# Coherent Approximation of Distributed Expert Assessments



by

Peter B. Jones

B.S., Electrical and Computer Engineering
Brigham Young University, 2002

S.M., Electrical Engineering and Computer Science
Massachusetts Institute of Technology, 2005

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2011



Author ..... .....
Department of Electrical Engineering and Computer Science
May 20, 2011

Certified by..... .....
Sanjoy K. Mitter
Professor of Electrical Engineering and Engineering Systems, MIT
Thesis Supervisor

Certified by... .....
Venkatesh Saligrama
Associate Professor of Electrical Engineering, Boston University
Thesis Supervisor

Accepted by..... .....
Leslie Kolodziejski
Graduate Officer

# Coherent Approximation of Distributed Expert Assessments

by

Peter B. Jones

Submitted to the Department of Electrical Engineering and Computer Science
on May 20, 2011, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

**Abstract**

Expert judgments of probability and expectation play an integral role in many systems. Financial markets, public policy, medical diagnostics and more rely on the ability of informed experts (both human and machine) to make educated assessments of the likelihood of various outcomes. Experts however are not immune to errors in judgment (due to bias, quantization effects, finite information or many other factors). One way to compensate for errors in individual judgments is to elicit estimates from multiple experts and then fuse the estimates together. If the experts act sufficiently independently to form their assessments, it is reasonable to assume that individual errors in judgment can be negated by pooling the experts' opinions.

Determining when experts' opinions are in error is not always a simple matter. However, one common way in which experts' opinions may be seen to be in error is through inconsistency with the known underlying structure of the space of events. Not only is structure useful in identifying expert error, it should also be taken into account when designing algorithms to approximate or fuse conflicting expert assessments. This thesis generalizes previously proposed constrained optimization methods for fusing expert assessments of uncertain events and quantities. The major development consists of a set of information geometric tools for reconciling assessments that are inconsistent with the assumed structure of the space of events.

This work was sponsored by the U.S. Air Force under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

Thesis Supervisor: Sanjoy K. Mitter
Title: Professor of Electrical Engineering and Engineering Systems, MIT

Thesis Supervisor: Venkatesh Saligrama
Title: Associate Professor of Electrical Engineering, Boston University

## Acknowledgments

It has given me enormous pleasure to have worked with and studied under Sanjoy Mitter for the past six years, first as a Master's and then as a Doctoral Student. His erudition is unparalleled, and his ability to leap from philosophy to mathematics to physics, sometimes within a single sentence, is one that I particularly admire. Beyond his intellectual mentorship, I particularly appreciate his willingness to engage in broader discussions of political and social significance. While our opinions haven't always coincided, I feel we've always been able to cohere on a view of the significant "moral dimension."

The role Venkatesh Saligrama has played as co-advisor of my Doctoral work is extremely appreciated. Beyond his own significant academic abilities, he brought to our frequent discussions an ability to see beyond theory to the real-world applications. He also often served as interpreter, being able to clarify complicated (and at times convoluted) arguments into more accessible ideas. I will always fondly recall sitting in Sanjoy's office as Venkatesh and Sanjoy seemingly effortlessly flew through an immensely diverse set of subjects, finding the interrelations as they went.

I'd also like to acknowledge the support and help of Devavrat Shah who served as the third member of my committee. His gentle critiques of my work served as an extremely useful corrective, providing an objective sounding board against which to develop and debate the ideas generated in my meetings with my advisors.

My time in the Laboratory for Information and Decision Systems has been absolutely wonderful. I deeply appreciate all the interactions with faculty, post-docs, students, staff and visiting scholars. I'm particularly cognizant of the great debt I owe to the friendship and help provided by Vincent Tan, Lav Varshney, Mukul Agarwal, Alex Olshevsky, Mesrob Ohannessian, Julius Kusuma, Hari Narayanan, Emily Fox, Jason Johnson, Kush Varshney, Matt Johnson, Venkat Chandrasekaran, Parikshit Shah, and Noah Stein. I'm especially grateful for feedback on early drafts of this thesis given by Vincent, Lav, and Julius.

Lincoln Laboratory has exhibited a huge amount of faith in me, providing me not only the financial ability to pursue my doctorate, but intellectual, social and emotional support as well. The Lincoln Scholars Committee, particularly Bill Keicher, Ken Estabrook and Jennifer Watson, not only made the choice to provide me with financial support, but did an admirable job of keeping me focused and on track. I particularly wish to acknowledge Michael Hurley who served not only as my Lincoln Mentor, but was often the first sounding board for concepts I was developing. The strong support of my Group Leaders, particularly Gary Condon who shepherded me through the Lincoln Scholars process, is very appreciated. The other members of Group 104 have given me frequent helpful advice and feedback on various developments, particularly Edward Kao, Andy Wang, Matt Daggett, Larry Bush, Ari Kobren, Rhonda Philips, and all the members of the CAMEL group.

My parents, Monti and Greg Jones, endowed me not only with the desire and drive to obtain the best education I could but with the abilities to achieve those goals.

Finally, this thesis owes more to my wife Cami than perhaps anyone else. She has been unwavering in her support for me, sacrificing years of her life for this goal. Over

the past seven years she sometimes had to function practically as a single parent when I disappeared for conferences, or spent mornings, afternoons and evenings working on a writing deadline. Her love and support mean the world to me.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

## 1.1 Motivation

A supercomputer called Multivac plays a central role in a series of short stories by Isaac Asimov. Multivac processes phenomenal amounts of data and outputs probabilities of various events of social significance. As a result of Multivac's accuracy of prediction, society is largely able to eradicate crime. In the story "All the Troubles of the World" Asimov presents a scenario in which a certain man is predicted with high probability to commit a murder. Police, acting to avert the crime, engage in an escalating series of actions, including detention, investigation, and questioning. However, each action simply increases Multivac's assessed probability of the crime occurring. The crime is eventually averted (in the nick of time), by the realization that the system's representational systems were incapable of discriminating between the father and his minor son. This fictional story points out real difficulties of large scale probabilistic inference involving autonomy, bias, and what probability really means.

A similar theme is evoked, more sinisterly, in Philip K. Dick's "The Minority Report," in which three mutant humans act as oracles of future crime. Termed 'precogs,' the mutants' predictions of future events help curtail crime. However the predictions are not absolute, and occasionally one of the three precogs has a different vision from the others, termed a 'minority report.' The police division charged with acting on the precogs' visions chooses to ignore the minority reports, resulting in troubling outcomes where truth is sacrificed in the interest of efficiency.

These themes of eradicating evil through super-human foresight are certainly entertaining (as evidenced by the success of recent Hollywood movies based on each of the short stories), but reality is not science fiction. However, in an instance of fiction prefiguring reality, 'predictive policing' was the theme of a 2009 National Institute of Justice symposium. As data on crime proliferates and as high-dimensional inference improves, science fiction is moving closer to reality. With this increased predictive capability comes the same challenges for those decision makers tasked with acting on expert advice in risky situations that were illuminated by the fictional stories: what action to take when experts (human or artificial) express uncertainty, when the advice of one expert contradicts or is inconsistent with that of another, and how to deal with fundamental representational limits of expert systems.

When we're called to act under uncertainty or risk, a reasonable way of improving

the outcome of our actions is to gather recommendations and evaluations from groups of experts (sometimes called Subject Matter Experts, or SMEs, in the literature). For example:

- A military commander, prior to engaging in a strategic action, will request input from his senior officers
- An individual, diagnosed with a deadly illness, will consult multiple doctors
- A company, wishing to protect its communications network, may employ multiple commercial Intrusion Detection and Prevention Systems (IDPSs)
- A police detective, investigating a series of thefts, will consult other detectives, data mining algorithms working on a variety of databases, and profiling specialists to identify the culprit

As a general model, each expert provides an *assessment* of some *event*. Assessments may be qualitative, as in the case of the commander's senior officers, or quantitative, as in the case of the IDPSs. Events may be future (forecasting) or past (forensics). Experts may be human (doctors, officers, etc.) or machine (IDPSs, data mining algorithms, etc.). They may assess a single event or many. Each may be asked to assess the same set of events, or each may be asked to assess a different, but related, set of events.

The central question we address in this thesis is how to reconcile uncertain, inconsistent expert assessments. For now, the term 'inconsistent' can be taken loosely to mean logically irreconcilable, although more technical definitions will follow. When a decision maker receives conflicting assessments he must employ some method to reconcile them. He may:

- Reject some assessments out of hand, based on exogenous (reputation, complexity, etc.) or endogenous (internal consistency, neighbor nearness) criteria
- Modify some or all assessments until they become consistent with prior beliefs and then update his beliefs
- Update his beliefs with the inconsistent assessments, and then perform revision to create consistency

Recognizing that the right method of reconciliation may be situation dependent, we will develop some general principles for reconciling inconsistent assessments in a variety of situations.

## 1.2 Distributed Detection and Estimation

From an engineering perspective, the problem of reconciling diverse assessments is related to the problem of distributed detection and estimation. The fusion challenge is to create a rule that incorporates the information from the distributed assessments while obeying fundamental constraints that may exist in the world.

The sister problems of distributed and decentralized detection have received much attention in the academic engineering literature, particularly in the context of the design of sensor networks. An excellent overview of the subject is given in [1]. Early work by Tenney [2,3], Tsistiklis [4–6], Varshney [7] and others has grown into a robust community.

One particularly fruitful area of research has been on belief dynamics in networks. See particularly [8–10]. The challenges of coming to a consensus on a state of the world under limited and time-varying communication has been applied to cooperative vehicle dynamics [11,12], distributed function computation [13], leader election [14,15], clock synchronization [16,17] and more. The consensus literature focuses primarily on engineered solutions to the problem of convergence on a single value across the network.

In contrast to previous literature, in this thesis we will focus on the estimation of a set of *consistent* values rather than a single *consensus* value. Also, the engineering focus on limitations imposed by constrained communications in networks in relation is left largely to future work.

### 1.2.1 Distributed Filtering and Consensus

A natural outgrowth of the decentralized detection literature has been to consider estimating the state of a time-varying process. Optimal filtering theory [18], with early developments by Wiener [19], Kalman [20] and Kalman and Bucy [21] developed statistical methods for estimating process states given a sequence of uncertain observations.

Attempts to generalize optimal filtering to the decentralized (indicating geographic diversity but no communication network constraints) and distributed (both geographically diverse and communication constrained) settings [22–25] has produced generalized algorithms for estimating the time-varying state of a centrally observable process. Assumed, however, in all these treatments is a single, globally agreed-upon dynamical model. The problem of how to do optimal distributed filtering when observational experts dissent on the proper model specifications has not been addressed in the literature.

### 1.2.2 Of Indian Villagers and Elephants

A classic example used in much of the distributed detection and estimation literature, is of observationally limited villagers (either blind or in the dark, depending on the telling) who must ascertain the nature and characteristics of an elephant. I take the following account from a classic text by the Sufi poet Rumi:

> Some Hindoos were exhibiting an elephant in a dark room, and many people collected to see it. But as the place was too dark to permit them to see the elephant, they all felt it with their hands, to gain an idea of what it was like. One felt its trunk, and declared that the beast resembled a water-pipe; another felt its ear, and said it must be a large fan; another its leg, and thought it must be a pillar; another felt its back, and declared

the beast must be like a great throne. According to the part which each felt, he gave a different description of the animal.
(*The Masnavi*, by Rumi, tr. by E.H. Whinfield, [1898])

Many lessons can (and have) been derived from this simple tale, but in the context of distributed detection and estimation the story is used to illustrate the difficulty of using multiple observations to construct a complete picture of the observed object. Key to our conception of distributed assessment is this picture of a single phenomenon that is being perceived in a limited way by a distributed set of assessors, each attempting to explicate his local observation as faithfully as possible.

### 1.2.3 Consistency vs. Consensus

Much of the previous engineering literature in distributed detection and estimation has focused on the consensus question. The two questions most often asked are: given a distributed set of assessors, each observing the same phenomenon, **when** and **how** can their beliefs about the phenomenon converge to a single, global belief.

Such a model is excellent when the phenomenon under observation is sufficiently confined that experts can extend their local knowledge to an assessment of the phenomenon as a whole, or when locality conditions on the phenomenon itself result in separability guarantees among the groups of assessors. However, in an entangled world with phenomena that are 'elephant' sized, a broader criteria needs to be adopted.

We suggest that the appropriate question in a global environment with "large" phenomena is consistency. To strain the elephant analogy, suppose we have two experts trying to classify an unknown animal. One expert, using keen scientific intellect, takes the animal's temperature under various external conditions. The temperature remains constant to within a threshold, and the expert proclaims, "this animal is warm-blooded." The second expert, observing mating and reproduction, states that "this animal's young are hatched from eggs." Both of these statements represent 'soft' decisions about the identity of the animal (i.e. they don't uniquely identify the animal, but narrow the set somewhat). The assessments are consistent with one another in the sense that there exist animals that are both warm-blooded and whose young hatch from eggs. In essence, if there exists at least one animal consistent with the decisions of all the assessors, then the assessments are consistent.

The goal of distributed systems for detection, estimation and assessment should be to make consistent rather than consensus decisions.

