{E}_{t+1}} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_{t+1}} \\ P_{\mathcal{E}_t} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_t}}} P_{\mathcal{E}_t}(A \mid E_j) \\
&\quad \cdot \inf_{\substack{P_{\mathcal{E}_{t+1}} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_{t+1}} \\ P_{\mathcal{E}_t} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_t}}} \left[\beta(n_t) P_{\mathcal{E}_t}(E_j) + (1 - \beta(n_t)) P^{emp}_{t+1}(E_j) \right] \qquad \text{(B.16)} \\
&= \sum_{E_j \in \mathcal{E}_{t+1}} \inf_{\substack{P_{\mathcal{E}_{t+1}} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_{t+1}} \\ P_{\mathcal{E}_t} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_t}}} P_{\mathcal{E}_t}(A \mid E_j) \\
&\quad \cdot \left[\beta(n_t) \inf_{\substack{P_{\mathcal{E}_{t+1}} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_{t+1}} \\ P_{\mathcal{E}_t} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_t}}} P_{\mathcal{E}_t}(E_j) + (1 - \beta(n_t)) P^{emp}_{t+1}(E_j) \right] \\
&= \sum_{E_j \in \mathcal{E}_{t+1}} \underline{P}^{B}_{\mathcal{E}_t}(A \mid E_j) \left[\beta(n_t) \underline{P}_{\mathcal{E}_t}(E_j) + (1 - \beta(n_t)) P^{emp}_{t+1}(E_j) \right].
\end{aligned}
$$

The inequality in (B.14) comes from the well known fact that the sum of the infima is at most equal to the infimum of the sum. The inequality in (B.15) comes from the fact that for differentiable functions, the product of the infima is at most equal to the infimum of the product. Equation (B.16) comes from equation (3.7). A similar argument – together with the facts that the supremum of the sum is at most equal to the sum of the suprema, and that for differentiable functions, the supremum of the product is at most equal to the product of the suprema – gives us the stated upper bound for $\overline{P}_{\mathcal{E}_{t+1}}(A)$. □

Proof of Corollary 35. Immediate from Proposition 34 and the definitions of upper and lower probabilities. □

Proof of Proposition 36. Pick any $A \in \mathcal{F}$ and any $t \in \mathbb{N}$. Then, we have that

$$\begin{aligned}
\underline{P}_{\mathcal{E}_{t+1}}(A) &\geqslant \sum_{E_j \in \mathcal{E}_{t+1}} \inf_{\substack{P_{\mathcal{E}_{t+1}} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_{t+1}} \\ P_{\mathcal{E}_t} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_t}}} P_{\mathcal{E}_t}(A \mid E_j) \\
&\quad \cdot \inf_{\substack{P_{\mathcal{E}_{t+1}} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_{t+1}} \\ P_{\mathcal{E}_t} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_t}}} \left[\beta(n_t) P_{\mathcal{E}_t}(E_j) + (1-\beta(n_t)) P^{emp}_{t+1}(E_j)\right] \qquad \text{(B.17)} \\
&= \sum_{E_j \in \mathcal{E}_{t+1} : P_{\mathcal{E}_t}(E_j) \neq 0} \inf_{\substack{P_{\mathcal{E}_{t+1}} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_{t+1}} \\ P_{\mathcal{E}_t} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_t}}} P_{\mathcal{E}_t}(A \mid E_j) \\
&\quad \cdot \inf_{\substack{P_{\mathcal{E}_{t+1}} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_{t+1}} \\ P_{\mathcal{E}_t} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_t}}} \left[\beta(n_t) P_{\mathcal{E}_t}(E_j) + (1-\beta(n_t)) P^{emp}_{t+1}(E_j)\right] \qquad \text{(B.18)} \\
&= \sum_{E_j \in \mathcal{E}_{t+1} : P_{\mathcal{E}_t}(E_j) \neq 0} \inf_{\substack{P_{\mathcal{E}_{t+1}} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_{t+1}} \\ P_{\mathcal{E}_t} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_t}}} P_{\mathcal{E}_t}(A \mid E_j) \\
&\quad \cdot \left[\beta(n_t) \underline{P}_{\mathcal{E}_t}(E_j) + (1-\beta(n_t)) P^{emp}_{t+1}(E_j)\right] \\
&= \sum_{E_j \in \mathcal{E}_{t+1} : P_{\mathcal{E}_t}(E_j) \neq 0} \inf_{\substack{P_{\mathcal{E}_{t+1}} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_{t+1}} \\ P_{\mathcal{E}_t} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_t}}} \frac{P_{\mathcal{E}_t}(A \cap E_j)}{P_{\mathcal{E}_t}(E_j)} \\
&\quad \cdot \left[\beta(n_t) \underline{P}_{\mathcal{E}_t}(E_j) + (1-\beta(n_t)) P^{emp}_{t+1}(E_j)\right] \\
&\geqslant \sum_{E_j \in \mathcal{E}_{t+1} : P_{\mathcal{E}_t}(E_j) \neq 0} \frac{\underline{P}_{\mathcal{E}_t}(A \cap E_j)}{\underline{P}_{\mathcal{E}_t}(E_j)} \left[\beta(n_t) \underline{P}_{\mathcal{E}_t}(E_j) + (1-\beta(n_t)) P^{emp}_{t+1}(E_j)\right] \qquad \text{(B.19)} \\
&= \sum_{E_j \in \mathcal{E}_{t+1} : P_{\mathcal{E}_t}(E_j) \neq 0} \underline{P}^{G}_{\mathcal{E}_t}(A \mid E_j) \left[\beta(n_t) \underline{P}_{\mathcal{E}_t}(E_j) + (1-\beta(n_t)) P^{emp}_{t+1}(E_j)\right].
\end{aligned}$$

The inequality in (B.17) comes from (B.16). Equality (B.18) comes from Remark 26. The inequality in (B.19) comes from the fact that for differentiable functions, the

product of the infima is at most equal to the infimum of the product. In particular,

$$\inf_{\substack{P_{\mathcal{E}_{t+1}} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_{t+1}} \\ P_{\mathcal{E}_t} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_t}}} P_{\mathcal{E}_t}(A \cap E_j) \frac{1}{P_{\mathcal{E}_t}(E_j)} \geqslant \underline{P}_{\mathcal{E}_t}(A \cap E_j) \frac{1}{\underline{P}_{\mathcal{E}_t}(E_j)},$$

for all $A \in \mathcal{F}$, all $E_j \in \mathcal{E}_{t+1}$, and all $t \in \mathbb{N}$. Notice that differentiability of $1/P_{\mathcal{E}_t}(E_j)$ is guaranteed by summing over the E_j's in $\mathcal{E}_{t+1}$ having nonzero $P_{\mathcal{E}_t}$-probability.

A similar argument gives us the stated upper bound for $\overline{P}_{\mathcal{E}_{t+1}}(A)$. □

Proof of Corollary 37. The interval in (3.20) comes from inequalities (3.18) and (3.19). The inequalities in (3.21) and (3.22) come from the inequalities in (3.17). □

Proof of Proposition 38. By Proposition 34, we have that

$$\underline{P}_{\mathcal{E}_t}(A) \geqslant \sum_{E_j \in \mathcal{E}_t} \underline{P}^B_{\mathcal{E}_{t-1}}(A \mid E_j) \left[\beta(n_t) \underline{P}_{\mathcal{E}_{t-1}}(E_j) + (1 - \beta(n_t)) P^{emp}_t(E_j)\right],$$

so if

$$\sum_{E_j \in \mathcal{E}_t} \underline{P}^B_{\mathcal{E}_{t-1}}(A \mid E_j) \left[\beta(n_t) \underline{P}_{\mathcal{E}_{t-1}}(E_j) + (1 - \beta(n_t)) P^{emp}_t(E_j)\right] \geqslant \underline{P}_{\mathcal{E}_{t-1}}(A),$$

then $\underline{P}_{\mathcal{E}_t}(A) \geqslant \underline{P}_{\mathcal{E}_{t-1}}(A)$. A similar reasoning gives us that

$$\sum_{E_j \in \mathcal{E}_t} \overline{P}^B_{\mathcal{E}_{t-1}}(A \mid E_j) \left[\beta(n_t) \overline{P}_{\mathcal{E}_{t-1}}(E_j) + (1 - \beta(n_t)) P^{emp}_t(E_j)\right] \leqslant \overline{P}_{\mathcal{E}_{t-1}}(A)$$

implies $\overline{P}_{\mathcal{E}_t}(A) \leqslant \overline{P}_{\mathcal{E}_{t-1}}(A)$. In turn, we obtain the desired DIPK-contraction. □

Proof of Proposition 39. By Proposition 34, we have that

$$\underline{P}_{\mathcal{E}_t}(A) \geqslant \sum_{E_j \in \mathcal{E}_t} \underline{P}^B_{\mathcal{E}_{t-1}}(A \mid E_j) \left[\beta(n_t) \underline{P}_{\mathcal{E}_{t-1}}(E_j) + (1 - \beta(n_t)) P^{emp}_t(E_j)\right],$$

so if

$$\sum_{E_j \in \mathcal{E}_t} \underline{P}^B_{\mathcal{E}_{t-1}}(A \mid E_j) \left[\beta(n_t) \underline{P}_{\mathcal{E}_{t-1}}(E_j) + (1 - \beta(n_t)) P^{emp}_t(E_j)\right] > \overline{P}_{\mathcal{E}_{t-1}}(A),$$

then $\underline{P}_{\mathcal{E}_t}(A) > \overline{P}_{\mathcal{E}_{t-1}}(A)$. A similar reasoning gives us that

$$\sum_{E_j \in \mathcal{E}_t} \overline{P}^B_{\mathcal{E}_{t-1}}(A \mid E_j) \left[\beta(n_t)\overline{P}_{\mathcal{E}_{t-1}}(E_j) + (1-\beta(n_t))P^{emp}_t(E_j)\right] < \underline{P}_{\mathcal{E}_{t-1}}(A)$$

implies $\overline{P}_{\mathcal{E}_t}(A) < \underline{P}_{\mathcal{E}_{t-1}}(A)$. In turn, we obtain the desired DIPK-sure loss. □

Proof of Proposition 40. Fix any $t \in \mathbb{N}$ and consider some $A \in \mathcal{F}$. By the definitions of lower and upper probabilities, we have that $P_{s,\mathcal{E}_t}(A'), P_{k,\mathcal{E}_t}(A') \in [\underline{P}_{\mathcal{E}_t}(A'), \overline{P}_{\mathcal{E}_t}(A')]$, for all $A' \in \mathcal{F}$. Then, if our hypotheses hold, we have that, for the set A we have chosen,

$$\underline{P}_{\mathcal{E}_{t-1}}(A) \geqslant P_{s,\mathcal{E}_t}(A) \geqslant \underline{P}_{\mathcal{E}_t}(A)$$

and

$$\overline{P}_{\mathcal{E}_{t-1}}(A) \leqslant P_{k,\mathcal{E}_t}(A) \leqslant \overline{P}_{\mathcal{E}_t}(A).$$

This concludes the proof. □

Proof of Proposition 41. Fix any $t \in \mathbb{N}$ and consider some $A \in \mathcal{F}$. By the definitions of lower and upper probabilities, we have that

$$P_{s,\mathcal{E}_{t-1}}(A'), P_{k,\mathcal{E}_{t-1}}(A') \in [\underline{P}_{\mathcal{E}_{t-1}}(A'), \overline{P}_{\mathcal{E}_{t-1}}(A')],$$

for all $A' \in \mathcal{F}$. Then, if our hypotheses hold, we have that, for the set A we have chosen,

$$\underline{P}_{\mathcal{E}_t}(A) \geqslant P_{k,\mathcal{E}_{t-1}}(A) \geqslant \underline{P}_{\mathcal{E}_{t-1}}(A)$$

and

$$\overline{P}_{\mathcal{E}_t}(A) \leqslant P_{s,\mathcal{E}_{t-1}}(A) \leqslant \overline{P}_{\mathcal{E}_{t-1}}(A).$$

This concludes the proof. □

Proof of Lemma 43. Given our assumptions, we have that for any $A' \in \mathcal{F}$ we can always find $P^{A'}_{\mathcal{E}_t}(A') \in \{P^A_{\mathcal{E}_t}\}_{A \in \mathcal{F}}$ such that

$$\underline{P}_{\mathcal{E}_t}(A') = P^{A'}_{\mathcal{E}_t}(A') = P^{A'}_{\mathcal{E}_t}(T^{-1}(A')) = \underline{P}_{\mathcal{E}_t}(T^{-1}(A')),$$

so $\underline{P}_{\mathcal{E}_t}$ is T-invariant. Because this holds for all $t \geqslant T$, it also holds for a collection $\{P^A_{\tilde{\mathcal{E}}}\}_{A\in\mathcal{F}}$ belonging to the almost sure limit $\mathcal{P}_{\tilde{\mathcal{E}}}$ of sequence $(\mathcal{P}_{\mathcal{E}_t})$. In turn, this implies that $\underline{P}_{\tilde{\mathcal{E}}}$ is T-invariant. □

Proof of Lemma 44. Suppose that there exist $T \in \mathbb{N}_0$ and $P'_{\mathcal{E}_T} \in \mathcal{P}_{\mathcal{E}_T}$ such that $P'_{\mathcal{E}_T}(A) = 0$, for all $A \in \mathcal{G}$. This implies that $\underline{P}_{\mathcal{E}_T}(A) = 0$. Then, we have that

$$\begin{aligned} P'_{\mathcal{E}_{T+1}}(A) &= \sum_{E\in\mathcal{E}_{T+1}} \frac{P'_{\mathcal{E}_T}(A\cap E)}{P'_{\mathcal{E}_T}(E)} P'_{\mathcal{E}_{T+1}}(E) \\ &\leqslant \sum_{E\in\mathcal{E}_{T+1}} \frac{P'_{\mathcal{E}_T}(A)}{P'_{\mathcal{E}_T}(E)} P'_{\mathcal{E}_{T+1}}(E) = 0. \end{aligned}$$

So $P'_{\mathcal{E}_{T+1}}(A) = 0$, which implies $\underline{P}_{\mathcal{E}_{T+1}}(A) = 0$. A similar argument shows that $P'_{\mathcal{E}_t}(A) = 0$, for all $t \geqslant T$, which implies that $P'_{\tilde{\mathcal{E}}}(A) = 0$. In turn, this implies that $\underline{P}_{\tilde{\mathcal{E}}}(A) = 0$. But because this is true for all $A \in \mathcal{G}$, we have that $\underline{P}_{\tilde{\mathcal{E}}}$ is ergodic. □

Proof of Theorem 45. From (Cerreia-Vioglio et al., 2015, Corollary 1), we know that if $\underline{P}_{\tilde{\mathcal{E}}}$ is invariant, then $\text{core}(\underline{P}_{\tilde{\mathcal{E}}}) = \mathcal{P}^{\text{co}}_{\tilde{\mathcal{E}}} \subset \mathcal{PI}$, where $\mathcal{PI}$ is the set of potentially invariant probability measures; that is, a probability measure P belongs to $\mathcal{PI}$ if and only if

$$\exists \hat{P} \in \mathcal{I} : P(E) = \hat{P}(E), \quad \forall E \in \mathcal{G}.$$

Then, (Cerreia-Vioglio et al., 2015, Theorem 5) ensures us that $\mathcal{P}^{\text{co}}_{\tilde{\mathcal{E}}} \subset \mathcal{PI}$ is equivalent to the fact that for all $f \in B(\Omega, \mathcal{F})$, there exists $f^\star \in B(\Omega, \mathcal{G})$ such that

$$\lim_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} f(T^{j-1}(\omega)) = f^\star(\omega) \quad \underline{P}_{\tilde{\mathcal{E}}} - a.s.$$

In particular, $f^\star \in B(\Omega, \mathcal{G})$ is defined as

$$\omega \mapsto f^\star(\omega) := \limsup_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} f(T^{j-1}(\omega)).$$

So we retrieve (4.1): the limit of the empirical averages exists and is finite $\underline{P}_{\tilde{\mathcal{E}}}$-almost surely.

Let us now show that, if $\underline{P}_{\tilde{\mathcal{E}}}$ is also ergodic, then $f^\star(\omega) \in \mathcal{A}(\omega)$ $\underline{P}_{\tilde{\mathcal{E}}}$-almost surely. Suppose for now that $f^\star \geqslant 0$. Since $\underline{P}_{\tilde{\mathcal{E}}}$ is a lower probability such that $\underline{P}_{\tilde{\mathcal{E}}}(\mathcal{G}) = \{0,1\}$, and $0 \leqslant f^\star \leqslant \lambda$ for some $\lambda \in \mathbb{R}$ (because $f^\star$ is bounded), then

$$I := \{t \in \mathbb{R}_+ : \underline{P}_{\tilde{\mathcal{E}}}\left(\{\omega \in \Omega : f^\star(\omega) \geqslant t\}\right) = 1\}$$

and

$$J := \{t \in \mathbb{R}_- : \underline{P}_{\tilde{\mathcal{E}}}\left(\{\omega \in \Omega : -f^\star(\omega) \geqslant t\}\right) = 1\}$$

are well defined nonempty intervals. I is bounded from above and such that $0 \in I$, and J is bounded from below and such that $-\lambda \in J$. Notice also that since $\underline{P}_{\tilde{\mathcal{E}}}$ is a lower probability, then it is continuous. We can conclude that $\sup I =: t^\star \in I$ and $\sup J =: t_\star \in J$. Since $\underline{P}_{\tilde{\mathcal{E}}}(\mathcal{G}) = \{0,1\}$, we have that

$$\sum_{\omega\in\Omega} f^\star(\omega)\underline{P}_{\tilde{\mathcal{E}}}(\{\omega\}) \leqslant \int_\Omega f^\star \, \mathrm{d}\underline{P}_{\tilde{\mathcal{E}}} = \int_0^\infty \underline{P}_{\tilde{\mathcal{E}}}\left(\{\omega \in \Omega : f^\star(\omega) \geqslant t\}\right) \, \mathrm{d}t = \int_0^{\sup I} \mathrm{d}t = t^\star$$

and

$$\begin{aligned}\sum_{\omega\in\Omega}[-f^\star(\omega)]\underline{P}_{\tilde{\mathcal{E}}}(\{\omega\}) &\geqslant \int_\Omega (-f^\star) \, \mathrm{d}\underline{P}_{\tilde{\mathcal{E}}} \\ &= \int_{-\infty}^0 \left[\underline{P}_{\tilde{\mathcal{E}}}\left(\{\omega \in \Omega : -f^\star(\omega) \geqslant t\}\right) - \underline{P}_{\tilde{\mathcal{E}}}(\Omega)\right] \, \mathrm{d}t = \int_{\sup J}^0 (-1) \, \mathrm{d}t = t_\star.\end{aligned}$$

So $t^\star \geqslant \sum_{\omega\in\Omega} f^\star(\omega)\underline{P}_{\tilde{\mathcal{E}}}(\{\omega\})$ and $t_\star \leqslant \sum_{\omega\in\Omega}[-f^\star(\omega)]\underline{P}_{\tilde{\mathcal{E}}}(\{\omega\})$. Now, since $t^\star \in I$ and $t_\star \in J$, we also have that

$$\underline{P}_{\tilde{\mathcal{E}}}\left(\{\omega \in \Omega : f^\star(\omega) \geqslant t^\star\}\right) = 1 = \underline{P}_{\tilde{\mathcal{E}}}\left(\{\omega \in \Omega : f^\star(\omega) \leqslant -t_\star\}\right).$$

Since $\underline{P}_{\tilde{\mathcal{E}}}$ is a lower probability, this implies that

$$\begin{aligned}&\underline{P}_{\tilde{\mathcal{E}}}\left(\left\{\omega \in \Omega : \sum_{\omega\in\Omega} f^\star(\omega)\underline{P}_{\tilde{\mathcal{E}}}(\{\omega\}) \leqslant f^\star(\omega) \leqslant \sum_{\omega\in\Omega} f^\star(\omega)\overline{P}_{\tilde{\mathcal{E}}}(\{\omega\})\right\}\right) \\ &\quad = \underline{P}_{\tilde{\mathcal{E}}}\left(\{\omega \in \Omega : t^\star \leqslant f^\star(\omega) \leqslant -t_\star\}\right) = 1.\end{aligned} \tag{B.20}$$

Let us now relax the assumption that $f^\star \geqslant 0$. Since $f^\star \in B(\Omega, \mathcal{G})$, then there exists $c \in \mathbb{R}$ such that $f^\star + c\mathbb{1}_\Omega \geqslant 0$. By (B.20), we have that

$$\underline{P}_{\tilde{\mathcal{E}}}\Bigg(\Bigg\{\omega \in \Omega : \sum_{\omega\in\Omega}(f^\star + c\mathbb{1}_\Omega)(\omega)\underline{P}_{\tilde{\mathcal{E}}}(\{\omega\}) \leqslant f^\star(\omega) + c \\ \leqslant \sum_{\omega\in\Omega}(f^\star + c\mathbb{1}_\Omega)(\omega)\overline{P}_{\tilde{\mathcal{E}}}(\{\omega\})\Bigg\}\Bigg)$$

$$= \underline{P}_{\tilde{\mathcal{E}}}\Bigg(\Bigg\{\omega \in \Omega : \Bigg[\sum_{\omega\in\Omega} f^\star(\omega)\underline{P}_{\tilde{\mathcal{E}}}(\{\omega\})\Bigg] + c \leqslant f^\star(\omega) + c \leqslant \Bigg[\sum_{\omega\in\Omega} f^\star(\omega)\overline{P}_{\tilde{\mathcal{E}}}(\{\omega\})\Bigg] + c\Bigg\}\Bigg)$$

$$= \underline{P}_{\tilde{\mathcal{E}}}\Bigg(\Bigg\{\omega \in \Omega : \sum_{\omega\in\Omega} f^\star(\omega)\underline{P}_{\tilde{\mathcal{E}}}(\{\omega\}) \leqslant f^\star(\omega) \leqslant \sum_{\omega\in\Omega} f^\star(\omega)\overline{P}_{\tilde{\mathcal{E}}}(\{\omega\})\Bigg\}\Bigg) = 1.$$

To conclude the proof, since by (4.1) we have that

$$\underline{P}_{\tilde{\mathcal{E}}}\left(\left\{\omega \in \Omega : f^\star(\omega) = \lim_{k\to\infty}\sum_{j=1}^{k} f(T^{j-1}(\omega))\right\}\right) = 1$$

and since $\underline{P}_{\tilde{\mathcal{E}}}$ is a lower probability, this implies that

$$\underline{P}_{\tilde{\mathcal{E}}}\left(\left\{\omega \in \Omega : \sum_{\omega\in\Omega} f^\star(\omega)\underline{P}_{\tilde{\mathcal{E}}}(\{\omega\}) \leqslant \lim_{k\to\infty}\frac{1}{k}\sum_{j=1}^{k} f(T^{j-1}(\omega)) \leqslant \sum_{\omega\in\Omega} f^\star(\omega)\overline{P}_{\tilde{\mathcal{E}}}(\{\omega\})\right\}\right)$$

is equal to 1, retrieving equation (4.2). □

Proof of Lemma 46. Pick any $A, B \in \mathcal{F}$ such that $A \subset B$. If (i) holds, it is immediate to see that for all $t \geqslant T$, $P'_{\mathcal{E}_{t+1}}(A) \leqslant P_{\mathcal{E}_{t+1}}(A)$ and $P'_{\mathcal{E}_{t+1}}(B) \leqslant P_{\mathcal{E}_{t+1}}(B)$, for all $P_{\mathcal{E}_{t+1}} \in \mathcal{P}^{\text{co}}_{\mathcal{E}_{t+1}}$. This implies that $P'_{\mathcal{E}_{t+1}}(A) = \underline{P}_{\mathcal{E}_{t+1}}(A)$ and $P'_{\mathcal{E}_{t+1}}(B) = \underline{P}_{\mathcal{E}_{t+1}}(B)$. In turn, because this holds for all $t \geqslant T$, we have that the almost sure limit $P'_{\tilde{\mathcal{E}}}$ of $P'_{\mathcal{E}_{t+1}}$ is such that $P'_{\tilde{\mathcal{E}}}(A) = \underline{P}_{\tilde{\mathcal{E}}}(A)$ and $P'_{\tilde{\mathcal{E}}}(B) = \underline{P}_{\tilde{\mathcal{E}}}(B)$. This implies that $\underline{P}_{\tilde{\mathcal{E}}}$ is convex by (Marinacci and Montrucchio, 2004b, Theorem 38.(ii)).