## 1.3 Coherence

The fundamental concept to which this thesis will speak is **coherence**. With respect to assessments, coherence is the requirement that assessments be consistent with one another, that they cohere. In terms of the Indian story of the elephant, the various statements of the villagers were coherent because an object exists (an elephant) with all the assessed characteristics. Had no such object existed, the assessments would have been incoherent.

Coherence can be viewed as a relaxation of the concept of consensus. If the object or event under assessment is identical across all assessors, then to be internally coherent, all the assessments must be the same. However, if the assessors are working locally, as opposed to globally, their assessments may differ (as per the elephant example). But this does not represent a *de facto* falsification of the assessments; rather, such falsification would depend on the (non)existence of an object or event that exhibited all assessed local behaviors.

Coherence-like concepts, as will be shown, are prevalent in many academic disciplines where they are often contrasted with foundationalist approaches. Concepts of coherence are often (but not always) associated with ideas of subjectivism, nonrationality, and systemic thinking while fundamentalist views are focused on objectivisim, rationality, and axiomatization. The tension between these two viewpoints, objective/foundational versus subjective/coherent, can be seen in many different academic fields.

In the following cross-disciplinary development we do not attempt to give a complete view of any one topic. We recognize that many of the covered fields have rich and deep findings not included in these brief summaries, and that even through the limited lens of coherence, none of these caricatures constitute a complete picture. Together, however, they indicate the pervasiveness of the dichotomy between foundationalism and coherentism.

### 1.3.1 Probabilistic Coherence

In his seminal work on probability theory [26], de Finetti defined a concept of probabilistic coherence. Probabilistic coherence will be the fundamental concept in our theoretical development, and we leave a lengthy discussion of its principles to Chapter 2. Here we will refer briefly to the theory, its motivations and implications.

Early approaches to probability theory were based on a frequentist notion that a probability encodes an average long-term outcome of some repeatable trial. In this approach to probability, the values reflect a fundamental truth about objective reality. However, this theory of probability is insufficient to explicate how probability may be used to represent non-repeatable events. This shortcoming of the frequentist/objectivist view of probability led to a subjectivist approach, pioneered by Ramsey [27], de Finetti [26], and Savage [28]. Generally speaking, the subjectivist view of probability was that a probability is an encoding of personal uncertainty about the outcome of an event. The coherence principle, posited by de Finetti but drawing on earlier work by Ramsey, states that probabilities when viewed as odds must not allow a risk-free gamble to be made against them. This so-called "Dutch book" argument views probabilities not as representative of long run averages, but as a system of consistent beliefs.

### 1.3.2 Coherence in Philosophy

A fundamental objective of philosophy is to explicate the concept of 'truth,' including what is meant when we claim a proposition is true, and how we justify such belief.

One class of theories of truth is referred to as "Coherence Theories" [29]. Coherence theories have been espoused by such philosophers as Leibniz, Spinoza, Hegel, Bradley, Blanshard, Neurath, Hempel, Dummett, and Putnam. Generally, these theories "account for the truth of a proposition as arising out of a relationship between that proposition and other propositions." The core principle of coherentism is that truth arise not from agreement with 'objective reality' (which a coherentist may dispute is a meaningful concept), but through internal cohesion within a belief set.

#### Epistemology

Due to some degree to prominent criticisms of the coherentist view of truth (particularly by Russell [30]), modern coherentists have shifted from coherence as an attempt to *define* truth to coherence as a method of *justifying* epistemic statements. In this strain of philosophical coherentism, an individual's belief system is justifiable only if it is internally consistent. A stronger version of justificational coherentism states that coherence is both necessary **and sufficient** to justify belief.

#### Analytic Philosophy

One of the most important innovators in analytic philosophy in the 20th century was Ludwig Wittgenstein. In his *Tractatus Logico-Philosophicus* [31], he states "Wovon man nicht sprechen kann, darüber muss man schweigen." Translation: whereof man can not speak, he must keep silent. Wittgenstein argued that man's necessarily finite descriptive ability was insufficient to describe the infinite variety of experience of reality. As such, the set of truths that can be communicated (and hence verified by examination) is a small subset of the total set of truths.

The concept of verifiability was taken up by Wittgenstein's sometime friend and collaborator, and a founding member of the influential Vienna Circle, Waismann. In his 1946 paper "Verifiability," [32] Waismann espouses a very Wittegenstein-ian view of the limits of verifiable truth. He coins the term "open texture" to refer to the diversity of ways in which the same language can be justifiably applied. In [33], Hilary Putnam takes up the discussion of open texture as a challenge to the computational reductionist view of language and truth. The strict rules of logic and computation are insufficient to span the breadth of infinite experience and meaning

#### Philosophy of Science

Perhaps the two most currently influential philosophies of scientific progress are those of Karl Popper [34, 35] and Thomas Kuhn [36]. Popper's view, first posited as a counter to the logical positivism of the Vienna Circle (of which Friedrich Waismann was a member), is that scientific theories should be treated critically. In *Logik der Forschung* [34] he proposed *falsifiability* as the fundamental invariant of scientific knowledge. This was in contrast to the positivists, who believed in constructive empiricism rather than destructive empiricism. However, like the positivists, Popper's theory of falsifiability is a fundamentalist view of science, that scientific knowledge exists only through tests and experiments.

In contrast, Kuhn's proposition in *The Structure of Scientific Revolutions* of 'revolutionary science' in which old modes of thinking are rapidly replaced by new paradigms is premised on the idea that scientific progress is not solely dependent on rational comparison, but can be strongly influenced by gestalt shift. Underlying this premise is the idea that there is something akin to *taste* in the scientific endeavor, which is only possible if scientific knowledge exhibits a degree of openness, or porosity.

#### Philosophy of Law

The concept of porosity, related to the 'open texture' of Waismann, was fundamental in HLA Hart's formulation of legal positivism [37]. The theory of legal positivism views legal structures as reflective not of fundamental truths, but as an outgrowth of social development. Although not explicit in Hart's formulation, a coherentist view of law would suggest that a law base is justified not by its reflection of reality, nor by its consequences, but by its internal consistency.

### ■ 1.3.3 Coherence in Political Science

In [38] evidence is given for an objectivist/subjectivist divide within political theory. Specifically, the author proposes what he refers to as the 'economic' and 'sociological' schools of thought within political science. By 'economic' he means rational, foundational and axiomatic in nature and by 'sociological' he means irrational (or at least arational), systems-centric and coherent in nature. The author suggests the modern sociological approach has its roots in Hegel, Coleridge, and other European theorists responding to the excesses of the French Revolution, and he points to Talcott Parsons as its primary modern purveyor.

### ■ 1.3.4 Coherence in Cognitive Linguistics

Perhaps the most influential figure in 20th century linguistics has been Noam Chomsky. Among his many contributions is the formulation of the concept of Universal Grammar [39] as the common structural basis for all spoken language, introduced in the 1950s and early 1960s. At the time, this theory displaced the then dominant behaviourist view of language as a learned (rather than innate) set of rules of communication. The heart of Chomsky's development of the universal grammar is the concept of Deep Structure, an overarching conceptual framework from which multiple grammatical constructs can be derived. In this sense it is an absolutist view of linguistics, relying on a common structure, or first principle for its motive effect.

In response to dissatisfaction with Chomsky's theory, several influential linguists (including some of Chomsky's former students) developed a theory of generative semantics, which posited that any universal grammar was insufficiently complex to account for the variety of linguistic constructs. This led, in turn, to the development of the concept of cognitive 'frames' [40,41] which are, in the words of Charles Fillmore, that which "identifies the experience as a type and gives structure and coherence - in short, meaning - to the points and relationships, the objects and events, within the experience." Cognitive frames are collections of concepts which form meaning through

interrelation rather than deriving meaning from atomic rules. A key cognitive developmental step, according to the theory, is the ability to switch frames, which creates a new structure for understanding and interpreting experience.

### 1.3.5 Coherence in Economics and Finance

Perhaps nowhere is the concept of coherence and distributed consistency more important that in economics. The law of one price [42] and the closely related Efficient Market Hypothesis [43] and Arbitrage Pricing Theory [44] are essentially consistency principles on the assessment of financial worth of assets, including commodities, stocks, options, etc. Furthermore, the concepts of distributed equilibrium in games [45–47] can be seen as strategic coherence.

Despite its centrality (or, likely, because of it), the ability of these coherence principles to describe real economic behavior has frequently been brought into question. The seminal work of Kahneman and Tversky [48,49] introduced the field of Prospect Theory which demonstrates empirically that coherent pricing frequently fails to hold, and suggests psychological and behavioral models as alternatives to expected utility models. Their work started a cascade of results and the establishment of the field of behavioral economics [50] which explores the various ways in which human economic assessments violate the principle of coherence.

The psychological and behavioral realities such as loss aversion, bounded rationality, cognitive biases, herd behavior, and the impact of distributed information, are exactly why the coherent approximation principles developed in this thesis are necessary. If assessments were coherently generated then no approximation would be necessary. But given the wealth of empirical evidence to the contrary, for expert assessments to be maximally useful to decision makers, systematic techniques for approximating 'true' values given incoherent assessments must be developed.

## 1.4 Thesis Outline and Major Contributions

This thesis will develop methods for combining or approximating expert assessments coherently. The major contributions include:

- The formulation and justification of an Information Geometric Coherent Approximation Principle (IGCAP) and a comparison with other methods of coherent approximation (Chapter 3)

- The development of mechanisms for coherently approximating sequences of assessments generated by mismatched likelihood models (Chapter 4)

- An application of IGCAP to perform approximate Bayesian filtering based on an incoherent sequence of expert assessments (Chapter 4)

- A method for coherently fusing the outputs of a distributed risk assessment process based on the IGCAP (Chapter 5)

- An alternative interpretation of incoherence as a structural limitation and two suggested methods for relaxing the structure accordingly (Chapter 6)

First, in Chapter 2, we summarize the subjective view of probability theory, from which the concept of probabilistic coherence naturally emerges. We give more firm mathematical definitions to the concepts of 'probability', 'assessment', and 'coherence' and provide both a historical perspective on some of the issues and expand on the philosophy justifying the focus on coherence as a fundamental principle of assessment.

Then, in Chapter 3 we state, critique, and generalize the Coherent Approximation Principle (CAP), a method previously suggested in the literature for coherently approximating incoherent assessments. We investigate a single-shot problem, in which a set of experts generate assessments of some set of events, and analyze the computability and cost-behavior of the CAP for the problem. We introduce an alternative formulation, termed the Information Geometric CAP (IGCAP), justify it via a particular assessor model, and analyze its behavior in various regimes. Finally, we consider the impact on the set of coherent assessments of additional structural constraints.

Next, in Chapter 4 we switch from a static assessment problem to a dynamic one. In this case, the problem is to define a method for approximating sequences of potentially incoherent assessments with coherent ones. We analyze three distinct sequential problems: the Subjective Likelihood (SL) model, Conditional Assessments (CA) model, and Markov Chain (MC) model. In SL, two definitions of coherent likelihood functions are proposed and one is shown to be strictly weaker than the other. Then, based on the asymptotic coherence definition, an approximation method is introduced which preserves the predictive uncertainty of the incoherent assessment. We verify the efficacy of the approximation method in simulation. In CA we analyze the problem of coherently approximating mixtures of conditional and unconditional assessments, and demonstrate how the information geometric formulation of IGCAP allows conditional and unconditional assessments to be fused. We also demonstrate an equivalence between applying IGCAP to conditional assessments and applying conditioning to IGCAP approximations of unconditional assessments. Finally in MC we develop an approximate method of Bayesian filtering when observations and likelihood models are observable only through the distributed assessments of a set of incoherent assessors.

We then turn our attention to the role of subjectivity and coherence in financial markets. In Chapter 5 we review the principles of coherent and convex risk measures and analyze the minimal convex extension of the popular Value-at-Risk (VaR) risk measure. We then analyze the robustness of several risk measures (coherent, convex, and non-convex) to small variations in the probability distribution over outcomes. Finally we employ the IGCAP developed in earlier chapters to the problem of approximating coherently a distributed risk assessment, and then analyze a similar problem of fusing risk assessments across a global financial enterprise.

In the penultimate chapter, Chapter 6, we revise the assumption that an incoherent assessment need be coherently approximated. Instead we suggest two mechanisms by which the assessment is not in error, but the assumed structure is insufficient to describe the phenomenon under assessment. For each of these mechanisms we develop

an associated method for minimally relaxing the structural constraints such that the assessment becomes coherent under the relaxed structure.

Finally, in Chapter 7, we will summarize the main developments of the thesis, particularly the new contributions, provide thoughts on future applications and extensions, and attempt to place the work in a broader academic context.

# Chapter 2

# Background

## 2.1 Introduction

In this chapter we set the stage for the technical developments to come. The fundamental task we set for ourselves is one of justification.

## 2.2 What is Probability

Probability, as a concept, is both accessible to children in their earliest stages of mathematical development and simultaneously a source of perpetual disagreement among the extremely wise and learned.

Fundamentally, probability is a way of expressing uncertainty about the outcome of some event. While there is a notion of 'qualitative' probability, in general we restrict probability to mean a quantitative expression of uncertainty. By convention, a probability is a number between zero and one that expresses something about the nature of a not-yet-determined event.