Pick any finite chain $(A_i)_{i=1}^n \subset \mathcal{F}$. If (ii) holds, it is immediate to see that for all $t \geqslant T$, $P'_{\mathcal{E}_{t+1}}(A_i) \leqslant P_{\mathcal{E}_{t+1}}(A_i)$, for all $i \in \{1, \ldots, n\}$. This implies that $P'_{\mathcal{E}_{t+1}}(A_i) =$

$\underline{P}_{\mathcal{E}_{t+1}}(A_i)$, for all $i \in \{1, \ldots, n\}$. In turn, because this holds for all $t \geqslant T$, we have that the almost sure limit $P'_{\tilde{\mathcal{E}}}$ of $P'_{\mathcal{E}_{t+1}}$ is such that $P'_{\tilde{\mathcal{E}}}(A_i) = \underline{P}_{\tilde{\mathcal{E}}}(A_i)$, for all $i \in \{1, \ldots, n\}$. This implies that $\underline{P}_{\tilde{\mathcal{E}}}$ is convex by (Marinacci and Montrucchio, 2004b, Theorem 38.(iii)). A similar argument combined with (Marinacci and Montrucchio, 2004b, Theorem 38.(iv)) gives us that condition (iii) implies $\underline{P}_{\tilde{\mathcal{E}}}$ being convex. □

Proof of Lemma 47. (i) By (Marinacci and Montrucchio, 2004b, Theorem 10), we have that if $\text{core}(\underline{P}_{\tilde{\mathcal{E}}}) = \mathcal{P}^{\text{co}}_{\tilde{\mathcal{E}}}$ is nonempty, $\underline{P}_{\tilde{\mathcal{E}}}(A) = \min_{P \in \mathcal{P}^{\text{co}}_{\tilde{\mathcal{E}}}} P(A)$, for all $A \in \mathcal{F}$, and $\mathcal{P}^{\text{co}}_{\tilde{\mathcal{E}}}$ is a weakly compact subset of the space $ca(\mathcal{F})$ of all measures having finite total variation norm, then $\underline{P}_{\tilde{\mathcal{E}}}$ is continuous at Ω. Because these conditions are always satisfied, we conclude that $\underline{P}_{\tilde{\mathcal{E}}}$ is always continuous at Ω.

(ii) Immediate from our assumption and the fact that $(\mathcal{P}_{\mathcal{E}_t})$ converges (almost surely) to $\mathcal{P}_{\tilde{\mathcal{E}}}$.

(iii) If the assumptions in (iii) hold, we have that

$$\begin{aligned} P_{\mathcal{E}_{t+1}}(A) &= \sum_{E \in \mathcal{E}_{t+1}} \frac{P_{\mathcal{E}_t}(A \cap E)}{P_{\mathcal{E}_t}(E)} \left[\beta(n) P_{\mathcal{E}_t}(E) + (1 - \beta(n)) P^{emp}_{t+1}(E)\right] \\ &= \sum_{E \in \mathcal{E}_{t+1}} \frac{P_{\mathcal{E}_t}(T^{-1}(A) \cap E)}{P_{\mathcal{E}_t}(E)} \left[\beta(n) P_{\mathcal{E}_t}(E) + (1 - \beta(n)) P^{emp}_{t+1}(E)\right] \\ &= P_{\mathcal{E}_{t+1}}(T^{-1}(A)), \end{aligned}$$

for all $A \in \mathcal{F}$ and all $P_{\mathcal{E}_t} \in \mathcal{P}_{\mathcal{E}_t}$. This implies that $\mathcal{P}_{\mathcal{E}_t} \subset \mathcal{I}$. Because this is true for all $t \geqslant T$, we have that the almost sure limit $\mathcal{P}_{\tilde{\mathcal{E}}}$ too is a subset of $\mathcal{I}$, which implies that $\underline{P}_{\tilde{\mathcal{E}}}$ is functionally invariant.

□

Proof of Lemma 48. We know that (S_k) satisfies (4.4), so $(S_k) \subset B(\Omega, \mathcal{F})$. If (S_k) is

superadditive and $\mathcal{P}_{\tilde{\mathcal{E}}} \subset \mathcal{I}$, then, for all $k, \ell \in \mathbb{N}$, we have the following

$$\begin{aligned}
-a_{k+\ell} &= \sum_{\omega\in\Omega} \inf_{P\in\mathcal{P}_{\tilde{\mathcal{E}}}} S_{k+\ell}(\omega)P(\{\omega\}) \geqslant \sum_{\omega\in\Omega} \inf_{P\in\mathcal{P}_{\tilde{\mathcal{E}}}} \left[S_k(\omega) + (S_\ell \circ T^k)(\omega)\right] P(\{\omega\}) \\
&\geqslant \sum_{\omega\in\Omega} \inf_{P\in\mathcal{P}_{\tilde{\mathcal{E}}}} S_k(\omega)P(\{\omega\}) + \sum_{\omega\in\Omega} \inf_{P\in\mathcal{P}_{\tilde{\mathcal{E}}}} (S_\ell \circ T^k)(\omega)P(\{\omega\}) \\
&= \sum_{\omega\in\Omega} \inf_{P\in\mathcal{P}_{\tilde{\mathcal{E}}}} S_k(\omega)P(\{\omega\}) + \sum_{\omega\in\Omega} \inf_{P\in\mathcal{P}_{\tilde{\mathcal{E}}}} S_\ell(\omega)P(\{\omega\}) = -a_k - a_\ell.
\end{aligned}$$

A similar procedure shows the result when (S_k) is subadditive and a_k is the sum of the suprema over $\mathcal{P}_{\tilde{\mathcal{E}}}$ of $S_k(\omega)P(\{\omega\})$. □

Proof of Theorem 49. Given our assumption that $\underline{P}_{\tilde{\mathcal{E}}}$ is functionally invariant, we have that $\mathcal{P}^{\mathrm{co}}_{\tilde{\mathcal{E}}} \subset \mathcal{I}$. Define the sequence (f_k) such that $f_k(\omega) := \frac{1}{k} S_k(\omega)$, for all $k \in \mathbb{N}$ and all $\omega \in \Omega$. By (S_k) satisfying (4.4), this implies that $f_k \in B(\Omega, \mathcal{F})$, for all k. Consider then a function $p : \mathcal{F} \times \Omega \to [0, 1]$ such that

- for all $P \in \mathcal{P}_{\tilde{\mathcal{E}}}$ and all $A \in \mathcal{F}$, $p(A, \cdot) : \Omega \to [0, 1]$ is a version of the conditional probability of A given $\mathcal{G}$;
- for all $\omega \in \Omega$, $p(\cdot, \omega) : \mathcal{F} \to [0, 1]$ is a probability measure;
- for all $\omega \in \Omega$, $p(\cdot, \omega) \in \mathcal{P}_{\tilde{\mathcal{E}}}$.

For all $k \in \mathbb{N}$, define then

$$\hat{f}_k : \Omega \to \mathbb{R}, \quad \omega \mapsto \hat{f}_k(\omega) := \sum_{\omega'\in\Omega} f_k(\omega')p(\{\omega'\}, \omega). \tag{B.21}$$

Given that $f_k \in B(\Omega, \mathcal{F})$, for all k, this implies that $\hat{f}_k \in B(\Omega, \mathcal{G})$, for all k. Since (S_k) satisfies (4.4), it follows that there exists $\lambda \in \mathbb{R}$ such that $-\lambda \leqslant f_k, \hat{f}_k \leqslant \lambda$, for all $k \in \mathbb{N}$. Define now $f^\star \in B(\Omega, \mathcal{G})$ by $f^\star := \sup_{k\in\mathbb{N}} \hat{f}_k$. By Kingman's Subadditive Ergodic Theorem (Dudley, 2002, Theorem 10.7.1) and (Gray, 2009, Theorem 8.4), we have that $f^\star = \lim_{k\to\infty} \hat{f}_k$ and $\lim_{k\to\infty} \frac{1}{k} S_k(\omega) = f^\star(\omega)$ P-almost surely, for all

$P \in \mathcal{P}_{\tilde{\mathcal{E}}}$. Since $\underline{P}_{\tilde{\mathcal{E}}}$ is a lower probability, it follows that $\lim_{k\to\infty} \frac{1}{k} S_k(\omega) = f^\star(\omega)$ $\underline{P}_{\tilde{\mathcal{E}}}$-almost surely. This shows the first part of the theorem. Let us now show claims (1)–(3).

(1). If $\underline{P}_{\tilde{\mathcal{E}}}$ is convex and strongly invariant, by (Cerreia-Vioglio et al., 2015, Theorem 1) we have that $\text{core}(\underline{P}_{\tilde{\mathcal{E}}}) = \mathcal{P}^{\text{co}}_{\tilde{\mathcal{E}}} \subset \mathcal{I}$. We also have that

$$\sum_{\omega\in\Omega} f(\omega)\underline{P}_{\tilde{\mathcal{E}}}(\{\omega\}) = \sum_{\omega\in\Omega} \inf_{P\in\mathcal{P}^{\text{co}}_{\tilde{\mathcal{E}}}} f(\omega)P(\{\omega\}), \quad \forall f \in B(\Omega,\mathcal{F}). \tag{B.22}$$

Consider now the sequence (a_k) defined as $a_k := -\sum_{\omega\in\Omega} S_k(\omega)\underline{P}_{\tilde{\mathcal{E}}}(\{\omega\})$, for all $k \in \mathbb{N}$. By (B.22) and Lemma 48, it follows that (a_k) is subadditive. By (Gray, 2009, Lemma 8.3), this implies that

$$\lim_{k\to\infty} \frac{1}{k}(-a_k) = \sup_{k\in\mathbb{N}} \frac{1}{k}(-a_k). \tag{B.23}$$

Now, by the fact that $(\hat{f}_n)$ is uniformly bounded, (B.23), the first part of the theorem, and the fact that $\mathcal{P}^{\text{co}}_{\tilde{\mathcal{E}}} \subset \mathcal{I}$, the following equalities hold

$$\begin{aligned}
\sum_{\omega\in\Omega} f^\star(\omega)\underline{P}_{\tilde{\mathcal{E}}}(\{\omega\}) &= \sum_{\omega\in\Omega} \lim_{k\to\infty} \hat{f}_k(\omega)\underline{P}_{\tilde{\mathcal{E}}}(\{\omega\}) = \lim_{k\to\infty} \sum_{\omega\in\Omega} \hat{f}_k(\omega)\underline{P}_{\tilde{\mathcal{E}}}(\{\omega\}) \\
&= \lim_{k\to\infty} \left[\sum_{\omega\in\Omega} \inf_{P\in\mathcal{P}^{\text{co}}_{\tilde{\mathcal{E}}}} \hat{f}_k(\omega)P(\{\omega\}) \right] \\
&= \lim_{k\to\infty} \left[\sum_{\omega\in\Omega} \inf_{P\in\mathcal{P}^{\text{co}}_{\tilde{\mathcal{E}}}} f_k(\omega)P(\{\omega\}) \right] \\
&= \lim_{k\to\infty} \sum_{\omega\in\Omega} f_k(\omega)\underline{P}_{\tilde{\mathcal{E}}}(\{\omega\}) \\
&= \lim_{k\to\infty} \frac{1}{k} \sum_{\omega\in\Omega} S_k(\omega)\underline{P}_{\tilde{\mathcal{E}}}(\{\omega\}) \\
&= \lim_{k\to\infty} \frac{1}{k}(-a_k) = \sup_{k\in\mathbb{N}} \frac{1}{k}(-a_k) \\
&= \sup_{k\in\mathbb{N}} \frac{1}{k} \sum_{\omega\in\Omega} S_k(\omega)\underline{P}_{\tilde{\mathcal{E}}}(\{\omega\}),
\end{aligned}$$

concluding the proof of (1).