Within this framework, there is ample room for disagreement about the precise nature of probability. Indeed, how to interpret probability has long been a contentious issue and several schools of thought have arisen over the years. We give here a summary account here of four major historical schools of interpretation of the term 'probability.' More complete accounts can be found in [28, 51, 52].

### 2.2.1 Classical

The classical view of probability arose in the work of Fermat, the Bernoullis (Jakob and Daniel), Pascal and de Moivre. David [53] terms this view "the mathematical theory of arrangements" and explains the classical view of probability as follows:

> The probability of an event happening is, in a general way, then, the ratio of the number of ways in which the event *may* happen, divided by the total number of ways in which the event may or may not happen.

Thus, when asking the probability of a spin of the roulette wheel or the roll of a die or the flip of a coin, the probability is determined by the cardinality of the set of possible outcomes and the cardinality of the set of outcomes where the event of interest occurs, or obtains. The fundamental invariant in the classical view is one

of symmetry: the set of possible outcomes is defined by repeated division until the remaining atoms are symmetric.

Significant objections, both practical and philosophical, can be raised to the classical view. One practical objection has to do with the connection between reality and theory. For probability to have an operational meaning, it must be translated into something with physical meaning, but the classical basis of probability in symmetric outcomes, while based in the physical nature, doesn't have an immediate physical implication. Another objection, more philosophical, is how to interpret symmetry consistently. In an example given in [54], when uniformly distributed dice are replaced with bones, as in the ancient Mediterranean culture, how to appropriately understand the set of atomic, or symmetric, outcomes is unclear.

### ■ 2.2.2 Frequentist

The frequentist (sometimes objectivist, mathematical, numerical or statistical) school of thought supplanted the classical view in the early 1900s. Influential proponents of the frequentist view include Richard von Mises [55, 56], R.A. Fisher [57], Jerzey Neyman and Egon Pearson [58]. The fundamental invariant of the frequentist school is the repeated trial of some experiment. In the frequentist school of thought, probability is the limit as the number of repeated trials goes to infinity of the frequency of occurrence of the event of interest.

An obvious limitation of the frequentist view of probability is that it only allows probabilities to be ascribed to events which are subjectable to repeated trial. This limitation is at odds with natural language, where it is perfectly natural to assert that it's somewhat probable the Red Sox will win the World Series this year. The 2011 Major League Baseball season will only be played once and is not amenable to repeated trial. However, the ascription of probability to such an uncertain event seems quite natural to most people.

A further technical limitation of the frequentist view is in the definition of a repeated trial. What is meant, exactly, when one refers to a repeated trial? Certainly if a trial is repeated exactly then the outcome will be identical (ignoring quantum effects). If a coin is tossed from exactly the same point with exactly the same rotational force and caught in exactly the same position, under exactly the same environmental conditions, physics would suggest the outcome will be precisely the same. It seems obvious that by 'repeated trial' what is meant is a sequence of trials which are deemed 'close enough'. But frequentist probability theory gives no guide for how to differentiate between experiments which are sufficiently similar to be considered repetitions and those that aren't.

### ■ 2.2.3 Necessary

The necessary, or logical, view of probability is perhaps the least understood, and certainly the least developed. Originally proposed by Keynes [59] and further developed by Carnap [60], the necessary view of probability holds that probability represents (in Carnap's phrase) a "degree of confirmation" of a truth value.

The use of the identifier "logical" (used by Keynes in his Treatise on Probability) suggests a strong connection between this theory and the work in analytic philosophy of Bertrand Russell and Alfred Whitehead. Indeed, one author refers to Keynes' development of probability as "a lineal descendant of Russell and Whitehead's *Principia Mathematica*" [61]. However the project was never fully realized, and Keynes recognized himself that the mechanisms derived in the theory would be difficult to bring into practice. Carnap's attempt at further development met with similar challenges, and the logical view of probability remains largely of philosophical rather than practical importance.

### ■ 2.2.4 Subjective

The subjectivist school of probability was set forth partially as an answer to the problem of ascribing probability to a non-repeatable event. Subjectivism interprets probabilities not as long-run frequencies of events, but as statements of internal uncertainty of an event's outcome by individuals. Early proponents of the subjectivist view of probability included Frank Ramsey [27] and Bruno de Finetti [26]. The beginning of a widespread acceptance of subjectivist probability theory as a viable alternative to the frequentist approach can probably be traced to Savage's seminal development [28,62] and the influential work of Richard Jeffreys [63].

Subjectivism is not without its critics, or its valid critiques. One criticism is that there is no framework within the subjectivist model to refute probabilities that are at odds with observed frequencies of repeated events. For example, given a six-sided die one knows (via physical inspection or repeated observation) to be uniformly balanced, there is no subjectivist reason to reject an assessment that places unequal probability on the six possible outcomes of a randomizing throw. As a result of this perceived short-coming, subsequent authors were at some pains to attempt to unite subjectivist and objectivist views of probability [64–66]. The kernel of thought that runs through such attempts seems to be that there exist "true" probabilities of events, and probability assessments ought to reflect such probabilities in a meaningful way.

One result of this combined view of subjective and objective probability is the perception of subjective probabilities as "noisy" measurements of the objective probabilities of events. This suggests that in order to determine the objective probabilities, one could collect a set of subjective probabilities across a population and appropriately fuse them in order to generate an improved estimate of some true world state.

Another possible way to reconcile objectivist and subjectivist views of probability theory is to view them as, respectively, time-average and ensemble-average interpretations of events. In the frequentist view, an event space is visited repeatedly and a time average is taken of when an event achieves. In the subjectivist view, when the event space cannot be revisited, subjects mentally construct an ensemble of the event and report (hopefully accurately) the ensemble average outcome of the experiment.

Despite the criticisms of the subjectivist view, it seems to offer the best framework for interpreting probability as it applies to natural usage. The conceptualization which underlies de Finetti and Ramsey's original development of probabilities as points of indifference to wagers on uncertain events will be employed throughout this thesis

when it is necessary to fundamentally define probability.

### 2.2.5 Other Representations of Uncertainty

Probability is not the only proposed mathematical representation of uncertainty. Several other methods have been proposed, including upper and lower probabilities, Dempster-Shafer theory, possibility theory, fuzzy sets and fuzzy logic, and many more. Each of these methods has been introduced in order to address some perceived shortcoming of the probability framework. While the identified shortcomings of probability are indeed valid, the proposed solutions often suffer from their own shortcomings. Furthermore, the increased complexity of most of these theories has limited their practical applicability. They certainly may play a role in certain applications, but they won't be the primary focus of this thesis. For further discussion and references, see [67, 68].

## 2.3 What is an Assessment

Now that we have settled on an understanding of what is meant by 'probability,' we next move to a definition of the concept of 'assessment.' Informally, an assessment is some value that is representative of all the possible outcomes of an uncertain quantity. This may also be termed a valuation or, in a limited sense, an expectation or (in de Finetti's terms) a prevision, although the concept of assessment is more general than that of prevision, as will be explained in Section 2.4.

### 2.3.1 Mathematical Model

To formalize the concept of an assessment, we need to introduce some mathematical concepts and notation. In this section we introduce mathematical notation that will be used throughout the thesis.

#### Outcomes

We denote by $\Omega$ the set of all possible *outcomes*, where each outcome (in the case where $\Omega$ is countable) is sometimes referred to as an *atom*. This set is defined by the problem at hand. So, for instance, if our interest is in the outcome of the 2011 World Series, we might consider a set of 32 atoms, with one atom representing each of the MLB teams that might potentially win the World Series. The outcome space $\Omega$ is also referred to at times as the sample space, or the event space.

The definition of the outcome space is not unique. To expand on the World Series example, we could instead define the set of outcomes by the events Team A defeats Team B. Under this new definition, there would be $32 * 16$ possible outcomes corresponding to each of the possible winning teams defeating a team from the opposing league. We may also want to include an outcome for there being no winner of the World Series, if the season is unexpectedly cut short as happened in 1993. Furthermore, by taking conjunctions of outcomes of interest with other outcomes (e.g. 'Cubs win the World Series and I eat Grape Nuts for breakfast next Tuesday') the outcome

space can be made arbitrarily large. The correct choice of outcome space plays a significant role in the ascription of probability by demarking the considered outcomes from the unconsidered (and perhaps inconsiderable) outcomes. We won't have much to say further about the importance of the proper definition of the outcome space, but recognize the non-trivial nature of its proper definition for any given problem.

**Assumption 2.1.** *We will generally assume in the sequel that our outcome space is finite with* $|\Omega| = N < \infty$ *and will specifically note any instances where we generalize this assumption.*

In general we will associate with $\Omega$ a sigma-algebra $\mathcal{F}$ and thus $(\Omega, \mathcal{F})$ is a measurable space. Given Assumption 2.1, $\mathcal{F}$ is generally understood to be the power set $2^{\Omega}$.

#### Events

Next, we define an *event* (denoted by $A$) as a subset of outcomes ($A \subseteq \Omega$). In the case when $\Omega$ is infinite the definition of event needs to be handled with more care, by specifying first a sigma-algebra on $\Omega$ and then defining an event as a measurable set. However, since our default is Assumption 2.1, the question of measurability will largely be moot.

As an example of an event, turning again to the World Series outcome space, we might identify event $A$ as 'the winning team is from the American League'. This event $A$ would thus include all $\omega \in \Omega$ s.t. the event is satisfied. For instance, all $\omega$ corresponding to the Red Sox winning the World Series would belong to event $A$, as would all $\omega$ corresponding to the Yankees winning the World Series.

#### Random Variables

A random variable $X$ is defined as a mapping from $\Omega$ to $\mathbb{R}$ (again, neglecting questions of measurability given assumptions of finiteness). When it is clear from context, we will suppress the argument of the random variable, denoting it as $X$ rather than $X(\omega)$. An event $A$ can be identified with its characteristic (or indicator) random variable

$$\mathbb{1}_A(\omega) \triangleq \begin{cases} 1 & \omega \in A \\ 0 & \text{o.w.} \end{cases}.$$

Simple random variables are those which take on only finitely many values. Equivalently, any simple random variable $X$ can be represented as

$$X = \sum_{i=1}^{M} x_i \mathbb{1}_{A_i}$$

for some finite set of events. By Assumption 2.1, we will be dealing only with simple random variables.

For a set of random variables $\{X_i\}_{i=1}^{M}$, we will denote the random vector created

by the random variables as

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_M \end{bmatrix}$$

Thus as a random vector, $X$ is a mapping from $\Omega$ to $\mathbb{R}^M$. For a random vector $X$ we define the **outcome matrix** $\mathbf{X} = [x_{ij}]$ as $x_{ij} = X_i(\omega_j)$. Thus $x_{ij}$ is the realization of random variable $X_i$ under outcome $\omega_j$. When $X$ is a characteristic random variable (meaning for all $i, j$, $x_{ij} \in \{0, 1\}$), we will refer to $\mathbf{X}$ as the characteristic matrix and sometimes denote it by $\chi$.

The set of unique outcomes for a random variable $X$ will be called its 'alphabet' and denoted $\mathcal{X}$. For random vectors, the alphabet is the set of unique outcome vectors. As such, for a random vector $X$, $\mathcal{X} \subseteq \mathcal{X}_1 \otimes \mathcal{X}_2 \otimes \cdots \otimes \mathcal{X}_M$.

### ■ 2.3.2 Assessments

We can now give a mathematical definition of an assessment. Given an outcome matrix $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^{M \times N}$, an assessment $P$ is simply a mapping from $\mathcal{X}$ to the real numbers $\mathbb{R}^M$. As with random variables, when it is not needed for clarity we will suppress the random variable argument, denoting assessments merely as $P$ rather than $P(\mathbf{X})$. Also, when dealing with characteristic random variables, we will sometimes denote the assessment as $P(A)$, which should be taken to mean $P(\mathbb{1}_A)$.

Mathematical expectation is an example of an assessment. Given a personal probability distribution $Q$ over the set of outcomes, the expectation is a mapping from the set of possible outcomes of a random variable $X$ to the real numbers given by the equation

$$\mathbb{E}_Q[X] = \sum_i Q\{\omega_i\} X(\omega_i)$$

While we leave the mathematical definition deliberately loose, in practical terms an assessment is an attempt to summarize in some way the set of outcomes of the random variable. As such, we generally take $P(X) \in [\min x_i, \max x_i]$ where $x_i$ is a realization of the random variable $X$. An equivalent way of saying this is that generally for an assessment $P(X)$, $\exists \lambda \in [0, 1]^N$ s.t. $\sum_i \lambda_i = 1$ and $P(X) = \mathbf{X}\lambda$.

It is important to note that this general assumption about assessments is stated for a *single* random variable. We will generalize and strengthen this assumption when we introduce the concept of coherence in Section 2.4.

## ■ 2.4 What is Coherence

In Section 2.3 we mathematically defined an assessment as a mapping from a set of possible outcomes of a random vector of length $M$ to the vector set of real numbers $\mathbb{R}^M$. Here we introduce a fundamental property of assessments, called *coherence*. We begin by considering only characteristic random vectors, and then expand the discussion to all simple random variables.