(2). If $\underline{P}_{\tilde{\mathcal{E}}}$ is convex and strongly invariant, by (Cerreia-Vioglio et al., 2015, Theorem 1) we have that $\text{core}(\underline{P}_{\tilde{\mathcal{E}}}) \subset \mathcal{I}$. We also have that

$$\sum_{\omega\in\Omega} f(\omega)\overline{P}_{\tilde{\mathcal{E}}}(\{\omega\}) = \sum_{\omega\in\Omega} \sup_{P\in\mathcal{P}^{\text{co}}_{\tilde{\mathcal{E}}}} f(\omega)P(\{\omega\}), \quad \forall f \in B(\Omega,\mathcal{F}). \tag{B.24}$$

Consider now the sequence (a_k) defined as $a_k := \sum_{\omega\in\Omega} S_k(\omega)\overline{P}_{\tilde{\mathcal{E}}}(\{\omega\})$, for all $k \in \mathbb{N}$. By (B.24) and Lemma 48, it follows that (a_k) is subadditive. By (Gray, 2009, Lemma 8.3), this implies that

$$\lim_{k\to\infty} \frac{1}{k} a_k = \inf_{k\in\mathbb{N}} \frac{1}{k} a_k. \tag{B.25}$$

Now, by the fact that $(\hat{f}_n)$ is uniformly bounded, (B.25), the first part of the theorem, and the fact that $\mathcal{P}^{\text{co}}_{\tilde{\mathcal{E}}} \subset \mathcal{I}$, the following equalities hold

$$\begin{aligned}
\sum_{\omega\in\Omega} f^\star(\omega)\overline{P}_{\tilde{\mathcal{E}}}(\{\omega\}) &= \sum_{\omega\in\Omega} \lim_{k\to\infty} \hat{f}_k(\omega)\overline{P}_{\tilde{\mathcal{E}}}(\{\omega\}) = \lim_{k\to\infty} \sum_{\omega\in\Omega} \hat{f}_k(\omega)\overline{P}_{\tilde{\mathcal{E}}}(\{\omega\}) \\
&= \lim_{k\to\infty} \left[\sum_{\omega\in\Omega} \sup_{P\in\mathcal{P}^{\text{co}}_{\tilde{\mathcal{E}}}} \hat{f}_k(\omega)P(\{\omega\}) \right] \\
&= \lim_{k\to\infty} \left[\sum_{\omega\in\Omega} \sup_{P\in\mathcal{P}^{\text{co}}_{\tilde{\mathcal{E}}}} f_k(\omega)P(\{\omega\}) \right] \\
&= \lim_{k\to\infty} \sum_{\omega\in\Omega} f_k(\omega)\overline{P}_{\tilde{\mathcal{E}}}(\{\omega\}) \\
&= \lim_{k\to\infty} \frac{1}{k} \sum_{\omega\in\Omega} S_k(\omega)\overline{P}_{\tilde{\mathcal{E}}}(\{\omega\}) \\
&= \lim_{k\to\infty} \frac{1}{k} a_k = \inf_{k\in\mathbb{N}} \frac{1}{k} a_k \\
&= \inf_{k\in\mathbb{N}} \frac{1}{k} \sum_{\omega\in\Omega} S_k(\omega)\overline{P}_{\tilde{\mathcal{E}}}(\{\omega\}),
\end{aligned}$$

concluding the proof of (2).

(3). If $\underline{P}_{\tilde{\mathcal{E}}}$ is ergodic, using the same technique as in the proof of Theorem 45 we can show that

$$\underline{P}_{\tilde{\mathcal{E}}}\left(\left\{\omega\in\Omega:\sum_{\omega\in\Omega}f^{\star}(\omega)\underline{P}_{\tilde{\mathcal{E}}}(\{\omega\})\leqslant f^{\star}(\omega)\leqslant\sum_{\omega\in\Omega}f^{\star}(\omega)\overline{P}_{\tilde{\mathcal{E}}}(\{\omega\})\right\}\right)=1.$$

By the first part of the theorem, we have that

$$\underline{P}_{\tilde{\mathcal{E}}}\left(\left\{\omega\in\Omega:\lim_{k\to\infty}\frac{1}{k}S_k(\omega)=f^{\star}(\omega)\right\}\right)=1.$$

Since $\underline{P}_{\tilde{\mathcal{E}}}$ is a lower probability, this implies that

$$\underline{P}_{\tilde{\mathcal{E}}}\left(\left\{\omega\in\Omega:\sum_{\omega\in\Omega}f^{\star}(\omega)\underline{P}_{\tilde{\mathcal{E}}}(\{\omega\})\leqslant\lim_{k\to\infty}\frac{1}{k}S_k(\omega)\leqslant\sum_{\omega\in\Omega}f^{\star}(\omega)\overline{P}_{\tilde{\mathcal{E}}}(\{\omega\})\right\}\right)=1.$$

□

Proof of Corollary 50. Consider $f\in B(\Omega,\mathcal{F})$. We first notice that (S_k) such that $S_k:=\sum_{j=1}^{k}f\circ T^{j-1}$, for all $k\in\mathbb{N}$, is an additive sequence satisfying (4.4). Now, since $\underline{P}_{\tilde{\mathcal{E}}}$ is strongly invariant, then it is also functionally invariant by (Cerreia-Vioglio et al., 2015, Theorem 1). Consider now the sequence (f_k) where $f_k:=S_k/k$, for all $k\in\mathbb{N}$. Notice that $\hat{f}_k=\hat{f}$, for all $k\in\mathbb{N}$, where $\hat{f}_k$ is defined as in (B.21), $\hat{f}:\Omega\to\mathbb{R}$, $\omega\mapsto\hat{f}(\omega):=\sum_{\omega'\in\Omega}f(\omega')p(\{\omega'\},\omega)$, and $p:\mathcal{F}\times\Omega\to[0,1]$ is defined as in the proof of Theorem 49. Then, by the proof of Theorem 49, it follows that $\lim_{k\to\infty}\frac{1}{k}S_k=\lim_{k\to\infty}\hat{f}_k=\hat{f}$, $\underline{P}_{\tilde{\mathcal{E}}}$-almost surely. This proves the first part of the corollary, and also point (1) by setting $f^{\star}=\hat{f}$. Let us now show claims (2) and (3).

(2). Since $\underline{P}_{\tilde{\mathcal{E}}}$ is convex and strongly invariant, then $\mathcal{P}^{\mathrm{co}}_{\tilde{\mathcal{E}}}\subset\mathcal{I}$ and

$$\sum_{\omega\in\Omega}f(\omega)\underline{P}_{\tilde{\mathcal{E}}}(\{\omega\})=\sum_{\omega\in\Omega}\inf_{P\in\mathcal{P}^{\mathrm{co}}_{\tilde{\mathcal{E}}}}f(\omega)P(\{\omega\}).$$

By (1) and $\mathcal{P}^{\text{co}}_{\tilde{\mathcal{E}}} \subset \mathcal{I}$, we have that

$$\sum_{\omega\in\Omega} f(\omega)\underline{P}_{\tilde{\mathcal{E}}}(\{\omega\}) = \sum_{\omega\in\Omega} \inf_{P\in\mathcal{P}^{\text{co}}_{\tilde{\mathcal{E}}}} f(\omega)P(\{\omega\})$$

$$= \sum_{\omega\in\Omega} \inf_{P\in\mathcal{P}^{\text{co}}_{\tilde{\mathcal{E}}}} \hat{f}(\omega)P(\{\omega\}) = \sum_{\omega\in\Omega} \hat{f}(\omega)\underline{P}_{\tilde{\mathcal{E}}}(\{\omega\}),$$

concluding the proof of (2). Notice also that

$$\sum_{\omega\in\Omega} f(\omega)\overline{P}_{\tilde{\mathcal{E}}}(\{\omega\}) = \sum_{\omega\in\Omega} \sup_{P\in\mathcal{P}^{\text{co}}_{\tilde{\mathcal{E}}}} f(\omega)P(\{\omega\})$$

$$= \sum_{\omega\in\Omega} \sup_{P\in\mathcal{P}^{\text{co}}_{\tilde{\mathcal{E}}}} \hat{f}(\omega)P(\{\omega\}) = \sum_{\omega\in\Omega} \hat{f}(\omega)\overline{P}_{\tilde{\mathcal{E}}}(\{\omega\}).$$

(3). By Theorem 49.(3) and the proof of claim (2), claim (3) follows. □

Proof of Theorem 52. This proof is very similar to the one of (Cerreia-Vioglio et al., 2015, Theorem 4). The main difference is that we expressly request that Ω is finite or countable. By assumption, we have that $\mathbf{f}$ is stationary. Then, by a mathematical induction argument, we have that for all $k \in \mathbb{N}$ and all Borel subset B of $\mathbb{R}$,

$$\underline{P}_{\tilde{\mathcal{E}}}(\{\omega \in \Omega : f_1(\omega) \in B\}) = \underline{P}_{\tilde{\mathcal{E}}}(\{\omega \in \Omega : f_k(\omega) \in B\}). \tag{B.26}$$

Equation (B.26) implies that for all $k \in \mathbb{N}$ and all Borel subset B of $\mathbb{R}$,

$$\underline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}}(\{x \in \mathbb{R}^{\mathbb{N}} : x_k \in B\}) = \underline{P}_{\tilde{\mathcal{E}}}(\{\omega \in \Omega : f_k(\omega) \in B\}) = \underline{P}_{\tilde{\mathcal{E}}}(\{\omega \in \Omega : f_1(\omega) \in B\}).$$

Now, since $(f_k) \subset B(\Omega, \mathcal{F})$, we hve that there exists $m \in \mathbb{R}$ such that $-m\mathbb{1}_\Omega \leqslant f_1 \leqslant m\mathbb{1}_\Omega$. By replacing B with $[-m, m]$, we obtain that

$$\underline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}}(\{x \in \mathbb{R}^{\mathbb{N}} : x_k \in [-m, m]\}) = \underline{P}_{\tilde{\mathcal{E}}}(\{\omega \in \Omega : f_1(\omega) \in [-m, m]\}) = 1, \quad \forall k \in \mathbb{N}. \tag{B.27}$$

Let us now define the function $\pi : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ as

$$x \mapsto \pi(x) := \begin{cases} x_1 & \text{if } x_1 \in [-m, m] \\ 0 & \text{otherwise} \end{cases}.$$

We immediately see that π belongs to $B(\mathbb{R}^{\mathbb{N}}, \sigma(\mathcal{C}))$. Notice also that

$$\bigcap_{k=1}^{\infty} \left\{x \in \mathbb{R}^{\mathbb{N}} : x_k \in [-m, m]\right\} \subset \bigcap_{k=1}^{\infty} \left\{x \in \mathbb{R}^{\mathbb{N}} : \frac{1}{k}\sum_{j=1}^{k} \pi(s^{j-1}(x)) = \frac{1}{k}\sum_{j=1}^{k} x_j\right\}. \quad \text{(B.28)}$$

Given (B.27) and (B.28), and since $\underline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}}$ is both convex and continuous at Ω, we have that

$$\underline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}}\left(\bigcap_{k=1}^{\infty} \left\{x \in \mathbb{R}^{\mathbb{N}} : \frac{1}{k}\sum_{j=1}^{k} \pi(s^{j-1}(x)) = \frac{1}{k}\sum_{j=1}^{k} x_j\right\}\right) = 1. \quad \text{(B.29)}$$

By (Cerreia-Vioglio et al., 2015, Theorem 2) and the fact that $\underline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}}$ is shift invariant and ergodic, then there exists $\pi^* \in B(\mathbb{R}^{\mathbb{N}}, \mathcal{G})$ such that

$$\underline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}}\left(\left\{x \in \mathbb{R}^{\mathbb{N}} : \int_{\mathbb{R}^{\mathbb{N}}} \pi^\star \, \mathrm{d}\underline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}} \leqslant \lim_{k\to\infty} \frac{1}{k}\sum_{j=1}^{k} \pi(s^{j-1}(x)) = \pi^\star(x) \leqslant \int_{\mathbb{R}^{\mathbb{N}}} \pi^\star \, \mathrm{d}\overline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}}\right\}\right) = 1. \quad \text{(B.30)}$$

Then, by (B.29), (B.30), and the fact that $\underline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}}$ is convex, we have that

$$\underline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}}\left(\left\{x \in \mathbb{R}^{\mathbb{N}} : \int_{\mathbb{R}^{\mathbb{N}}} \pi^\star \, \mathrm{d}\underline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}} \leqslant \lim_{k\to\infty} \frac{1}{k}\sum_{j=1}^{k} x_j = \pi^\star(x) \leqslant \int_{\mathbb{R}^{\mathbb{N}}} \pi^\star \, \mathrm{d}\overline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}}\right\}\right) = 1. \quad \text{(B.31)}$$