### 2.4.1 Probabilistic Coherence

Coherence is perhaps the central tenet of subjectivist probability theory. Denote a vector of assessments of events $A_1, A_2, \ldots, A_M$ as $P(A_i), i = 1, 2, \ldots, M$. In de Finetti's formulation [26], this set of assessments is said to be coherent if and only if there does not exist a wager at the odds given by the assessments such that a non-negative outcome is guaranteed and a positive outcome is possible (a so-called Dutch Book). Savage [28] eschews the gambling narrative, relying on von Neumann-Morgenstern style utility theory combined with a few foundational axioms about preference among actions, but arrives at the same conclusions. In both cases, the subsequent theory (probability theory for de Finetti and decision theory for Savage) relies on this philosophical principle: incoherent assessments are fundamentally flawed. A more complete development of the similitaries and differences between de Finetti and Savage's theoretical developments can be found in Appendix A.

We follow de Finetti in identifying coherent assessments of characteristic random variables as *probabilities.*

**Definition 2.1.** *A **probability** is a coherent assessment of a characteristic random variable.*

Returning to our favorite World Series example, suppose I were asked to generate an assessment of the two events $A_1 = \{$A National League team wins the World Series$\}$ and $A_2 = \{$An American League team wins the World Series$\}$. We take as the outcome space $\Omega = \{\omega_1, \omega_2, \omega_3\}$ where $\omega_1$ corresponds to an NL team winning, $\omega_2$ corresponds to an AL team winning and $\omega_3$ corresponds to no team winning (for whatever reason). We can thus define the outcome matrix of events $A_i$ as

$$\chi = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

Suppose we have the following assessment of events $A_i$

$$P = \begin{bmatrix} 0.6 \\ 0.6 \end{bmatrix}$$

or equivalently $3:2$ odds for both events. A wise bettor chooses to take the short side of both sets of odds, at a dollar a piece (meaning he pays \$2 if the event **does** occur but receives \$3 if it **does not**). We now calculate the payoffs under each possible outcome, which are summarized in Table 2.1. Under $\omega_1$ (an NL team wins) the bettor

| Outcome | $\text{Gain}_1$ | $\text{Gain}_2$ | Total Gain |
|---|---|---|---|
| $\omega_1$ | -\$2 | \$3 | \$1 |
| $\omega_2$ | \$3 | -\$2 | \$1 |
| $\omega_3$ | \$2 | \$3 | \$5 |

**Table 2.1.** Gains under a Dutch Book wager

loses \$2 on the first bet but gains \$3 under the second bet (since event $A_1$ obtained but event $A_2$ did not). Thus the total gain under $\omega_1$ is \$1. Similarly, under $\omega_2$ the bettor gains \$3 on the first bet but loses \$2 on the second, for a total gain of \$1. Finally, under $\omega_3$ when neither team wins the bettor gains \$3 on the first and \$2 on the second for a total gain of \$5. Thus there exists a wager with a guaranteed positive outcome (specifically shorting both positions) and the assessment is incoherent, by definition.

We can write the mathematical condition of coherence as

$$\forall c \in \mathbb{R}^M, \ \min_j \left[ \sum_i c_i \left( \mathbb{1}_{A_i}(\omega_j) - P_i \right) \right] \leq 0 \leq \max_j \left[ \sum_i c_i \left( \mathbb{1}_{A_i}(\omega_j) - P_i \right) \right] \quad (2.1)$$

with equality iff $\sum_i c_i \left( \mathbb{1}_{A_i}(\omega_j) - P_i \right) = 0$ for all $\omega_j$. Note that this is somewhat redundant, as each inequality implies the other by considering $-c$, but writing it so builds intuition that for any position taken ($c$), there must be some possibility of gain or loss. More formal definitions replace the min and max with inf and sup, but this definition will suffice for our current development.

Parsing this equation, $c$ is the wager, with $c_i > 0$ meaning a bet that event $A_i$ occurs and $c_i < 0$ a bet that event $A_i$ does not occur (with $c_i = 0$ refusing the bet). The expression under minimization is the payoff under event $\omega_j$. Therefore, for any wager there is some outcome in which the gain of the wager is negative.

#### A Geometric Interpretation

Section 3.4 of [26] develops a secondary, geometric interpretation of coherence. Specifically de Finetti demonstrates that the philosophical interpretation of coherence is equivalent to requiring the assessment to lie in the space of all convex combinations of outcomes. Returning to our previous example, we can treat each column of the outcome matrix $\chi$ as a vertex of the 2-dimensional hypercube. To be coherent, in this geometrical interpretation $P$ must be representable as a convex combination of the points. So, for example, the assessment $P = [0.6 \ 0.6]^T$ is again seen to be **not** coherent because the assessment does not lie in the convex hull of the set of possible outcomes (shown in Figure 2-1).

#### A Measure-Theoretic Interpretation

An additional equivalence between coherence and the existence of a Kolmogorov-style probability measure was demonstrated in [69]. It was shown that de Finetti's coherence principle was equivalent to the existence of a finitely additive probability measure, i.e. a mapping $P$ from the outcome space to the real numbers such that:

1. $P(A_i) \geq 0$
2. $P(\cup A_i) = 1$
3. For disjoint, finite unions, $P(\cup A_i) = \sum P(A_i)$

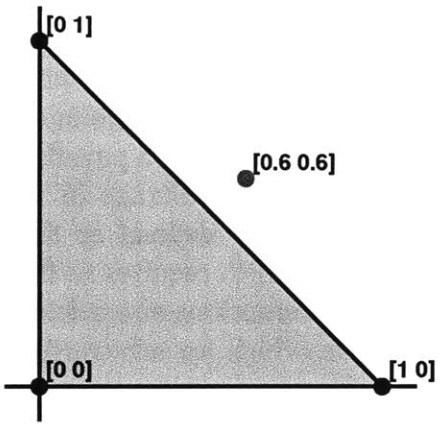



**Figure 2-1.** Coherent hull of outcomes

More generally, Kolmogorov probability measures must satisfy the stricter condition of *countable* additivity (i.e. $P(\cup A_i) = \sum P(A_i)$ for all countable unions of disjoint events). In [69] it was also shown that coherence in conjunction with a certain consistency condition on the events was sufficient to guarantee the existence of a countably additive probability measure that extended the assessment over the $\sigma$-algebra of the sample space. Since our default assumption is finiteness of the outcome space, the difference between countable and finite probability measure will largely be moot for our purposes.

### ■ 2.4.2 Coherence and Arbitrage

Several authors (e.g. [42, 70–72]) mention the similarity between the concepts of coherence and no arbitrage. Conceptually, this is perhaps best suggested through the example of the 'money pump.' A money pump, as introduced by Ramsey [27], is a philosophical justification for requiring transitivity among preferences (i.e. if option A is preferred to option B and option B is preferred to option C then option A must be preferred to option C). If not, suggests Ramsey, a wily agent could take advantage of your intransitivity to create unlimited wealth. Specifically, suppose that A is preferred to B is preferred to C is preferred to A (ans so on recursively). Then, suppose I first give you C, then offer to trade your C, plus some marginal utility, for my B. Since you prefer B to C, there is some amount of marginal utility such that a rational actor will engage in the trade. Next I offer to trade your B (again, plus some marginal utility) for my A. And then trade your A (plus marginal utility) for the original C. We are now in the original state, except I am richer by the (necessarily positive) differences of marginal utilities between A and B, B and C, and C and A. This process could be repeated *ad infinitum* to generate arbitrarily large amounts of wealth for the advantaging agent.

Ramsey's description of the money pump is quite similar in nature to the philo-

sophical rationale for coherence given by de Finetti [26]. De Finetti himself recognized Ramsey's work as influential on his thinking. Thus, the Dutch book argument put forward by de Finetti to justify requiring coherence among assessments is essentially a 'money pump' style argument: if assessments are incoherent than a wily bettor can create a book with arbitrarily large guaranteed payoff.

The same is true of the concept of arbitrage in markets. Arbitrage, following Ross's influential work [44] is generally defined as the existence of a portfolio, or mixture over primitive investments, which requires no financing but has almost certain positive return. Given standard assumptions about frictionless trading, arbitrage can be considered a money pump, in which an arbitrarily large payoff can be realized risk-free.

#### Formal Connections between Arbitrage and Incoherence

In a formal context, the close relationship between arbitrage and incoherence was noted in [73]. In this paper it is shown that, treating a market of contingent claims as a vector lattice, the no arbitrage condition is equivalent to the price of contingent claims being a strictly positive linear functional. It is then demonstrated that, treating events as unit-value contingent claims, de Finetti's coherence axiom is equivalent to the probabilities of events under assessment being a positive linear functional. Thus the no arbitrage condition in a contingent claims market where each claim has unit value is a slight strengthening of de Finetti's coherence axiom.

Clark's argument is persuasive, but is given in terms of a highly abstracted mathematical model. In this section we make an argument similar to Clark's, that probabilistic incoherence is a special case of arbitrage, but do so in terms of the stochastic calculus model given in [70].

Adapting a simplified version of Kartzas' model, consider a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$ and an adapted process $S(t)$ representing $N$ asset prices over time, where $S_0(t)$ is the "risk-free" asset. We'll assume $S_0(t) \equiv 1$ (meaning that it's just cash, no interest). Define $R(t)$ as the "excess yield (over the interest rate) process." In our market of interest, with no interest rate, no dividends, etc. this is simply $\int_0^t \sigma(u)dW(u)$ where $W$ is the Brownian motion process and $\sigma(u)$ is the volatility process that together underlie the change in asset prices. Next, let $\pi(t)$ be the portfolio process (i.e. the amount invested in each of the $N$ assets at time $t$). Assume this is progressively measurable (no trading off future information). Then we can derive the gains process $G(t)$ as

$$G(t) = \int_0^t \pi(u)^T dR(u)$$

An arbitrage opportunity exists if there is some progressively measurable $\pi$ s.t. $G(t) \geq 0$ a.s. and $G(t) > 0$ with positive probability.

Now consider a set of $M$ idealized unitary bets (i.e. that pay unit amount when the event obtains and nothing otherwise) that are made and instantaneously realized, like betting on the outcome of a coin-flip immediately before the outcome is revealed. This matches the market assumptions above of $S_0(t) \equiv 1$, no mean rate of return on

investments, etc. The differential on the excess yield process, $dR(t)$ at the instant the bet is made and revealed is exactly $\mathbb{1}_{A_i}(\omega) - \tilde{P}(A_i)$ where $A_i \subseteq \Omega$ is an event, $\mathbb{1}_A$ is the indicator of event $A$ and $\tilde{P}(A)$ is the assessment of event $A$. So for this instantaneous, idealized market, the no arbitrage condition becomes that there is no $\pi \in R^N$ s.t.

$$\sum_i^N \pi_i \left[\mathbb{1}_{A_i}(\omega) - \tilde{P}(A_i)\right] \geq 0 \ (a.s.)$$

and

$$\sum_i^N \pi_i \left[\mathbb{1}_{A_i}(\omega) - \tilde{P}(A_i)\right] > 0 \ (w.p.p)$$

For finite $\Omega$ with $P(\omega) > 0$ this can be stated as

$$\forall \pi, \ \max_{\omega \in \Omega} \sum_{i=1}^N \pi_i \left[\mathbb{1}_{A_i}(\omega) - \tilde{P}(A_i)\right] \geq 0$$

which is exactly the definition of coherence given in Equation 2.1.

This argues that coherence could be thought of as the non-existence of instantaneous arbitrage in instantaneous opinion markets (i.e. markets whose assets are unity bets and whose prices are assessments).

### ■ 2.4.3 Coherence of Non-Characteristic Random Variables

Thus far we have developed the concept of coherence as pertains to characteristic random variables, yielding a definition of probability based on the non-existence of a Dutch book or, equivalently, an arbitrage-free single-stage opinion market. We will now expand the concept of coherence to non-characteristic (but simple, keeping in mind the finiteness assumption) random variables.

We begin by revising Equation 2.1 to reflect a Dutch Book argument for random variables with non-unitary values.

$$\forall c \in \mathbb{R}^M, \ \min_j \left[\sum_i c_i \left(X_i(\omega_j) - P_i\right)\right] \leq 0 \tag{2.2}$$

with equality iff $\sum_i c_i \left(X_i(\omega_j) - P_i\right) = 0$ for all $\omega_j$. Comparing Equation 2.1 to Equation 2.2 we see that the 'payoff' portion of the wager has changed from $\mathbb{1}_{A_i}(\omega_j) - P_i$ to $X_i(\omega_j) - P_i$, and the rest of the equation has remained the same. The interpretation of the condition is nearly identical: $c$ now represents the purchase of some number of contracts at price $P$ whose outcomes are $X$. An assessment is coherent iff, for any position taken $c$, there is non-zero probability of both positive and negative gain.

De Finetti referred to general coherent assessments as *previsions*, of which probabilities are a special class, i.e. previsions of characteristic random variables. In his formulation, a prevision "consists in considering, after careful reflection, all the possible alternatives, in order to distribute among them, in the way which will appear most

appropriate, one's own expectations, one's own sensations of probability" [26]. One of de Finetti's purposes in employing the term 'prevision' was to maintain notational consistency with 'probability' and the assessment denotation $P$. We believe $P$ could also accurately be considered a *price* for reasons that will be shown in Section 2.4.4. De Finetti also used the term 'price' to refer to this quantity when applied to market goods, but preferred prevision when applied to non-marketable quantities.

#### An example of non-characteristic incoherence

Returning to our World Series example, suppose we define two random variables $X_1$ and $X_2$. $X_1$ is a contract with a Cubs fan to share in his feelings of joy or despair following the World Series and $X_2$ is an analogous contract with a Red Sox fan (note: the Cubs are an NL team and the Red Sox are an AL team). Since Cubs and Red Sox fans have some affinity for each other, and for the other teams in their own league (but not teams from the opposing league), suppose the outcome matrix can be given as

$$\mathbf{X} = \begin{bmatrix} 3 & 1 & 1 & 0 & 0 \\ 1 & 2 & 0 & 1 & 0 \end{bmatrix}$$

where the outcomes can be interpreted as follows: $\omega_1$ = Cubs win, $\omega_2$ = Red Sox win, $\omega_3$ = NL team other than Cubs wins, $\omega_4$ = AL team other than Red Sox wins, $\omega_5$ = no one wins (obviously given their long history of failure, Cubs fans will derive more utility from their team winning the World Series than will Red Sox fans).

Suppose one is called on to give an assessment on this outcome space, and the resulting assessment is

$$P = \begin{bmatrix} 2.2 \\ 1.5 \end{bmatrix}.$$

As in the previous example, a wily investor analyzes the assessment and decides to short both positions (in this case, since the assessments no longer correspond to odds, we take the gains directly from Equation 2.2; it is mathematically equivalent to the previous example). Table 2.2 summarizes the payoffs under each outcome, demonstrating that the assessment is incoherent (and thus not a prevision).

| Outcome | $\text{Gain}_1$ | $\text{Gain}_2$ | Total Gain |
|---|---|---|---|
| $\omega_1$ | -\$0.8 | \$1.5 | \$0.7 |
| $\omega_2$ | \$2.2 | -\$0.5 | \$1.7 |
| $\omega_3$ | \$1.2 | \$1.5 | \$2.7 |
| $\omega_4$ | \$2.2 | \$0.5 | \$2.7 |
| $\omega_5$ | \$2.2 | \$1.5 | \$3.7 |

**Table 2.2.** Gains under incoherent pricing

#### Geometric and Measure-Theoretic Equivalences

The geometric equivalence first introduced for characteristic random variables can be extended analogously to non-characteristic random variables. Thus in Figure 2-2 we

see that the incoherent assessment from the above example lies outside the convex hull of the columns of the outcome matrix $\mathbf{X}$. It is interesting to note that in the case of

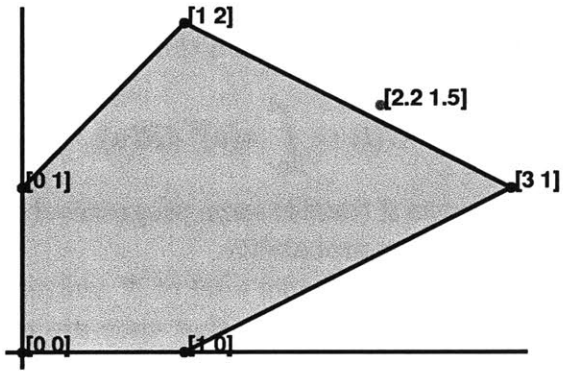



**Figure 2-2.** Convex hull of outcomes for non-characteristic random vector

characteristic random vectors each unique outcome necessarily lies on the boundary of the set of coherent assessments. In the more general case of non-characteristic random vectors it is possible for one of the outcome points to lie strictly within the interior of the convex hull. For example, given an outcome matrix

$$\mathbf{X} = \begin{bmatrix} 2 & 1 & 0 \\ 2 & 1 & 0 \end{bmatrix}$$

the outcome $X(\omega_2) = \begin{bmatrix} 1 & 1 \end{bmatrix}^T$ lies strictly in the interior of the convex hull. As such it is superfluous to determining the coherence of any given assessment.

Since the assessment is no longer a vector of probabilities, the measure-theoretic interpretation does not directly extend. However, as is evident from the geometric interpretation, coherent assessments are equivalent to expectations of the random vector.

**Proposition 2.1.** *An assessment is coherent if and only if* $\exists Q$ *such that* $(\Omega, \mathcal{F}, Q)$ *is a probability space and* $P = \mathbb{E}_Q[X]$.

where this proposition should be understood as only applying on finite outcome spaces.

This proposition subsumes the measure-theoretic interpretation of coherent assessments of characteristic random variables as probabilities on a finite outcome space through the identity that, for a probability measure $Q$ and $A \in \mathcal{F}$, $Q(A) = \mathbb{E}_Q[\mathbb{1}_A]$.

### 2.4.4 Arbitrage and Coherence for Non-characteristic Random Variables

Unsurprisingly, the equivalence between incoherence for characteristic random vectors and arbitrage in a single-shot unit-value payoff market can be carried through analogously to the case of non-characteristic random variables. As before consider a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, P)$ the N-valued price process $S(t)$ with $S_0(t) \equiv 1$. All other stochastic quantities, and the definition of arbitrage, are as given in Section 2.4.2.

$$G(t) = \int_0^t \pi(u)^T dR(u)$$

An arbitrage opportunity exists if there is some progressively measurable $\pi$ s.t. $G(t) \geq 0$ a.s. and $G(t) > 0$ with positive probability.

Now, rather than unitary bets, consider a set of $N$ contracts with prices $\tilde{P}_i$ that pay $x_{ij}$ when the event $\omega_j$ obtains, and assume that they are made and instantaneously realized. Then differential on the excess yield process, $dR(t)$ at the instant the bet is made and revealed is exactly $X_i(\omega) - P_i$ and the corresponding gains process

$$G(t) = \int_0^t \pi(u)^T dR(u) = \sum_i \pi_i (X_i(\omega) - \tilde{P}_i)$$

and therefore the no arbitrage condition becomes that there is no $\pi \in R^N$ s.t.

$$\sum_i^N \pi_i \left[ X(\omega) - \tilde{P}_i \right] \geq 0 \ (a.s.)$$

and

$$\sum_i^N \pi_i \left[ X(\omega) - \tilde{P}_i \right] > 0 \ (w.p.p)$$

For finite $\Omega$, if we assume as before that $P(\omega) > 0$ this can be stated as

$$\forall \pi, \ \max_{\omega \in \Omega} \sum_{i=1}^N \pi_i \left[ X(\omega) - \tilde{P}_i \right] \geq 0$$

which matches the definition of coherence given in Equation 2.2.

Thus, as in the case of characteristic random variables, coherence can be thought of as a no-arbitrage condition in a single-shot market.

## 2.5 Approximation and Approximation Cost

This thesis will develop a method of coherent approximation of incoherent assessments. At the core of this endeavor is a cost minimization problem. In Chapter 3 an existing method of coherent approximation is introduced and analyzed. Certain deficiencies are noted and an alternative formulation is introduced.

As an example of the centrality of cost to an approximation method, consider the

following example: given some observations $y$, and a distribution $P(x|y)$ over a set of possible values $x \in \mathbb{R}$, we wish to choose an approximation, or estimate $\hat{x}(y)$ that "best" represents the true value. Depending on how the concept of "best" is defined, this estimator may take on significantly different forms. For instance, if we adopt the minimum absolute error (MAE) cost function

$$C(\hat{x}, x) = |\hat{x} - x|$$

it can be shown that the optimal estimate is the median of $P(x|y)$. If instead we adopt the minimum uniform cost

$$C(\hat{(x)}, x) = \begin{cases} 0 & |\hat{x} - x| < \epsilon \\ 1 & \text{o.w.} \end{cases}$$

and let $\epsilon \to 0$, the optimal estimate is the mode of $P(x|y)$ (the so-called Maximum a Posteriori, or MAP, estimate). Finally, if we take

$$C(\hat{(x)}, x) = ||\hat{x} - x||_2^2$$

the optimal estimate is the mean value under $P(x|y)$, or Bayes' Least Squares (BLS) estimator.

Which of these cost functions, and hence which estimator should be used, is situation dependent. If the object under estimation is the position of an enemy force which we wish to eliminate with a bomb with destruction radius $\epsilon$, the minimum uniform cost may be the right cost function. Whereas if we are trying to estimate the amount of time we will have to wait for the next bus, the maximum absolute error cost function may be more appropriate (unless, as happens for some of us, time moves increasingly non-linearly as the wait period increases, in which case we might consider the least squares cost).

In Chapter 3 a new framework for performing coherent approximation is suggested. While we are at pains to demonstrate many of the advantages of this alternative formulation, it should always be recognized that the best cost function is the one that most closely models the true costs of the system under analysis. As such, the true value of this formulation will be in providing an alternative mechanism for coherent approximation that may be "better" than the previous suggestions in certain circumstances.

# Chapter 3

# Coherent Approximation

## 3.1 Introduction

In the previous chapter we developed the concept of assessments, which are properly viewed as mappings from sets of potential outcomes of random vectors to real numbers. We also introduced arbitrage/Dutch Book arguments for why such assessments should be coherent.

However, the human experience tells us that when such assessments are made in a distributed fashion they are often incoherent. As a documented example, consider the challenges that off-track betting created for the horse race industry. In the 1980s changes to regulations allowed horse tracks to accept bets for races run at other tracks. In effect, deregulation created a distributed assessment problem. In [74, 75] the impact of this distributed assessment problem is analyzed in depth. It is shown that, even accounting for racetracks' fees, the distributed assessments of racetracks were sufficiently incoherent to provide guaranteed positive outcomes for savvy bettors.

While arbitrage among horse-race odds setters is of minor importance in the overall scope of world affairs, less trivial is the use of 'opinion markets' to provide assessments to decision makers about the outcomes of uncertain events. Somewhat more will be said about opinion markets in Chapter 4, but here we suffice it to say that using distributed assessment to estimate probabilities of uncertain events plays an increasingly important role in the decision making processes of governments, businesses and individuals. While the calibration of such markets has been investigated (and, at times, criticized), a critical question is how to identify *a priori* whether the assessments are poorly calibrated or not. A further critical question would be if it's possible to identify miscalibrated assessments, how should they be corrected to better reflect the true uncertainty about outcomes.

If distributed assessments are incoherent then, by the theory developed in Chapter 2, they are fundamentally flawed as a representation of the true uncertainty about the events or random variables under assessment. If coherence is the answer to the question of how to identify mis-calibrated assessments *a priori*, the question of what to do when such assessments are identified is less obvious. Previous literature [76–78] has suggested an optimization framework for approximating incoherent assessments of characteristic random variables by probabilities.

In this chapter we develop a method of coherent approximation applicable to all random vectors with finite alphabets. We first report on some results relating

to the computability of coherent approximations in the framework of [76]. We will then turn our attention to a re-formulation of the Coherent Approximation Principle relying on information geometry. There are several benefits to this re-formulation, and we will spend some time justifying it both theoretically and empirically. We conclude the Section with a brief discussion of the impact of additional structural information (e.g. Markovianity w.r.t. a graph) on the feasible sets of assessments over characteristic random variables.

### 3.1.1 Non-optimization Methods of Assessment Fusion

The idea of coherent approximation, or fusion, advanced in this thesis is to modify assessments minimally until they are coherent. Several non-optimization based methods of coherent approximation have been previously advanced in the literature.

One well-explored method for fusing assessments is the pari-mutuel betting system. Beginning with the work of Eisenberg and Gale [79] there have been a series of papers exploring the ability of the pari-mutuel system to fuse experts' subjective judgments of probability [80,81]. This particular fusion method is interesting in that it both elicits the experts' assessments and simultaneously fuses them into the overall group assessment. There are questions, however, as to whether observed biases within the parimutuel system might be attributable to suboptimality of the "race track" fusion method. Furthermore, the creation of a simple betting market is not always a feasible solution to the problem of generating a joint estimate of probability.

A second group of researchers, predominantly in the field of psychology, have sought to used improved elicitation techniques, such as feedback, in order to remove assessment biases (and hence miscalibration) [82–84]. Another approach is to put the experts in contact with each other and allow them to come to a consensus through discussion and debate, while others have questioned the wisdom of fusing assessments at all, particularly when the range of opinion is large (c.f. [85] and references therein).

A more mathematical approach is to elicit experts' probability assessments first and then functionally fuse them in away that agrees with all known contextual information. Background surveys on such mathematical techniques can be found in [86–88]. Much work has also been done within the computer science and machine learning community on the aggregation of expert opinion (in this case, algorithms). Aggregation algorithms are often referred to under the general rubric of classifier or generalization combination [89,90]. In this case, the question is how to best judge the "correctness" of the fused solution.

Perhaps the most popular method for fusing experts' assessments is a simple or weighted linear average [86,91,92]. A similar approach is propounded in [93], wherein the "experts" evaluations are actually individual attribute percentile scores (e.g. student GPA, SAT scores, etc.). This is also a common technique used within the computer science literature on combining classifiers, including popular algorithms for "bagging" or "boosting" [90] and stacked generalization [94]. One of the challenges in such an approach is determining what weightings to provide each expert's opinion. Much work has gone into determining "optimal" weightings.

The linear averaging approach has the benefit of simplicity and intuitiveness, but

has been shown [77] to result in incoherent fused assessments in the presence of both abstaining and/or individually incoherent experts. Another common approach to fusing analysts' subjective probability assessments is by treating each assessment as an observation, and updating according to Bayes' Rule [95, 96]. This is sometimes referred to as the supra-Bayesian approach [87]. One of the disadvantages of this approach is the need to specify a likelihood for each expert's assessment, as well as a prior.