Define now define the set $E := \{x \in \mathbb{R}^{\mathbb{N}} : \lim_{k\to\infty} \frac{1}{k}\sum_{j=1}^{k} \pi(s^{j-1}(x)) = \pi^\star(x)\}$ and the function $\pi_k := \frac{1}{k}\sum_{j=1}^{k} \pi(s^{j-1})$, for all $k \in \mathbb{N}$. Then, by (B.30) we have that $P(E) = 1$, for all $P \in \mathrm{core}(\underline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}})$. By construction, $(\mathbb{1}_E \pi_k)_{k\in\mathbb{N}} \subset B(\mathbb{R}^{\mathbb{N}}, \sigma(\mathcal{C}))$ is uniformly bounded and converges (pointwise) to $\mathbb{1}_E \pi^\star$. By $\underline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}}$ being convex, $P(E)$ being 1 for all $P \in \mathrm{core}(\underline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}})$, and (Cerreia-Vioglio et al., 2012, Theorem 22), it follows that that

$$\begin{aligned}\int_{\mathbb{R}^{\mathbb{N}}} \pi^\star \, \mathrm{d}\underline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}} &= \int_{\mathbb{R}^{\mathbb{N}}} \mathbb{1}_E \pi^\star \, \mathrm{d}\underline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}} = \int_{\mathbb{R}^{\mathbb{N}}} \lim_{k\to\infty} \mathbb{1}_E \pi_k \, \mathrm{d}\underline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}} \\ &= \lim_{k\to\infty} \int_{\mathbb{R}^{\mathbb{N}}} \mathbb{1}_E \pi_k \, \mathrm{d}\underline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}} = \lim_{k\to\infty} \int_{\mathbb{R}^{\mathbb{N}}} \pi_k \, \mathrm{d}\underline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}}.\end{aligned} \quad \text{(B.32)}$$

Then, because $\underline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}}$ is convex and shift invariant, we have that, for all $n \in \mathbb{N}$,

$$\int_{\mathbb{R}^{\mathbb{N}}} \pi_k \, \mathrm{d}\underline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}} = \int_{\mathbb{R}^{\mathbb{N}}} \frac{1}{k} \sum_{j=1}^{k} \pi(s^{j-1}) \, \mathrm{d}\underline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}} \geqslant \frac{1}{k} \sum_{j=1}^{k} \int_{\mathbb{R}^{\mathbb{N}}} \pi(s^{j-1}) \, \mathrm{d}\underline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}} = \int_{\mathbb{R}^{\mathbb{N}}} \pi \, \mathrm{d}\underline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}}.$$

Then, by (B.32), we have that $\int_{\mathbb{R}^{\mathbb{N}}} \pi^{\star} \, \mathrm{d}\underline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}} \geqslant \int_{\mathbb{R}^{\mathbb{N}}} \pi \, \mathrm{d}\underline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}}$. A similar argument shows that $\int_{\mathbb{R}^{\mathbb{N}}} \pi^{\star} \, \mathrm{d}\overline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}} \leqslant \int_{\mathbb{R}^{\mathbb{N}}} \pi \, \mathrm{d}\overline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}}$. Now, since by construction

$$\int_{\mathbb{R}^{\mathbb{N}}} \pi \, \mathrm{d}\underline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}} \geqslant \sum_{\omega \in \Omega} f_1(\omega) \underline{P}_{\tilde{\mathcal{E}}}(\{\omega\}) \quad \text{and} \quad \int_{\mathbb{R}^{\mathbb{N}}} \pi \, \mathrm{d}\overline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}} \leqslant \sum_{\omega \in \Omega} f_1(\omega) \overline{P}_{\tilde{\mathcal{E}}}(\{\omega\}),$$

we have that (B.31) gives us

$$\begin{aligned} 1 &= \underline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}} \left(\left\{ x \in \mathbb{R}^{\mathbb{N}} : \int_{\mathbb{R}^{\mathbb{N}}} \pi \, \mathrm{d}\underline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}} \leqslant \lim_{k \to \infty} \frac{1}{k} \sum_{j=1}^{k} x_j \leqslant \int_{\mathbb{R}^{\mathbb{N}}} \pi \, \mathrm{d}\overline{P}^{\mathbf{f}}_{\tilde{\mathcal{E}}} \right\} \right) \\ &= \underline{P}_{\tilde{\mathcal{E}}} \left(\left\{ \omega \in \Omega : \sum_{\omega \in \Omega} f_1(\omega) \underline{P}_{\tilde{\mathcal{E}}}(\{\omega\}) \leqslant \lim_{k \to \infty} \frac{1}{k} \sum_{j=1}^{k} f_j(\omega) \leqslant \sum_{\omega \in \Omega} f_1(\omega) \overline{P}_{\tilde{\mathcal{E}}}(\{\omega\}) \right\} \right), \end{aligned}$$

which proves the statement. □

Proof of Proposition 54. Suppose for the sake of contradiction that for some $A = B$,

$$P^{ex}(A) \neq P^{ex}(B). \tag{B.33}$$

Consider then $A \backslash B = B \backslash B = \varnothing$; we have that

$$P^{ex}(A \backslash B) \neq P^{ex}(B \backslash B) = P^{ex}(\varnothing) = 0. \tag{B.34}$$

This is true because $A = \{A \backslash B\} \sqcup \{A \cap B\}$ and $B = \{B \backslash B\} \sqcup \{B \cap B\}$; by (ii*), this implies that $P^{ex}(A \backslash B) = P^{ex}(A) - P^{ex}(A \cap B)$ and $P^{ex}(B \backslash B) = P^{ex}(B) - P^{ex}(B \cap B)$. Then, we have that $P^{ex}(A \backslash B) = P^{ex}(B \backslash B)$ if and only if

$$P^{ex}(A) = P^{ex}(B) - P^{ex}(B \cap B) + P^{ex}(A \cap B);$$

but $B \cap B = B$, and $A \cap B = B$ because we assumed $A = B$, so $P^{ex}(A \backslash B) = P^{ex}(B \backslash B)$ if and only if $P^{ex}(A) = P^{ex}(B)$. But by (B.33) $P^{ex}(A) \neq P^{ex}(B)$, so we have that the inequality in (B.34) holds.

By assumption we know that $A = B$, so $A \backslash B = \varnothing$, and so $P^{ex}(A \backslash B) = P^{ex}(\varnothing)$. Then, by (B.34), we have that $P^{ex}(\varnothing) = P^{ex}(A \backslash B) \neq P^{ex}(B \backslash B) = P^{ex}(\varnothing)$, a contradiction. □

Proof of Proposition 55. Let $A \subset B$; then, we can write B as $B = A \sqcup (B \cap A^c)$. By (ii*) this implies that $P^{ex}(B) = P^{ex}(A) + P^{ex}(B \cap A^c)$. Then, if $P^{ex}(B)$, $P^{ex}(A)$, and $P^{ex}(B \cap A^c)$ are all nonnegative, then $P^{ex}(B) = P^{ex}(A) + P^{ex}(B \cap A^c) \geqslant P^{ex}(A)$. If instead they are all nonpositive, then $P^{ex}(B) = P^{ex}(A) + P^{ex}(B \cap A^c) \leqslant P^{ex}(A)$. □

Proof of Proposition 57. Pick $A, B \in \mathcal{F}$; if they are disjoint, the equation follows immediately from (ii*). If they are not, consider $A \cup B = \{A \backslash B\} \sqcup \{A \cap B\} \sqcup \{B \backslash A\}$. Then, again by (ii*), we have that

$$P^{ex}(A \cup B) = P^{ex}(A \backslash B) + P^{ex}(A \cap B) + P^{ex}(B \backslash A) \tag{B.35}$$

Notice then that $P^{ex}(A \backslash B) = P^{ex}(A) - P^{ex}(A \cap B)$; indeed

$$P^{ex}(A) = P^{ex}(\{A \cap B\} \sqcup \{A \backslash B\}) = P^{ex}(A \cap B) + P^{ex}(A \backslash B).$$

Similarly, $P^{ex}(B \backslash A) = P^{ex}(B) - P^{ex}(B \cap A)$. So, (B.35) becomes

$$\begin{aligned} P^{ex}(A \cup B) &= P^{ex}(A) - P^{ex}(A \cap B) + P^{ex}(A \cap B) + P^{ex}(B) - P^{ex}(B \cap A) \\ &= P^{ex}(A) + P^{ex}(B) - P^{ex}(A \cap B). \end{aligned}$$

□

Proof of Theorem 60. Notice that the condition in Definition 59 is equivalent to $\sup_{\omega \in \Omega} \mathfrak{f}(\omega) \geqslant 0$, for all finite collections $\{B_j\}_{j=1}^n \subset \mathscr{B}$, $\{s_j\}_{j=1}^n \subset \mathbb{R}$. Notice also that P^{ex} restricted to $\mathscr{B}$ is a finite signed measure. By the Hahn-Jordan decomposition theorem (Fischer, 2012), there exists a unique decomposition of P^{ex} into a difference $P^{ex} = P^{ex}_+ - P^{ex}_-$ of two finite positive measures. Pick any $\{B_j\}_{j=1}^n \subset \mathscr{B}$,

$\{s_j\}_{j=1}^n \subset \mathbb{R}$. We have that, for any payoff

$$\begin{aligned}\mathfrak{f}(\omega) &= \sum_{j=1}^n s_j \left[\mathbb{1}_{B_j}(\omega) - P^{ex}(B_j)\right] \\ &= \sum_{j=1}^n s_j \left[\mathbb{1}_{B_j}(\omega) - P^{ex}_+(B_j) + P^{ex}_-(B_j)\right],\end{aligned}$$

the following holds

$$\begin{aligned}\int_\Omega \mathfrak{f} \mathrm{d}P^{ex} &= \sum_{j=1}^n s_j \left[\int_\Omega \mathbb{1}_{B_j} \mathrm{d}P^{ex} - P^{ex}_+(B_j) + P^{ex}_-(B_j)\right] \\ &= \sum_{j=1}^n s_j \left[\int_\Omega \mathbb{1}_{B_j} \mathrm{d}P^{ex}_+ - \int_\Omega \mathbb{1}_{B_j} \mathrm{d}P^{ex}_- - P^{ex}_+(B_j) + P^{ex}_-(B_j)\right] \\ &= \sum_{j=1}^n s_j \left[\int_\Omega \mathbb{1}_{B_j} \mathrm{d}P^{ex}_+ - P^{ex}_+(B_j)\right] - \sum_{j=1}^n s_j \left[\int_\Omega \mathbb{1}_{B_j} \mathrm{d}P^{ex}_- - P^{ex}_-(B_j)\right] = 0.\end{aligned}$$

So $\mathfrak{f}$ cannot have a negative supremum. □

Proof of Proposition 63. We have that

$$\begin{aligned}P^{ex}_t(A) &= P^{ex}_t(A \cap \Omega^+_t) + P^{ex}_t(A \cap \Omega^-_t) \\ &= \sum_{E^+_{t,j} \in \mathcal{E}^+_t} P^{ex}_t(A \cap E^+_{t,j}) + \sum_{E^-_{t,j} \in \mathcal{E}^-_t} P^{ex}_t(A \cap E^-_{t,j}) \\ &= \sum_{E^+_{t,j} \in \mathcal{E}^+_t : P^{ex}_t(E^+_{t,j}) \neq 0} P^{ex}_t(A \mid E^+_{t,j}) P^{ex}_t(E^+_{t,j}) \\ &+ \sum_{E^-_{t,j} \in \mathcal{E}^-_t : P^{ex}_t(E^-_{t,j}) \neq 0} P^{ex}_t(A \mid E^-_{t,j}) P^{ex}_t(E^-_{t,j}),\end{aligned} \tag{B.36}$$

where the first equality comes from $A = (A \cap \Omega^+_t) \sqcup (A \cap \Omega^-_t)$ and (ii*), the third equality comes from (5.2), and the second equality comes from $\Omega^+_t = \sqcup_{E^+_{t,j} \in \mathcal{E}^+_t} E^+_{t,j}$, so

$$A \cap \Omega^+_t = A \cap (\sqcup_{E^+_{t,j} \in \mathcal{E}^+_t} E^+_{t,j}) = \sqcup_{E^+_{t,j} \in \mathcal{E}^+_t} (A \cap E^+_{t,j}),$$

where the last equality comes from De Morgan laws. □

Proof of Proposition 64. The first two results come immediately from the facts that $\Omega_t^+, \Omega_t^- \subset \Omega$, for all t, and that $\mathcal{F}_t^+ = 2^{\Omega_t^+}$ and $\mathcal{F}_t^- = 2^{\Omega_t^-}$.

Now, suppose for the sake of contradiction that $\mathcal{F}_t^+ \cup \mathcal{F}_t^- \subsetneq \mathcal{F}$. This implies that $\Omega_t^+ \sqcup \Omega_t^- \subsetneq \Omega$, a contradiction.