## ■ 3.2 Coherent Approximation Principle

In a series of papers [76, 77, 97] the geometric view of coherence demonstrated in Chapter 2 is used to formulate an optimization-based method for aggregating experts' assessments of probability. In [76, 97] the suggestion is to select as the fused probability assessment of a set of experts the point in the coherent hull that lies closest (in terms of the standard Euclidean norm) to the vector of expert assessments. Mathematically, we represent this as

$$\lambda^* = \arg\min_{\{\lambda | \sum_i \lambda_i = 1,\ \lambda_i \geq 0\}} ||P - \chi\lambda||_2 \tag{3.1}$$

with the optimal coherent approximation given by $P^* = \chi\lambda^*$. This is termed the *Coherent Approximation Principle (CAP).*

In [77] an approximation to the CAP is suggested to deal with the potential combinatorial growth (in the number of assessed events) of the computation of the exact CAP solution. The computational question was further developed in [98], in addition to analyses of the CAP under an alternate cost structure and its combination with additional constraints on the set of joint probability distributions. Much of the original work in [98] is reported here in Sections 3.2-3.3 and 3.6-3.7.

Note in Equation 3.1 that the optimization occurs in the space of atomic events $\Omega$, which may grow exponentially in the assessment space size. To mitigate this computational challenge, previous authors [77] proposed a hybrid approach between linear averaging and coherent approximation. Unfortunately, this approach will generally produce a fused estimate that is not coherent.

#### Example of Coherent Approximation

In Section 2.4.1 an assessment problem was formulated involving the characteristic matrix and assessment pair

$$\chi = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \qquad P = \begin{bmatrix} 0.6 \\ 0.6 \end{bmatrix}.$$

It was shown that the assessment was incoherent (and therefore could not represent the probabilities of the events). The CAP given in Equation 3.1 would suggest an optimal approximation of this incoherent assessment by the vector $\begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix}$ as depicted in Figure 3-1.

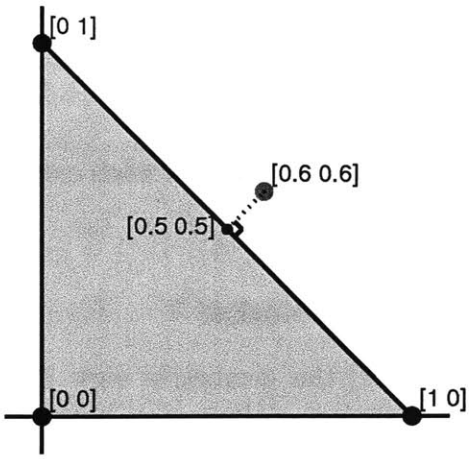



**Figure 3-1.** Coherent Approximation of an Incoherent Assessment

In the following subsections we develop a general fusion rule that operates in the assessment space and generates a coherent fused assessment.

### ■ 3.2.1 Monadic Structure

Consider the class of characteristic matrices such that

$$\sum_j \chi_{ij} \leq 1$$

We will refer to matrices in this class as *monadic*, meaning that each event under assessment is, at most, a singleton.

In this case, it is simple to show that a closed-form solution exists to the problem of finding a coherent approximation to an incoherent assessment. Let $P^* = \chi\lambda^*$ be defined by Equation 3.1 and let

$$\bar{P}_i = \frac{1}{n_i} \sum_{j \in \mathcal{N}_i} P_j \tag{3.2}$$

where $\mathcal{N}_i = \{j | A_j = A_i\}$ and $n_i = |\mathcal{N}_i|$. Define the probability excess/deficit as

$$D = 1 - \sum_{j=1}^{N} \frac{1}{n_j} \bar{P}_j$$

and assume wlog that $n_1\bar{P}_1 \leq n_2\bar{P}_2 \leq \cdots \leq n_N\bar{P}_N$ and $\forall i\ \mathcal{N}_i,\ j < k,\ j,k \in \mathcal{N}_i \Rightarrow \{j, j+1, \ldots, k\} \subseteq \mathcal{N}_i$.

**Proposition 3.1.** *If $\chi$ is monadic, then*

$$P^* = \bar{P} + \Delta$$

where we define $\Delta \in [0,1]^N$ as

$$\Delta_i = \begin{cases} f(\{\Delta_j\}_1^{i-1}, \{n_j\}, \bar{P}) & D < 0 \\ \frac{D}{n_i}\left(\sum_j \frac{1}{n_j^2}\right)^{-1} & D \geq 0 \;\&\; \nexists n_k = 0 \\ 0 & \text{o.w.} \end{cases}$$

with, for a given $i \in \{1, 2, \ldots, N\}$,

$$f(\cdot,\cdot,\cdot) = \begin{cases} \Delta_{i-1} & \mathcal{N}_i = \mathcal{N}_{i-1} \\ \max\left(-\bar{P}_i, \frac{n_{i-1}}{n_i}\Delta_{i-1} + g_i\right) & \text{o.w.} \end{cases} \tag{3.3}$$

with $g_i = \frac{1}{n_i}\left(\sum_{j=i}^N \frac{1}{n_j^2}\right)^{-1}\left(D - \sum_{j=1}^{i-1}\frac{\Delta_j}{n_j}\right)$ and $\Delta_0 = \bar{P}_0 = 0$.

When the characteristic matrix is monadic, each assessment event corresponds to a unique element of the sample space. All events corresponding to the same unique element of the sample space must have equal assessments to be coherent, leading to $\bar{P}$ in the coherent approximation solution. Starting from this point, we must satisfy the constraint that the assessments (or, equivalently, the corresponding weights on the unique elements of the sample space) must sum to one. If $D \geq 0$ (assessment deficit) this means we need to add a bit to $\bar{P}$; if $D < 0$ (assessment excess) we must take something away from the averaged assessments.

When $\exists n_k = 0$ it means there is some element of the sample space with no equivalent event under assessment. Therefore, if $D \geq 0$ we can meet the assessment deficit by weighting this unassessed element, which leaves the assessment unaffected ($\Delta = 0$). If no such element exists, then the optimal strategy is to add to each element of the sample space in proportion to the number of assessors that have it as their unique element ($\Delta = \frac{D}{n_i}\left(\sum_j \frac{1}{n_j^2}\right)^{-1}$).

In the case that $D < 0$ we must remove probability. The principle is the same; probability should be removed proportional to the number of events assessed against each of the sample elements. However there is a twist in that such a solution could remove "too much" weight from some element, resulting in a negative $\lambda$. The recursive formula given in Equation 3.3 is essentially a thresholding rule that prevents any element from becoming negatively weighted.

Proposition 3.1 can be extended to a broader class of matrices by using the complementary completeness of probabilities (i.e. each assessment $P$ of $A$ is also an indirect assessment $(1 - P)$ of $A^c = \Omega \backslash A$). A characteristic matrix is said to be *generalized* monadic if

$$\sum_j \chi_{ij} \leq 1 \qquad \text{or} \qquad \sum_j \chi_{ij} \geq N - 1$$

As before, let $P^*$ be defined by Equation 3.1. Let $\bar{P}_i$ be the mean of all assessments

of event $A_i$ (both directly over $A_i$ and indirectly over $A_i^c$).

**Corollary 3.1.** *If $\chi$ is generalized monadic, then*

$$\forall i \text{ s.t.} \sum_j \chi_{ij} \leq 1, P_i^* = \bar{P}_i + \Delta_i$$

*Also*

$$\forall i \text{ s.t.} \sum_j \chi_{ij} \geq N - 1, P_i^* = 1 - \bar{P}_i + \Delta_i$$

with $\Delta$ defined analogously to Proposition 3.1.

The benefit of Proposition 3.1 and its generalization is that there exist certain matrices for which the CAP can be solved exactly, with computational complexity proportional to the number of events under assessment (rather than the potentially exponentially larger number of atomic events, as in the direct solution to the CAP problem).

### ■ 3.2.2 Suboptimal Approximation

There is not a simple extension of the result from Section 3.2.1 to generally structured characteristic matrices. However, it is possible to use the solution for monadic matrices to approximate the solution for general characteristic matrices.

Consider again the coherence constraint: $P = \chi\lambda$. This can be rewritten in the following way

$$P = \sum_j [\chi_{i,S_j}]\lambda_{S_j}$$

where $\{S_j\}$ forms a partition over $\{1, 2, \ldots, |\Omega|\}$. Essentially, we've decomposed the original characteristic matrix columnwise. It is simple to show that for any characteristic matrix there exists a columnwise decomposition such that $\forall j$ $[\chi_{i,S_j}]$ is generalized monadic.

Also, Proposition 3.1 can be generalized to the constraint $\sum \lambda_i = \alpha_j$ where $\alpha_j$ is some given constant, rather than $\sum \lambda_i = 1$. Combining these two results gives a method of suboptimal coherent approximation

1. Decompose characteristic matrix columnwise into a set of monadic matrices
2. Apply Proposition 3.1 to each subproblem, with the constraint that $\sum_i (\lambda_{S_j})_i = \alpha_j$
3. Using Lagrangian analysis, determine optimal $\alpha_j$ coupling constants

The resulting solution is guaranteed coherent. Furthermore, the suboptimality of the result can be bounded by $\sum_i \sum_{j \neq i} \lambda_{S_i}^T \lambda_{S_j}$.

## 3.3 Divergence-based Coherent Approximation

In the following Section we will make two intuitive arguments for a cost function in Equation 3.1 based on binary KL divergence rather than quadratic cost. The first argument, given in Section 3.3.1 is based on a particular mathematical model of the experts' assessments; the second, given in Section 3.3.2 is based on an intuitive approach to a specific coherent approximation problem.

### 3.3.1 Opinion Deformation and Sanov's Theorem

Consider the following model for probability assessors: each assessor observes a sequence of realizations of her event over several periods of time and maintains an empirical distribution of event occurrence. When called upon to make an assessment, the assessor selects a distribution approximately equal to her empirical distribution from finite set $\mathcal{P}$.

When all assessors have reported it is noted the set of assessments is incoherent and therefore at least one assessor is in error, either due to approximation or due to miscalculation of the empirical distribution. Given that at least one reported assessment is in error, which is the most likely generating distribution to have caused the error(s)?

First, let's consider the most likely estimate of a true distribution for a single assessor. Let $p^*$ be the true generating distribution for the assessor, and $\hat{p}$ be the reported distribution. From Sanov's theorem, the probability of declaring $p_j$ when the true generating distribution is $p_i$ decays exponentially in $n$ and is approximately

$$P(\hat{p} = p_j | p^* = p_i) \doteq \exp(-nD_b(p_j||p_i)) \tag{3.4}$$

where

$$D_b(p_j||p_i) = p_j \log\left(\frac{p_j}{p_i}\right) + (1-p_j)\log\left(\frac{1-p_j}{1-p_i}\right) \tag{3.5}$$

and $\doteq$ denotes asymptotic equality to within a multiplicative factor with a slower rate of decay. (More accurately, the asymptotic rate of decay is $\inf_{p \in A(p_j)} D_b(p \,\|\, p_i)$ where $A(p_j)$ is the set of all distributions mapped to $p_j$ by $\hat{p}$. We've made the simplifying assumption that the acceptance regions $A(p_j)$ are uniformly small in a divergence sense).

Equation 3.4 gives an asymptotic, approximate expression for the posterior of an observation distribution given a specific generating distribution, but we wish to maximize the probability of a generating distribution given an observation distribution (i.e. $P(p_i = p^* | p_j = \hat{p})$). We make the assumption that the prior probability that $p^* = p_i$ is uniform over all $p_i$. Then, by Bayes' rule, the posterior distribution $P(p^* = p_i | \hat{p} = p_j)$ is (asymptotically, approximately) proportional to $\exp(-nD_b(p_j||p_i))$. Conditioning on the event $i \neq j$ (i.e. the assessor is in error), the likelihood can be given as

$$P(p^* = p_i | \hat{p} = p_j, i \neq j) \propto \begin{cases} \exp(-nD_b(p_j||p_i)) & i \neq j \\ 0 & i = j \end{cases}$$

Therefore the maximum likelihood estimate of $p^*$ given $\hat{p} = p_j$ and $\hat{p} \neq p^*$ is asymptotically

$$p_i^* = \underset{p_i \in \mathcal{P} \setminus \{p_j\}}{\arg\min} D_b(p_j || p_i) \tag{3.6}$$

Relating this to the multiple assessor case, given a set of incoherent assessments $p_1, p_2, \ldots, p_N$ we know at least one of the assessments is in error. If we knew which one, by the preceding development it should be revised following Equation 3.6. Instead, assume the assessors have independent observation processes. Then, by a development similar to the singe assessor case, we see that

$$P(p^* = p_i | p_{j_1}, p_{j_2}, \ldots, p_{j_M}) \propto \exp(\sum_{k=1}^{M} n_k D_b(p_{j_k} || p_i))$$

This expression has an intuitive interpretation; the weights $n_k$ represent reputational values, dependent on how many observations the assessor has made. Absent reputational information, we make the simplifying assumption that assessors have equal amounts of information, giving the following expression for the ML estimate of the generating distribution.

$$\lambda^* = \arg\min_{\{\lambda | \sum_i \lambda_i = 1,\ \lambda_i \geq 0\}} \sum_{i=1}^{N} D_b(p_i \,\|\, \chi_i \lambda) \tag{3.7}$$

where $\chi_i$ is the $i^{th}$ row of $\chi$.