For the opposite inclusion, suppose for the sake of contradiction that $\mathcal{F} \subsetneq \mathcal{F}_t^+ \cup \mathcal{F}_t^-$. Then, there exists $A \subset \Omega_t^+ \cup \Omega_t^-$ such that $A \notin \Omega$. But we know that, by hypothesis, $\Omega_t^+ \cup \Omega_t^- = \Omega$, for all t, so $A \subset \Omega_t^+ \cup \Omega_t^- = \Omega$, a contradiction. □

Proof of Proposition 65. Pick any $A, B \in \mathcal{F}$. Since $\Omega_t^+ \cap \Omega_t^- = \varnothing$, we have that

$$\begin{aligned} A \cup B &= ((A \cup B) \cap \Omega_t^+) \sqcup ((A \cup B) \cap \Omega_t^-) \\ &= ((A \cap \Omega_t^+) \cup (B \cap \Omega_t^+)) \sqcup ((A \cap \Omega_t^-) \cup (B \cap \Omega_t^-)) \\ &= (A_t^+ \cup B_t^+) \sqcup (A_t^- \cup B_t^-). \end{aligned}$$

A similar argument shows that $A \cap B = (A_t^+ \cap B_t^+) \cup (A_t^- \cap B_t^-)$, for all t.

For the second part, notice that $A \backslash B = \{\omega \in \Omega : \omega \in A, \omega \notin B\}$. Then, $A \backslash B \cap \Omega_t^+ \equiv A_t^+ \backslash B_t^+ = \{\omega \in \Omega_t^+ : \omega \in A, \omega \notin B\}$; similarly, $A \backslash B \cap \Omega_t^- \equiv A_t^- \backslash B_t^- = \{\omega \in \Omega_t^- : \omega \in A, \omega \notin B\}$. But $\Omega_t^+ \sqcup \Omega_t^- = \Omega$, so the claim follows. □

Proof of Proposition 66. Let us focus on $\mathcal{F}_t^+$; it is closed with respect to countable intersections because it is a sigma-algebra. Then, pick $A, B \in \mathcal{F}_t^+$ such that $A \neq B$. If $A \cap B = \varnothing$, then $A \backslash B = A \in \mathcal{F}_t^+$; if $A \cap B \neq \varnothing$, then $A \backslash B = A \cap B^c$. Then, $B \in \mathcal{F}_t^+$ implies $B^c \in \mathcal{F}_t^+$ because $\mathcal{F}_t^+$ is a sigma-algebra; also, $A \cap B^c \in \mathcal{F}_t^+$ because $\mathcal{F}_t^+$ is closed with respect to countable intersections. So $A \backslash B \in \mathcal{F}_t^+$. Finally, the unit element is Ω_t^+: for any $A \in \mathcal{F}_t^+$, $A \cap \Omega_t^+ = A$, because $A \in \mathcal{F}_t^+$. Hence $\mathcal{F}_t^+$ is a set algebra. We show in a similar fashion that $\mathcal{F}_t^-$ is a set algebra as well. □

Proof of Proposition 67. To ease notation, let $C \equiv \cup_{j \in \mathbb{N}_0} A_j$. C belongs to $\mathcal{F}$, because $\mathcal{F}$ is a sigma-algebra. Then, let $C \in \mathcal{F}_t^+$. This implies that $P_t^{ex}(C) \geqslant 0$, but also

that $P^{ex}(A) \geqslant 0$ and that $P^{ex}(C \cap A^c) \geqslant 0$, since $C = A \sqcup (C \cap A^c)$. Then,

$$P_t^{ex}(A) \leqslant P_t^{ex}(C) \leqslant \sum_{j \in \mathbb{N}_0} P_t^{ex}(A_j),$$

where the first inequality comes from Proposition 55, and the second one from Proposition 25.

If $C \in \mathcal{F}_t^-$, then $P_t^{ex}(C) \leqslant 0$, and also that $P^{ex}(A) \leqslant 0$ and that $P^{ex}(C \cap A^c) \leqslant 0$. Then,

$$P_t^{ex}(A) \geqslant P_t^{ex}(C) \geqslant \sum_{j \in \mathbb{N}_0} P_t^{ex}(A_j),$$

where again the first inequality comes from Proposition 55, and the second one from Proposition 57.

If $C \equiv \cup_{j \in \mathbb{N}_0} A_j$ is such that $C \cap \Omega_t^+ \neq \varnothing \neq C \cap \Omega_t^-$, then we cannot say anything general about the relation between $P_t^{ex}(A)$ and $\sum_{j \in \mathbb{N}_0} P_t^{ex}(A_j)$. □

Proof of Proposition 68. Notice that by (5.8) and(5.9), we have that, for all t,

$$P_t^{ex}(A \cap \Omega_t^+) = P(A \cap \Omega_t^+)$$

and

$$P_t^{ex}(A \cap \Omega_t^-) = -P(A \cap \Omega_t^-),$$

for all $A \in \mathcal{F}$, where $P \in \Delta(\Omega, \mathcal{F})$. Now, to save some notation, let us denote by $\Omega^+ \equiv \cup_{t \in \mathbb{N}_0} \Omega_t^+$ and by $\Omega^- \equiv \cap_{t \in \mathbb{N}_0} \Omega_t^-$. Then, we have the following.

$$d_{ETV}(P_t^{ex}, P_\infty^{ex}) = \sup_{A \in \mathcal{F}} |P_t^{ex}(A) - P_\infty^{ex}(A)| \equiv |P_t^{ex}(\mathbf{A}) - P_\infty^{ex}(\mathbf{A})|.$$

Let us denote by $\mathbf{A}^+ \equiv \mathbf{A} \cap \Omega^+$, and by $\mathbf{A}^- \equiv \mathbf{A} \cap \Omega^-$, so $\mathbf{A} = \mathbf{A}^+ \sqcup \mathbf{A}^-$. Notice that even though in the limit we fully discover $\mathbf{A}^+$, there may be some $t \in \mathbb{N}_0$ such

that $\mathbf{A}^+ \cap \Omega_t^+ \neq \varnothing \neq \mathbf{A}^+ \cap \Omega_t^-$. So we have

$$|P_t^{ex}(\mathbf{A}) - P_\infty^{ex}(\mathbf{A})| = |P_t^{ex}(\mathbf{A}^+) + P_t^{ex}(\mathbf{A}^-) - P_\infty^{ex}(\mathbf{A}^+) - P_\infty^{ex}(\mathbf{A}^-)| \tag{B.37}$$

$$= |P_t^{ex}(\mathbf{A}^+) - P_\infty^{ex}(\mathbf{A}^+)| \tag{B.38}$$

$$= |P(\mathbf{A}^+ \cap \Omega_t^+) - P(\mathbf{A}^+ \cap \Omega_t^-) - P(\mathbf{A}^+ \cap \Omega_t^+) - P(\mathbf{A}^+ \cap \Omega_t^-)| \tag{B.39}$$

$$= 2P(\mathbf{A}^+ \cap \Omega_t^-).$$

Equation (B.37) comes from $\mathbf{A} = \mathbf{A}^+ \sqcup \mathbf{A}^-$ and (ii*). Equation (B.38) comes from the fact that, for all t, $P_t^{ex}(\mathbf{A}^-) = P_\infty^{ex}(\mathbf{A}^-)$, because $\mathbf{A}^-$ is the portion of $\mathbf{A}$ that never leaves the latent space. Equation (B.39) comes from the updating procedure described in Section 5.3.2, from the countable additivity of P, and from $\mathbf{A}^+ = (\mathbf{A}^+ \cap \Omega_t^+) \sqcup (\mathbf{A}^+ \cap \Omega_t^-)$.

Now, notice that the limit as $t \to \infty$ of $\mathbf{A}^+ \cap \Omega_t^-$ is

$$\mathbf{A}^+ \cap \bigcap_{t \in \mathbb{N}_0} \Omega_t^- = \mathbf{A}^+ \cap \Omega^- = \varnothing,$$

where the last equality is by construction. Then, by the continuity of P we have that

$$\lim_{t\to\infty} d_{ETV}(P_t^{ex}, P_\infty^{ex}) = 2 \lim_{t\to\infty} P(\mathbf{A}^+ \cap \Omega_t^-) = 2P(\varnothing) = 0,$$

which concludes the proof. □

Proof of Proposition 69. Suppose that $P_\infty^{ex}(\Omega) \neq 1$. This of course implies that $P_\infty^{ex}(\Omega) < 1$ because, from the definition of extended probabilities, $P_\infty^{ex}(A) \in [-1, 1]$, for all $A \in \mathcal{F}$. Then, this means that there exists a set $A \in \mathcal{F}$ such that $P_\infty^{ex}(A) \leqslant 0$, which implies $A \subset \Omega_\infty^-$. But we know that, if $\Omega_t^+ \uparrow \Omega$, then $\Omega_t^- \downarrow \Omega_\infty^- = \varnothing$, which contradicts $A \subset \Omega_\infty^-$. □

Proof of Theorem 74. Pick any lower extended probability $\underline{P}^{ex}$ and suppose there exists $\varnothing \neq \mathcal{P}^{ex} \subset \Delta^{ex}(\Omega, \mathcal{F})$ such that $\underline{P}^{ex}(A) = \inf_{P^{ex} \in \mathcal{P}^{ex}} P^{ex}(A)$, for all $A \in \mathcal{F}$.

Now, by Theorem 60, we know that P^{ex} is coherent, for all $P^{ex} \in \mathcal{P}^{ex}$. This means that for all $\{B_j\}_{j=1}^n \subset \mathscr{B}'$, all $\{s_j\}_{j=1}^n \subset \mathbb{R}$, all $P^{ex} \in \mathcal{P}^{ex}$, and all $\omega \in \Omega$,

$$\sum_{j=1}^{n} s_j \left[\mathbb{1}_{B_j}(\omega) - P^{ex}(B_j)\right] \geqslant 0. \tag{B.40}$$

Notice that we use $\mathscr{B}'$ because that is the sigma-algebra generated by the events we can enter a bet about. Then, (B.40) holds if and only if

$$\sum_{j=1}^{n} s_j \mathbb{1}_{B_j}(\omega) \geqslant \sum_{j=1}^{n} s_j P^{ex}(B_j). \tag{B.41}$$

The inequality in (B.41) implies that

$$\begin{aligned} \sum_{j=1}^{n} s_j \mathbb{1}_{B_j}(\omega) &\geqslant \inf_{P^{ex} \in \mathcal{P}^{ex}} \sum_{j=1}^{n} s_j P^{ex}(B_j) \\ &\geqslant \sum_{j=1}^{n} s_j \inf_{P^{ex} \in \mathcal{P}^{ex}} P^{ex}(B_j) \\ &= \sum_{j=1}^{n} s_j \underline{P}^{ex}(B_j). \end{aligned} \tag{B.42}$$

The results in (B.42) hold if and only if $\sum_{j=1}^n s_j \left[\mathbb{1}_{B_j}(\omega) - \underline{P}^{ex}(B_j)\right] \geqslant 0$. The claim follows. □

Proof of Corollary 77. Pick a generic lower extended probability $\underline{P}^{ex}$, and suppose $\text{core}(\underline{P}^{ex}) \neq \varnothing$. Then, $\underline{P}^{ex}$ is coherent by Theorem 74. □

Proof of Proposition 78. We first show that $\text{core}(\underline{P}^{ex})$ is convex. Pick any $P_1^{ex}, P_2^{ex} \in \text{core}(\underline{P}^{ex})$, any $\alpha \in (0,1)$, and any $A \in \mathcal{F}$. We have

$$\alpha P_1^{ex}(A) + (1-\alpha) P_2^{ex}(A) \geqslant \alpha \underline{P}^{ex}(A) + (1-\alpha)\underline{P}^{ex}(A) = \underline{P}^{ex}(A),$$

so $\alpha P_1^{ex} + (1-\alpha) P_2^{ex} \in \text{core}(\underline{P}^{ex})$.