### ■ 3.3.2 Coherent Approximation of Probability Mass Functions

Consider a special case of Equation 3.1 in which the characteristic matrix $\chi$ is equal to the identity. In this case, we can rewrite the optimization problem as

$$Q^* = \underset{\{Q | \sum_i Q_i = 1,\ Q_i \geq 0\}}{\arg\min} C(P, Q) \tag{3.8}$$

#### Solution under Quadratic Cost Function

Consider an identity characteristic matrix. This would suggest an assessment regime in which each assessor is attempting to provide the probability of an atomic event. It is simple to see that the identity matrix is monadic and therefore the closed for solution to the CAP is given by Proposition 3.1. In this case, since each atom is under assessment by a single assessor, the solution will be increase or decrease each element of the assessment by an equal amount (respecting positivity requirements, of course) until the assessment is coherent.

As an example, take $N = 3$ and an assessment

$$P = \begin{bmatrix} .2 \\ .7 \\ .7 \end{bmatrix}.$$

| $i$ | $P_i$ | $D_i$ | $R_i$ | $Q_i$ |
|---|---|---|---|---|
| 0 | - | 0 | 0.7 | - |
| 1 | 0.2 | 0.2 | 0.1 | 0 |
| 2 | 0.7 | 0 | 0 | 0.5 |
| 3 | 0.7 | 0 | 0 | 0.5 |

**Table 3.1.** Optimal coherent approximation to quadratic approximation example

The solution given by Proposition 3.1 as shown in Table 3.1 would be

$$Q = \begin{bmatrix} 0 \\ 0.5 \\ 0.5 \end{bmatrix}$$

This demonstrates one of the limitations of using the quadratic cost in formulating the coherent approximation problem. The change from believing an event is 20% likely to believing an event has 0% chance of occuring *seems* more significant than the change from believing an event is 70% likely to 50% likely. Cognitively, the gap between moderate uncertainty and absolute certainty is larger than between moderate uncertainty and high uncertainty. This is more clearly seen in terms of odds: the odds assessments of $A_1$ and $A_2$ in our example are shifting from 7:3 to 1:1, while the assessment for $A_3$ is moving from 1:4 to 1:$\infty$. Or in other words, whereas the third expert states he'd be willing to accept either side of a wager that pays out \$4 if event $A_3$ occurs and \$1 if it does not, the coherent approximation would change his opinion that he'd be unwilling to take the long side of the wager *regardless of the potential payout.* Such an approximation represents a dramatic revision of the expert's assessment that the CAP is incapable of capturing because its objective is to minimize the Euclidean norm between the assessment and its approximation.

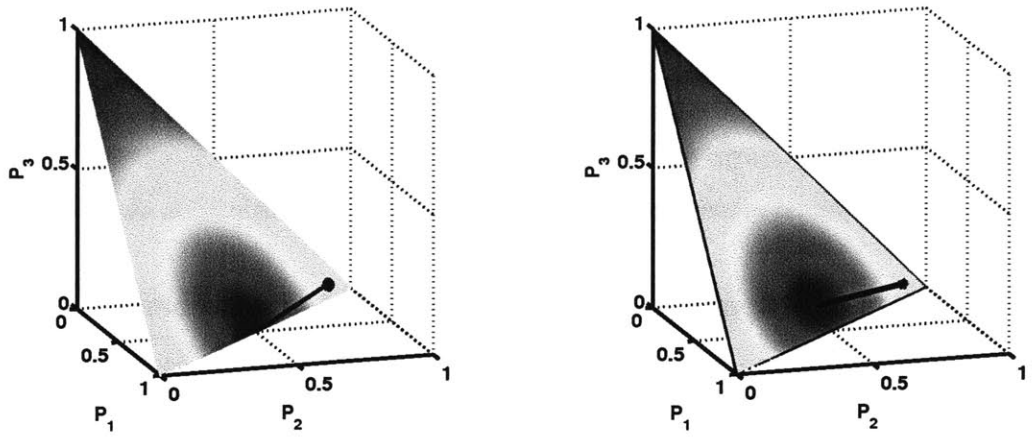
A more natural method for transforming a set of assessments into a probability mass function is simply to scale the assessments until they sum to one. In the following section it will be seen how the binary KL-divergence cost explains both why this is natural and why it is not the right thing to do (quite).

#### Solution under KL Divergence-based Cost Function

Now, instead of the quadratic cost function in Equation 3.8, take

$$C(P,Q) = \sum_{i=1}^{N} D_b(P_i||Q_i)$$

where $D_b(P_i||Q_i)$ is given by Equation 3.5. The rationale for choosing this cost function is given in Section 3.3.1. The difference between the two objective functions is depicted graphically in Figure 3-2. Notice particularly that the divergence-based cost function generates natural barriers around the boundaries of the simplex. This




(a) (b)

**Figure 3-2.** Optimal coherent revisions for example under (a) quadratic cost and (b) binary information divergence cost

barrier prevents coherent approximations of uncertain assessments taking on the form of certainties, as happens in the case of the quadratic objective.

Unfortunately, even with $\chi = I$, the analysis for this cost function is difficult. We can, however, derive a simple lower bound on the cost. Let

$$C_1(P,Q) = \sum_{i=1}^{N} P_i \log\left(\frac{P_i}{Q_i}\right)$$

and

$$C_2(P,Q) = \sum_{i=1}^{N} (1-P_i) \log\left(\frac{1-P_i}{1-Q_i}\right).$$

Then $C(P,Q) = C_1(P,Q) + C_2(P,Q)$ and therefore the cost of solving Equation 3.8 under $C(P,Q)$ is no smaller than the cost of solving Equation 3.8 under $C_1(P,Q)$ plus the cost of solving under $C_2(P,Q)$.

Substituting $C_1(P,Q) = \sum_{i=1}^{N} P_i \log\left(\frac{P_i}{Q_i}\right)$ into Equation 3.8 we have

$$\begin{aligned} Q_1^* &= \underset{\{Q \mid \sum_i Q_i = 1,\ Q_i \geq 0\}}{\arg\min} \sum_{i=1}^{N} P_i \log\left(\frac{P_i}{Q_i}\right) \\ &= \underset{\{Q \mid \sum_i Q_i = 1,\ Q_i \geq 0\}}{\arg\min} \sum_{i=1}^{N} P_i \left(\log P_i - \log Q_i\right) \\ &= \underset{\{Q \mid \sum_i Q_i = 1,\ Q_i \geq 0\}}{\arg\max} \sum_{i=1}^{N} P_i \log\left(Q_i\right) \end{aligned}$$

Applying Lagrangian analysis results in the unconstrained optimization

$$Q_1^* = \arg\max_{Q \geq 0} \sum_{i=1}^{N} P_i \log(Q_i) + \lambda \left(1 - \sum_i Q_i\right)$$

Taking the derivative with respect to $Q_i$ results in $Q_i = \frac{P_i}{\lambda}$ which immediately leads to the solution $Q^* = \frac{P}{\sum_i P_i}$; i.e. the optimal solution under $C_1(P,Q)$ is to scale the assessments until they form a probability mass function. This is the "natural" solution suggested at the end of Section 3.3.2.

To derive the other half of the lower bound, we substitute $C_2(P,Q)$ into Equation 3.8. Neglecting the range constraints on $Q$ would give an answer similar to that for $C_1(P,Q)$, i.e. $1 - Q = \frac{1-P}{\sum_i 1-P_i}$. However, unlike for $C_1(P,Q)$, the assumption that $P_i \in (0,1)\ \forall i$ does not in general imply the feasibility of this solution. Since we are only seeking to establish a lower bound, however, the unconstrained minimization is sufficient, and we can state a lower bound on the cost of Equation 3.8 using $C(P,Q) = \sum_{i=1}^{N} D_b(P||Q)$.

**Theorem 3.1.** *$C(P,P^*)$ where $P^*$ is the solution to Equation 3.8 is bounded from below by the relation*

$$C(P,P^*) \geq C_1\left(P, \frac{P}{\sum_i P_i}\right) + C_2\left(P, 1 - \frac{1-P}{\sum_i 1 - P_i}\right)$$

The derivation of the bound above gives insight into why the natural inclination to normalize the assessments isn't the right thing to do. Each expert's assessment $P_i$ specifies not only the subjective probability of event $A_i$, but also the subjective probability of event $A_i^c$. Taking $Q = \frac{P}{\sum_i P_i}$ neglects the implicit assessed probabilities of the complementary events. In using $C(P,Q) = \sum_{i=1}^{N} D_b(P_i||Q_i)$ as the cost function for Equation 3.8, the cost of normalizing the assessments is being balanced against the cost of normalizing their complements.

Another way of viewing this result is to disregard the implicit assessments of the complementary events. After all, the whole problem of incoherence occurs because experts' implicit assessments are incorrect; levying costs associated with implicit assessments seems unfair, given that we already known those implicit assessments are illogical. This perspective would suggest that the proper cost function wouldn't attempt to balance the costs of deforming explicit and implicit assessments, but that only $C_1(P,Q)$ (costs for explicit assessments) should be considered at all and that $C_2(P,Q)$ should receive no weight.

Expanding on this perspective, taking $C(P,Q)$ to be the generalized I-Divergence $C(P,Q) = \sum_i P_i \log\left(\frac{P_i}{Q_i}\right) - \sum_i (P_i - Q_i)$ would result in the intuitive scaling solution to Equation 3.8. It can also be shown (by a slight generalization of the analysis in [99]) that when $\chi_i(\omega) = \chi_1(\omega) \forall i$ (i.e. when all experts are assessing the same event) that the generalized I-Divergence cost function again results in the "intuitive" linear averaging mechanism for fusing probability assessments.

## 3.4 Alternative Formulation of the CAP

In Section 3.3 an argument was made that caution should be used in determining the proper objective function of the CAP. It was shown that the $L_2$ norm suggested in [76, 77] results in counter-intuitive results when applied to the estimation of probability mass functions. An alternative objective, based on the sum of binary divergences, was suggested and justified using a particular assessor model, as well as an empirical argument.

In this section we go further than simply considering alternative objective functions and consider the fundamental question of whether the formulation of Equation 3.1 is well-justified. We suggest an alternative formulation, based in information geometry, and compare it both to Equation 3.1, to the Divergence-based objective function formulated in Section 3.3.2 and to maximum entropy methods. We demonstrate that the alternative formulation is more flexible and more consistent with the underlying nature of assessment.

### 3.4.1 Criticisms of the CAP

In [78] the question of the proper scoring function for coherent approximation was considered. In that development the question was what posterior scoring functions would induce assessors to provide honest assessments. However no distinction was attempted among the family of proper scoring functions as to which led to the "best" (in some sense) approximation.

One criticism of the optimization formulation in 3.1 is that it treats assessments as points in Euclidean space. There's nothing mathematically troubling about doing so, but pragmatically it leads to results that don't match natural interpretations of probability. On a philosophical level, it ignores the fundamental formulation of an assessment as a subjective expectation of a random variable. We will discuss both the pragmatic and the philosophical objections to the formulation of 3.1 in greater depth below.

In an attempt to reformulate the CAP to better align both with natural interpretations of assessments as well as the fundamentals of subjective probability theory, consider the information provided by an assessment when viewed as a subjective expectation of a random variable. Specifically, this expectation gives us some information about the assessor's subjective probability distribution over the atoms of $\Omega$. Thus the question of coherently approximating a distributed assessment becomes a question of choosing a distribution over the atoms of $\Omega$ that best agrees with what is known about the subjective probability distributions of the set of experts. We formalize this information geometric view of coherent approximation in Section 3.4.2 below. As was alluded to in the introduction to this section, there are obvious deficiencies to the formulation of the CAP as a Euclidean projection. In this section we highlight two such criticisms, one based in natural interpretations of probability and one based in the philosophy of subjective probability.

### 3.4.2 An Information Geometric Reformulation of the CAP

Given the concerns with the formulation of the CAP in Equation 3.1 we suggest here a new formulation. The central idea of this reformulation is that an assessment is, paraphrasing de Finetti, a distribution among all possible options one's own "sensations of probability". Thus each assessment should be treated as a statement of subjective expectation. The key question in coherent approximation is what personal probabilities led to the set of assessments, and how to best approximate *those* by a single probability distribution over the atoms of $\Omega$. The formulation will draw extensively on basic concepts from the theory of information geometry [100, 101] such as the I-divergence and linear and exponential families.

Given that an assessment is a subjective expectation the natural question is what subjective probability distribution generated the assessment. An assessment alone is insufficient to define a unique probability distribution, but for each assessment we can consider the *family* of distributions which might have led to the assessment.

**Definition 3.1.** *For a given constant $\alpha$ and random variable $X$, the **linear family** is defined as $L_\alpha(X) \triangleq \{Q : \mathbb{E}_Q[X] = \alpha\}$.*

Thus each expert's assessment defines a linear family of probability distributions.