We then show that core($\underline{P}^{ex}$) is weak*-compact. Recall that, in the weak* topology, a net $(P^{ex}_\alpha)_{\alpha\in I}$ converges to P^{ex} if and only if $P^{ex}_\alpha(A) \to P^{ex}(A)$, for all $A \in \mathcal{F}$. This proof mimics the proof of (Marinacci and Montrucchio, 2004b, Proposition 3), where the authors prove the same claim for the core of a bounded game. Pick any $P^{ex} \in \text{core}(\underline{P}^{ex})$, and let $k := 2\sup_{A\in\mathcal{F}} |\underline{P}^{ex}(A)|$. For any $A \in \mathcal{F}$, it holds that $P^{ex}(A) \geqslant \underline{P}^{ex}(A) \geqslant -k$. On the other hand, for all $A \in \mathcal{F}$, we have that

$$P^{ex}(A) = P^{ex}(\Omega) - P^{ex}(A^c) \leqslant \underline{P}^{ex}(\Omega) - \underline{P}^{ex}(A^c) \leqslant 2\sup_{A\in\mathcal{F}} |\underline{P}^{ex}(A)|.$$

This implies that $|P^{ex}(A)| \leqslant k$, for all $A \in \mathcal{F}$. By (Dunford and Schwartz, 1958, Page 94), we have that

$$\|P^{ex}\| := \sup \sum_{j=1}^{n} |P^{ex}(A_j) - P^{ex}(A_{j-1})| \leqslant 2k, \tag{B.43}$$

where the supremum is taken over all finite chains $\varnothing = A_0 \subset A_1 \subset \cdots \subset A_n = \Omega$. Then, (B.43) implies that

$$\text{core}(\underline{P}^{ex}) \subset \{P^{ex} \in \Delta^{ex}(\Omega, \mathcal{F}) : \|P^{ex}\| \leqslant 2k\}.$$

By the Alaoglu Theorem (Dunford and Schwartz, 1958, Theorem 2, Page 424), we know that $\{P^{ex} \in \Delta^{ex}(\Omega, \mathcal{F}) : \|P^{ex}\| \leqslant 2k\}$ is weak*-compact. Hence, to complete the proof, we are left to show that core($\underline{P}^{ex}$) is weak*-closed. Let then $(P^{ex}_\alpha)_{\alpha\in I}$ be a net in core($\underline{P}^{ex}$) that weak*-converges to $P^{ex} \in \Delta^{ex}(\Omega, \mathcal{F})$. Using the properties of the weak* topology, it is easy to see that $P^{ex} \in \text{core}(\underline{P}^{ex})$. Hence, core($\underline{P}^{ex}$) is weak*-closed. The claim follows. □

Proof of Proposition 79. Equation (5.24) comes from the fact that, for all t,

$$P^{ex}_{t+1}(E^+_{t+1,\omega}) = |P^{ex}_t(E^-_{t,\omega})|,$$

that is, $P^{ex}_{t+1}(E^+_{t+1,\omega}) = -P^{ex}_t(E^-_{t,\omega})$, and that

$$-\overline{P}^{ex}_t(E^-_{t,\omega}) \leqslant -P^{ex}_t(E^-_{t,\omega}) \leqslant -\underline{P}^{ex}_t(E^-_{t,\omega}).$$

□

Proof of Proposition 80. To see that $\tilde{e} = J$ it is enough to notice that in Δ^1, the convex hull (which is a polytope) with the least amount of vertices is a line segment (also called dion), in Δ^2 it is a triangle, in Δ^3 it is a tetrahedron, and in Δ^4 it is a 5-cell. Then, an induction argument proves the claim. □

Appendix C

DeFinettian interpretation of subjective probability

De Finetti admits that it might have been better to adopt the seemingly more general approach of Ramsey and Savage of defining bets whose payoffs are in utils (de Finetti, 1974, Page 79):

The formulation [...] could be made watertight [...] by working in terms of the utility instead of with monetary value. This would undoubtedly be the best course from the theoretical point of view, because one could construct, in an integrated fashion, a theory of decision-making [...] whose meaning would be unexceptionable from an economic viewpoint, and which would establish simultaneously and in parallel the properties of probability and utility on which it depends.

Nevertheless, he found "other reasons for preferring" the money bet approach (de Finetti, 1974, Page 81):

The main motivation lies in being able to refer, in a natural way to combinations of bets, or any other economic transactions, understood in terms of monetary value (which is invariant). If we referred ourselves to the scale of utility, a transaction leading to a gain of amount S if the event E occurs would instead appear as a variety

of different transactions, depending on the outcome of other random transactions. These, in fact, cause variations in one's fortune, and therefore in the increment of utility resulting from the possible additional gain S: conversely, suppose that in order to avoid this one tried to consider bets, or economic transactions, expressed, let us say, in "utiles" (units of utility, definable as the increment between two fixed situations). In this case, it would be practically impossible to proceed with the transactions, because the real magnitudes in which they have to be expressed (monetary sums or quantities of goods, etc.) would have to be adjusted to the continuous and complex variations in a unit of measure that nobody would be able to observe.

Bibliography

Aldous, D. J. (2010), "More uses of exchangeability: representations of complex random structures," in *Probability and mathematical genetics*, vol. 378 of *London Math. Soc. Lecture Note Ser.*, pp. 35–63, Cambridge Univ. Press, Cambridge.

Allahverdyan, A. E. and Galstyan, A. (2014), "Opinion Dynamics with Confirmation Bias," *PLoS One*, 9.

Allen, E. H. (1976), "Negative Probabilities and the Uses of Signed Probability Theory," *Philosophy of Science*, 43, 53–70.

Amarante, M. and Maccheroni, F. (2006), "When an Event Makes a Difference," *Theory and Decision*, 60, 119–126.

Amarante, M., Maccheroni, F., Marinacci, M., and Montrucchio, L. (2006), "Cores of non-atomic market games," *International Journal of Game Theory*, 34, 399–424.

Amari, S.-i. (1983), "Differential geometry of statistical inference," in *Probability theory and mathematical statistics (Tbilisi, 1982)*, vol. 1021 of *Lecture Notes in Math.*, pp. 26–40, Springer, Berlin.

Antoniak, C. E. (1974), "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems," *Ann. Statist.*, 2, 1152–1174.

Arbabi, H. and Mezić, I. (2017), "Ergodic theory, Dynamic Mode Decomposition and Computation of Spectral Properties of the Koopman operator," *SIAM Journal on Applied Dynamical Systems*, 16, 2096–2126.

Bail, C. (2018), "Topic Modeling," *Available at cbail.github.io*.

Bárány, I. and Buchta, C. (1993), "Random polytopes in a convex polytope, independence of shape, and concentration of vertices," *Mathematische Annalen*, 297, 467–497.

Bartlett, M. S. (1945), "Negative Probability," *Mathematical Proceedings of the Cambridge Philosophical Society*, 41, 71–73.

Benavoli, A., Facchini, A., and Zaffalon, M. (2019), "Computational Complexity and the Nature of Quantum Mechanics (Extended version)," *Available at arXiv:1902.03513*.

Benavoli, A., Facchini, A., and Zaffalon, M. (2021), "The Weirdness Theorem and the Origin of Quantum Paradoxes," *Foundations of Physics*, 51.

Benferhat, S., Tabia, K., and Sedki, K. (2011), "Jeffrey's rule of conditioning in a possibilistic framework: an analysis of the existence and uniqueness of the solution," *Annals of Mathematics and Artificial Intelligence*, 61, 185–202.

Berger, J. O. (1984), "The robust Bayesian viewpoint," in *Robustness of Bayesian Analyses*, ed. J. B. Kadane, Amsterdam : North-Holland.

Berry, T., Giannakis, D., and Harlim, J. (2020), "Bridging data science and dynamical systems theory," *Notices of the American Mathematical Society*, 67, 1336–1348.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003), "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3, 993–1022.

Blume, L., Brandenburger, A., and Dekel, E. (1991), "Lexicographic Probabilities and Choice under Uncertainty," *Econometrica*, 59, 61–79.

Bock, J. D. and T'Joens, N. (2021), "Average Behaviour of Imprecise Markov Chains: A Single Pointwise Ergodic Theorem for Six Different Models," in *Proceedings of Machine Learning Research*, vol. 147, pp. 90–99.

Bollen, K. A. (2002), "Latent Variables in Psychology and the Social Sciences," *Annual Review of Psychology*, 53, 605–634.

Boltzmann, L. (1887), "Über die mechanischen Analogien des zweiten Hauptsatzes der Thermodynamik," *Journal für die reine und angewandte Mathematik*, 100, 201–212.

Burgin, M. (2012), "Integrating Random Properties and the Concept of Probability," *Integration: Mathematical Theory and Applications*, 3, 137–181.

Burgin, M. (2013), "Negative probability in the framework of combined probability," *Available at arXiv:1306.1166*.

Burgin, M. and Meissner, G. (2011), "Negative probabilities in modeling random financial processes," *Integration: Mathematical Theory and Applications*, 2, 305–322.

Caprio, M. and Mukherjee, S. (2021a), "Ergodic Theorems for Dynamic Imprecise Probability Kinematics," *Available at arXiv:2003.06502*.

Caprio, M. and Mukherjee, S. (2021b), "Ergodic Theorems in Dynamic Imprecise Probability Kinematics," *Available on ResearchGate*.

Caprio, M., Aveni, A., and Mukherjee, S. (2021a), "Concerning three classes of non-Diophantine arithmetics," *Involve, to appear*.

Caprio, M., Aveni, A., and Mukherjee, S. (2021b), "Concerning two classes of non-Diophantine arithmetics," *Proceedings of the 2021 summit of the International Society for the Study of Information, to appear*.

Cerreia-Vioglio, S., Maccheroni, F., Marinacci, M., and Montrucchio, L. (2012), "Signed integral representations of comonotonic additive functionals," *Journal of Mathematical Analysis and Applications*, 385, 895–912.

Cerreia-Vioglio, S., Maccheroni, F., and Marinacci, M. (2015), "Ergodic Theorems for Lower Probabilities," *Proceedings of the American Mathematical Society*, 144, 3381–3396.

Chen, J., Li, P., and Fu, Y. (2012), "Inference on the Order of a Normal Mixture," *J. Amer. Statist. Assoc.*, 107, 1096–1105.

Choquet, G. (1954), "Theory of capacities," *Annales de l'Institut Fourier*, 5, 131–295.

Coletti, G. and Scozzafava, R. (2002), *Probabilistic Logic in a Coherent Setting*, Trends in Logic, Dordrecht : Springer.

Cornfeld, I. P., Fomin, S. V., and Sinai, Y. G. (1982), *Ergodic Theory*, vol. 245 of *Grundlehren der mathematischen Wissenschaften*, New York: Springer-Verlag.

de Cooman, G., Hermans, F., and Quaeghebeur, E. (2009), "Imprecise Markov chains and their limit behaviour," *Probability in the Engineering and Informational Sciences*, 23, 597–635.

de Finetti, B. (1937), "La prévision : ses lois logiques, ses sources subjectives," *Annales de l'institut Henri Poincaré*, 7, 1–68.

de Finetti, B. (1974), *Theory of Probability*, vol. 1, New York : Wiley.

de Finetti, B. (1975), *Theory of Probability*, vol. 2, New York : Wiley.

Dempster, A. P. (1967), "Upper and lower probabilities induced by a multivalued mapping," *The Annals of Mathematical Statistics*, 38, 325–339.

Diaconis, P. and Freedman, D. (1980), "Finite exchangeable sequences," *Ann. Probab.*, 8, 745–764.

Diaconis, P. and Zabell, S. L. (1982), "Updating subjective probability," *Journal of the American Statistical Association*, 77, 822–830.

Dirac, P. A. M. (1930), "Note on exchange phenomena in the Thomas atom," *Mathematical Proceedings of the Cambridge Philosophical Society*, 26, 376–395.

Dudley, R. M. (2002), *Real Analysis and Probability*, vol. 74 of *Cambridge Studies in Advanced Mathematics*, Cambridge : Cambridge University Press, 2nd edn.

Dunford, N. and Schwartz, J. T. (1958), *Linear operators, part I: general theory*, London : Wiley Interscience.

Ellsberg, D. (1961), "Risk, Ambiguity, and the Savage Axioms," *The Quarterly Journal of Economics*, 75, 643–669.

Engelhardt, B. E. and Stephens, M. (2010), "Analysis of Population Structure: A Unifying Framework and Novel Methods Based on Sparse Factor Analysis," *PLOS Genetics*, 6, 1–12.

Everitt, B. S. and Hand, D. J. (1981), *Finite mixture distributions*, Chapman & Hall, London-New York, Monographs on Applied Probability and Statistics.

Feintzeig, B. H. and Fletcher, S. C. (2017), "On Noncontextual, Non-Kolmogorovian Hidden Variable Theories," *Foundations of Physics*, 47, 294—315.

Ferguson, T. S. (1974), "Prior distributions on spaces of probability measures," *Ann. Statist.*, 2, 615–629.

Ferrie, C. (2011), "Quasi-probability representations of quantum theory with applications to quantum information science," *Reports on Progress in Physics*, 74, 116001.

Fischer, T. (2012), "Existence, uniqueness, and minimality of the Jordan measure decomposition," *Available at arXiv:1206.5449*.

Fúquene, J., Steel, M., and Rossell, D. (2019), "On choosing mixture components via non-local priors," *Journal of the Royal Statistical Society: Series B*, 81, 809–837.

Gell-Mann, M. and Hartle, J. B. (2012), "Decoherent Histories Quantum Mechanics with One "Real" Fine-Grained History," *Physical Review A*, 85:062120.

Ghosal, S. and van der Vaart, A. (2017), *Fundamentals of Nonparametric Bayesian Inference*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge : Cambridge University Press.

Ghosh, J. K. and Ramamoorthi, R. V. (2003), *Bayesian nonparametrics*, Springer Series in Statistics, Springer-Verlag, New York.

Gong, R. (2018), "Modeling uncertainty with sets of probabilities." in *Foundations of Probability seminar series*, Rutgers University.

Gong, R. and Meng, X.-L. (2021), "Judicious judgment meets unsettling updating: dilation, sure loss, and Simpson's paradox," *Statistical Science*, 36, 169–190.

Gray, R. M. (2009), *Probability, random processes, and ergodic properties*, Dordrecht, Holland : Springer, 2nd edn.

Grün, B. and Hornik, K. (2011), "topicmodels: An R Package for Fitting Topic Models," *Journal of Statistical Software*, 40, 1–30.

Guha, A., Ho, N., and Nguyen, X. (2020), "On posterior contraction of parameters and interpretability in Bayesian mixture modeling," *Bernoulli*.

Harman, D. K. (1992), "Overview of the first text retrieval conference (TREC-1)," in *Proceedings of the First Text Retrieval Conference (TREC-1)*, pp. 1–20.

Hartle, J. B. (2004), "Linear Positivity and Virtual Probability," *Phisical Review A*, 70:022104.

Hartle, J. B. (2008), "Quantum Mechanics with Extended Probabilities," *Physical Review A*, 78:012108.

Heisenberg, W. K. (1931), "Über die inkohärente Streuung von Röntgenstrahlen," *Physikalische Zeitschrift*, 32, 737–740.

Henna, J. (2005), "Estimation of the number of components of finite mixtures of multivariate distributions," *Ann. Inst. Statist. Math.*, 57, 655–664.

Hoff, P. D. (2003), "Nonparametric estimation of convex models via mixtures," *Ann. Statist.*, 31, 174–200.

Hou, J. (2017), "Text Analysis with LDA on unknow topic structure data," *Available at Amazon AWS*.

Hu, Y. (2017), "The econometrics of unobservables: Applications of measurement error models in empirical industrial organization and labor economics," *Journal of Econometrics*, 200, 154–168.

Ichihashi, H. and Tanaka, H. (1989), "Jeffrey-like rules of conditioning for the Dempster-Shafer theory of evidence," *International Journal of Approximate Reasoning*, 3, 143–156.

Jarrow, R. A. and Turnbull, S. M. (1995), "Pricing derivatives on financial securities subject to credit risk," *Journal of Finance*, L, 53–85.

Jeffrey, R. C. (1957), *Contributions to the Theory of Inductive Probability*, PhD Thesis, Princeton University, Dept. of Philosophy.

Jeffrey, R. C. (1965), *The Logic of Decision*, Chicago : University of Chicago Press.

Jeffrey, R. C. (1968), "Probable Knowledge," in *The Problem of Inductive Logic*, ed. I. Lakatos, vol. 51 of *Studies in Logic and the Foundations of Mathematics*, pp. 166 – 190, Elsevier.

Kass, R. E. and Vos, P. W. (1997), *Geometrical foundations of asymptotic inference*, Wiley Series in Probability and Statistics: Probability and Statistics, John Wiley & Sons, Inc., New York, A Wiley-Interscience Publication.

Khrennikov, A. (2009), *Interpretations of probability*, Berlin/New York : Walter de Gruyter.

Kronz, F. (2007), "Non-monotonic probability theory and photon polarization," *Journal of Philosophical Logic*, 36, 449–472.

Lavine, M. (1992), "Some aspects of Pólya tree distributions for statistical modelling," *Ann. Statist.*, 20, 1222–1235.

Lavine, M. (1994), "More aspects of Pólya tree distributions for statistical modelling," *Ann. Statist.*, 22, 1161–1176.

Leonetti, P. and Caprio, M. (2021), "Turnpike in infinite dimension," *Canadian Mathematical Bulletin, to appear.*

Lewis, D. (1976), "Probabilities of Conditionals and Conditional Probabilities," *The Philosophical Review*, 85, 297–315.

Li, P. and Chen, J. (2010), "Testing the Order of a Finite Mixture," *J. Amer. Statist. Assoc.*, 105, 1084–1092.

Lindsay, B. G. (1995), *Mixture Models: Theory, Geometry and Applications (NSF-CBMS regional conference series in probability and statistics)*, Hayward, California : IMS.

Little, T. D. and Masyn, K. E. (2013), *Latent Class Analysis and Finite Mixture Modeling*, Oxford University Press.

Lowe, D. (2004/2007), "Machine Learning, Uncertain Information, and the Inevitability of Negative 'Probabilities'," *Available at Machine Learning Workshop 2004.*

Lv, H., Qiu, N., and Tang, Y. (2007), "Updating Probabilistic Knowledge Using Imprecise and Uncertain Evidence," in *Third International Conference on Natural Computation (ICNC 2007)*, vol. 4, pp. 624–628.

Ma, J., Lu, W., Dubois, D., and Prade, H. (2011), "Bridging Jeffrey's rule, AGM revision and Dempster conditioning in the theory of evidence," *International Journal on Artificial Intelligence Tools*, 20, 691–720.

Marchetti, S. and Antonucci, A. (2018), "Imaginary Kinematics," in *UAI 2018: Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence*, eds. A. Globerson and R. Silva, pp. 104–113, Monterey, California, USA, AUAI Press.

Marinacci, M. and Montrucchio, L. (2004a), *Introduction to the Mathematics of Ambiguity*, Uncertainty in Economic Theory, New York : Routledge.

Marinacci, M. and Montrucchio, L. (2004b), "Introduction to the mathematics of ambiguity," in *Uncertainty in economic theory: a collection of essays in honor of David Schmeidler's 65th birthday*, ed. I. Gilboa, London : Routledge.

Marriott, P. (2002), "On the local geometry of mixture models," *Biometrika*, 89, 77–93.

McLachlan, G. J. and Basford, K. E. (1988), *Mixture Models: Inference and Applications to Clustering*, vol. 38, New York : Marcel Dekker.

McNicholas, P. D. (2017), *Mixture model-based classification*, CRC Press, Boca Raton, FL.

Meng, X.-L. (2021), "Double Your Variance, Dirtify Your Bayes, Devour Your Pufferfish, and Draw Your Kidstogram," *Prepared for The New England Journal of Statistics in Data Science*.

Mengersen, K., Robert, C., and Titterington, D. M. (2011), *Mixtures: Estimation and Applications*, New York : Wiley.

Miller, J. W. and Harrison, M. T. (2018), "Mixture models with a prior on the number of components," *J. Amer. Statist. Assoc.*, 113, 340–356.

Nau, R. F. (2001), "De Finetti was Right: Probability Does Not Exist," *Theory and Decision*, 51, 89–124.

Ohn, I. and Lin, L. (2020), "Optimal Bayesian estimation of Gaussian mixtures with growing number of components," *Available at arXiv:2007.09284*.

Pearson, K. (1895), "Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material," *Philosophical Transactions of the Royal Society of London Series A*, 186, 343–414.

Pearson, K. and Erdmann III, H. O. M. F. (1894), "Contributions to the mathematical theory of evolution," *Philosophical Transactions of the Royal Society of London Series A*, 185, 71–110.

Phelps, R. R. (2001), *Lectures on Choquet's theorem*, vol. 1757 of *Lecture Notes in Mathematics*, Springer-Verlag, Berlin, second edn.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000), "Inference of Population Structure Using Multilocus Genotype Data," *Genetics*, 155, 945–959.

Rabe-Hesketh, S. and Skrondal, A. (2008), "Classical latent variable models for medical research," *Statistical Methods in Medical Research*, 17, 5–32.

Ramsey, F. P. (1964), "Truth and Probability," in *Studies in Subjective Probability*, eds. H. E. K. Jr. and H. E. Smokler, pp. 61–92, New York : Wiley.

Rao, K. P. S. B. and Rao, M. B. (1983), *Theory of charges, a study of finitely additive measures*, London : Academic Press.

Reitzner, M. (2005a), "Central limit theorems for random polytopes," *Probab. Theory Related Fields*, 133, 483–507.

Reitzner, M. (2005b), "The combinatorial structure of random polytopes," *Advances in Mathematics*, 191, 178–208.

Savage, L. J. (1954), *The Foundations of Statistics*, New York : John Wiley and Sons.

Shafer, G. (1981), "Jeffrey's Rule of Conditioning," *Philosophy of Science*, 48, 337–362.

Škulj, D. (2006), "Jeffrey's conditioning rule in neighbourhood models," *International Journal of Approximate Reasoning*, 42, 192–211.

Smets, P. (1993), "Jeffrey's rule of conditioning generalized to belief functions," in *Proceedings of the Ninth international conference on Uncertainty in Artificial Intelligence*, pp. 500–505.

Székely, G. J. (2005), "Half of a Coin: Negative Probabilities," *Wilmott Magazine*, pp. 66–68.

Tang, Y., Sun, S., and Li, Z. (2004), "Conditional evidence theory and its application in knowledge discovery," *Lecture Notes in Computer Sciences*, 3007, 500–505.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006), "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association*, 101, 1566–1581.

Tijms, H. C. and Staats, K. (2007), "Negative probabilities at work in the M/D/1 queue," *Probability in the Engineering and Informational Sciences*, 21, 67–76.

Titterington, D. M. (1990), "Some recent research in the analysis of mixture distributions," *Statistics*, 21, 619–641.

Tsitsiklis, J. (2018), "Sample spaces," in *Introduction to Probability (online course)*, Massachusetts Institute of Technology.

Vose, M. D. (1999), *The Simple Genetic Algorithm: Foundations and Theory*, Complex adaptive systems, Cambridge, Massachusetts : MIT Press.

Walley, P. (1991), *Statistical reasoning with imprecise probabilities*, vol. 42 of *Monographs on Statistics and Applied Probability*, London : Chapman and Hall.

Wasserman, L. A. and Kadane, J. B. (1990), "Bayes' Theorem for Choquet Capacities," *The Annals of Statistics*, 18, 1328–1339.

West, M., Müller, P., and Escobar, M. D. (1994), "Hierarchical priors and mixture models, with application in regression and density estimation," in *Aspects of uncertainty*, Wiley Ser. Probab. Math. Statist. Probab. Math. Statist., pp. 363–386, Wiley, Chichester.

Wiewiora, E. W. (2008), "Modeling probability distributions with predictive state representations," Ph.D. thesis, University of California, San Diego.

Wu, H. and Li, Z. (2022), "Ergodic Theorems for Capacity Preserving $\mathbb{Z}_+^d$-Actions," *Submitted to the International Journal of Approximate Reasoning*.

Zadeh, L. (1978), "Fuzzy sets as a basis for a theory of possibility," *Fuzzy sets and systems*, 1, 3–28.

Zhang, Y. (2018), "The Theory and Algorithm of Ergodic Inference," *Available on arXiv*.

Ziegler, G. M. (1995), *Lectures on Polytopes*, vol. 152 of *Graduate Texts in Mathematics*, Springer.

Biography

Michele Caprio obtained a Bachelor's degree in "Economic and Social Sciences" at Bocconi University in Milan, Italy. He continued his studies at Bocconi where he received a Master's degree in "Economic and Social Sciences". In Milan, he wrote a thesis on random sets supervised by Professor Pietro Muliere.

In August 2018, Michele started his graduate studies in the Department of Statistical Science at Duke University and has been advised by Sayan Mukherjee. During his time at Duke, he published a paper on turnpike theory (Leonetti and Caprio, 2021), and two on non-Diophantine arithmetics (Caprio et al., 2021a,b). He also studied finite mixture models and imprecise probabilities; the results he found constitute the chapters of his dissertation. He recently won the prestigious Aleane Webb Dissertation Research Fellowship and the IMS Hannan Graduate Student Travel Award to attend the 2022 IMS Annual Meeting. Next, he will be moving to the Department of Computer and Information Science of the University of Pennsylvania as a Postdoctoral Researcher. He will be working on the ARO-MURI project concerning robust concept learning and lifelong adaptation against adversarial attacks under the supervision of Professor Insup Lee.