**Proposition 3.2.** *If $\bigcap_i L_{P_i}(X_i) \neq \emptyset$ then the assessment $P$ is coherent*

Let $Q$ be a distribution over the atoms of $\Omega$ and let $Q \in \bigcap_i L_{P_i}(X_i)$. Thus there exists $\lambda = Q$ s.t. $P = \mathbb{E}_\lambda[X] = \mathbf{X}\lambda$ and therefore $P$ is coherent. $\square$

This view of assessments as linear families leads us to formulate the following Information Geometric Coherent Approximation Principle (IGCAP).

$$\lambda^* = \arg\min_{Q \in \Delta} \sum_{i=1}^{N} \min_{\pi \in L_{P_i}(X_i)} D(\pi || Q) \tag{3.9}$$

with $P^* = \mathbf{X}\lambda^*$. The IGCAP is depicted graphically in Figure 3-3

### 3.4.3 IGCAP Solution as an MAP estimate

In this section an assessor model is presented to justify the form chosen for the IGCAP.

Suppose there exists a probability distribution $\mu$ over the atoms of $\Omega$, and that each assessor $i$ has received $n_i$ samples drawn i.i.d. from $\mu$. From these samples, assessor $i$ forms an empirical distribution $\hat{p}^i_{n_i}$. In providing an assessment of random variable $X_i$, assessor $i$ follows his type, meaning

$$P_i = \mathbb{E}_{\hat{p}^i_{n_i}}[X_i] \tag{3.10}$$

Let $L^i \triangleq \{Q | \mathbb{E}_Q[X_i] = P_i\}$. Given assessment $P$, we wish to select a $\hat{\mu}$ from some arbitrary, but finite, set of distributions over $\Omega$. Let $\mathcal{M}$ be the set of such distributions. One reasonable method for doing so would be to choose the MAP estimate. If we

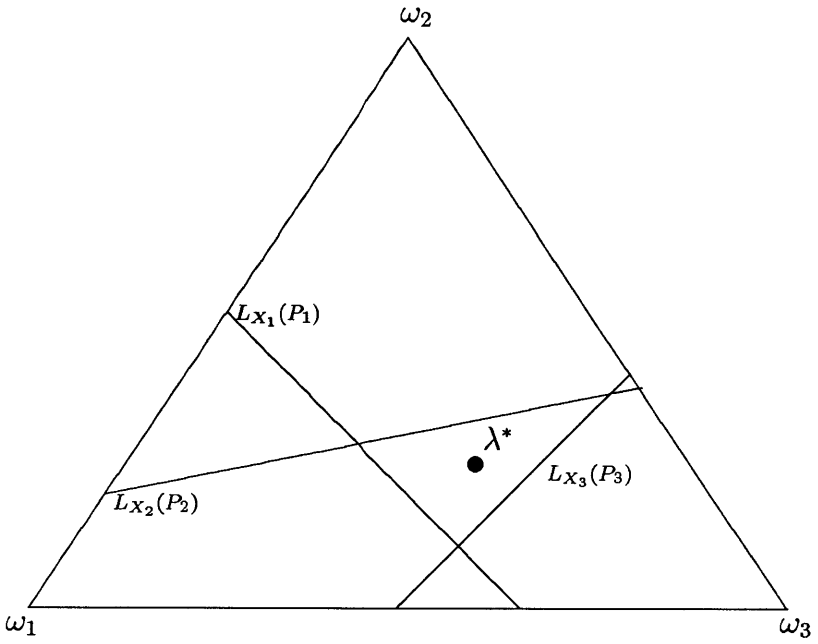



**Figure 3-3.** Graphical view of IGCAP

assume a uniform distribution over $\mathcal{M}$, the MAP estimator is equivalent to the ML estimator, i.e.

$$\begin{aligned}
\hat{\mu}_{MAP} &= \arg\max_{\nu \in \mathcal{M}} P(\nu | \forall i, \hat{p}^i_{n_i} \in L^i) \\
&= \arg\max_{\nu \in \mathcal{M}} \frac{P_\nu(\forall i, \hat{p}^i_{n_i} \in L^i)\frac{1}{|\mathcal{M}|}}{\sum_{\nu \in \mathcal{M}} P_\nu(\forall i, \hat{p}^i_{n_i} \in L^i)\frac{1}{|\mathcal{M}|}} \\
&= \arg\max_{\nu \in \mathcal{M}} P_\nu(\forall i, \hat{p}^i_{n_i} \in L^i) = \hat{\mu}_{ML}
\end{aligned}$$

We make the following assumption about our assessors:

**Assumption 3.1.** *Each assessor's sequence of observations is independent of every other assessors observations.*

Given this assumption, we can write the maximum likelihood estimate of $\mu$ as

$$\hat{\mu}_{ML} = \arg\max_{\nu \in \mathcal{M}} P_\nu(\forall i, \hat{p}^i_{n_i} \in L^i) = \arg\max_{\nu \in \mathcal{M}} \prod_i P_\nu(\hat{p}^i_{n_i} \in L^i) \tag{3.11}$$

We next use a 'method of types' argument to demonstrate that for $n_i$ sufficiently

large the ML estimate $\hat{\mu}_{ML}$ is given by

$$\hat{\mu}_{ML} = \min_{Q \in \mathcal{M}} \sum_{i=1}^{M} n_i D(\pi_i^* || Q) \tag{3.12}$$

i.e. a weighted form of the IGCAP. For this argument we will need the following fact

**Fact 3.1.** *Let $\hat{p}_n$ be the empirical distribution, or 'type', of an observation sequence of length $n$ and let $\mathcal{L}_n$ be the set of all such $n$-types. Then for any set of probability vectors $\Gamma \subseteq \Delta(\Omega)$ we have*

$$(n+1)^{-|\Omega|} e^{-n \inf_{\nu \in \Gamma \cap \mathcal{L}_n} D(\nu||\mu)} \leq P_\mu(\hat{p}_n \in \Gamma) \leq (n+1)^{|\Omega|} e^{-n \inf_{\nu \in \Gamma \cap \mathcal{L}_n} D(\nu||\mu)}$$

For proof, see Chapter 2 of [102]. Letting

$$C(n) = (\max_i(n_i) + 1)^{M|\Omega|} \tag{3.13}$$

we can apply these bounds to the argument of the maximization in Equation 3.12.

$$\begin{aligned} \prod_i P_\nu(\hat{p}^i_{n_i} \in L^i) &\geq \frac{1}{C(n)} e^{-\sum_i n_i \inf_{\pi \in L^i \cap \mathcal{L}_{n_i}} D(\pi||\nu)} \\ \prod_i P_\nu(\hat{p}^i_{n_i} \in L^i) &\leq C(n) e^{-\sum_i n_i \inf_{\pi \in L^i \cap \mathcal{L}_{n_i}} D(\pi||\nu)} \end{aligned}$$

Since $L^i$ is a linear family, the inf in the exponent of the above bounds will be achieved. We introduce another well-known fact.

**Fact 3.2.** *Given a linear family $L$ and some $\pi \in L$, for any $Q \notin L$,*

$$D(\pi||Q) = D(\pi||\pi^*) + D(\pi^*||Q)$$

*where $\pi^* = \arg\min_{\pi \in L} D(\pi||Q)$ (i.e. the I-projection).*

Using this fact we can rewrite the above bounds as

$$\begin{aligned} \prod_i P_\nu(\hat{p}^i_{n_i} \in L^i) &\geq \frac{1}{C(n)} e^{-\sum_i n_i \left( \min_{\pi \in L^i \cap \mathcal{L}_{n_i}} D(\pi||\pi_i^*) + D(\pi_i^*||\nu) \right)} \\ &= \frac{1}{C(n)} e^{-\sum_i n_i \min_{\pi \in L^i \cap \mathcal{L}_{n_i}} D(\pi||\pi_i^*)} e^{-\sum_i n_i D(\pi_i^*||\nu)} \\ \prod_i P_\nu(\hat{p}^i_{n_i} \in L^i) &\leq C(n) e^{-\sum_i n_i \left( \min_{\pi \in L^i \cap \mathcal{L}_{n_i}} D(\pi||\pi_i^*) + D(\pi_i^*||\nu) \right)} \\ &= C(n) e^{-\sum_i n_i \min_{\pi \in L^i \cap \mathcal{L}_{n_i}} D(\pi||\pi_i^*)} e^{-\sum_i n_i D(\pi_i^*||\nu)} \end{aligned}$$

As $n_i \to \infty$, $\min_{\pi \in L^i \cap \mathcal{L}_n} D(\pi||\pi_i^*) \to 0$ at a rate of approximately $\frac{1}{n_i}$. Therefore, for notational ease, we will assume $n_i$ is sufficiently large that the impact of the empirical

distribution quantization factor $e^{-\sum_i n_i \min_{\pi \in L^i \cap \mathcal{L}_{n_i}} D(\pi||\pi_i^*)}$ is negligible on the above bounds and drop it from the rest of the analysis.

Let $\nu_1 = \arg\min_{\nu \in \mathcal{M}} \sum_i n_i D(\pi_i^*||\nu)$ and consider the difference in log-likelihood, i.e. $\sum_i \log P_{\nu_1}(\hat{p}_{n_i}^i \in L^i) - \sum_i \log P_{\nu_2}(\hat{p}_{n_i}^i \in L^i)$ for some $\nu_2 \neq \nu_1$. Using the above bounds (*sans* quantization factor), we have

$$\begin{aligned}
&\sum_i \log P_{\nu_1}(\hat{p}_{n_i}^i \in L^i) - \sum_i \log P_{\nu_2}(\hat{p}_{n_i}^i \in L^i) \\
&\qquad \geq -\log C(n) - \sum_i n_i D(\pi_{i_1}^*||\nu_1) - \log C(n) + \sum_i n_i D(\pi_{i_2}^*||\nu_2) \\
&\qquad = -2\log C(n) - \sum_i n_i \left( D(\pi_{i_1}^*||\nu_1) - D(\pi_{i_2}^*||\nu_2) \right)
\end{aligned}$$

By the definition of $\nu_1$ the factor $\left( D(\pi_{i_1}^*||\nu_1) - D(\pi_{i_2}^*||\nu_2) \right) < 0$ and since $-\log C(n)$ is going to $-\infty$ logarithmically in $\max_i(n_i)$, we have that for all $n_i$ sufficiently large

$$-2\log C(n) - \sum_i n_i \left( D(\pi_{i_1}^*||\nu_1) - D(\pi_{i_2}^*||\nu_2) \right) > 0$$

and therefore $\hat{\mu}_{ML} = \nu_1$.

### ■ 3.4.4 Alternative Minimax Cost

Notice in Equation 3.12 above that it depends on the weighting factors $n_i$, which are essentially the amount of experience of assessor $i$. In the coherent approximation problem as formulated we don't have access to such reputational information. In the absence of it, we make the assumption that $n_i = n_j = n$ for all $i, j$. Under this assumption we retrieve the formula for the IGCAP in Equation 3.9.

A reasonable alternative assumption in the absence of knowledge of $n_i$ is that we are engaged in an adversarial game with nature. In this case, after choosing an estimate, nature chooses values for $n_i$ to maximize our expected estimation error. This assumption leads to the related formulation

$$Q_{mmx}^* = \arg\min_{Q \in \mathcal{M}} \max_i D(\pi_i^*||Q)$$

This formulation will choose $Q$ to be 'equidistant' (in the divergence sense) from all the linear families. Possible implications of this alternative cost formulation will be discussed somewhat further in Section 4.9.2. One limitation of the minimax cost structure is that it is more sensitive to outlier assessments than is the sum of divergences formulation. If almost all assessors are coherent with respect to each other, but a single assessor has a highly divergent assessment, the minimax cost structure will tend to give the outlier assessment outsized weight based on the level of agreement of all the other assessors.

### ■ 3.4.5 Application of the Conditional Limit Theorem

In the sequel, we will at times interpret the distributions $\pi_i^*$ which are the I-projections of $\mu = Q^*$ onto $L^i$, as estimates of the assessors' 'true' subjective distributions. Using the above assessment model, we demonstrate the reasonableness of that assumption. What follows is a special case of a conditional limit theorem; significantly more general treatments can be found in [103–105] and the references therein.

As per the assessor model suggested in Section 3.4.3, suppose the assessors are formulating types $\hat{p}_n^i$ based on independent streams of observations, generated according to distribution $\mu$. We will demonstrate that, for any $\epsilon > 0$, for any $i$, $P_\mu(D(\hat{p}_{n_i}^i||\pi_i^*) > \epsilon|\hat{p}_n^i \in L^i)$ can be made arbitrarily small for $n_i$ sufficiently large. Since we will be deriving the result for each $i$, we will drop the assessor identification in what follows (i.e. $n_i$ will be denoted $n$, $\hat{p}_{n_i}^i$ will be $\hat{p}_n$ and so forth).

Define $D_\epsilon \triangleq \{Q \in L|D(Q||\pi^*) < \epsilon\}$ and let $D_\epsilon^c = L \backslash D_\epsilon$. Letting $C(n) = (n+1)^{|\Omega|}$, $\nu_n^* = \arg\min_{\nu \in D_\epsilon \cap \mathcal{L}_n} D(\nu||\pi^*)$, and $\nu_n^{*c} = \arg\min_{\nu \in D_\epsilon^c \cap \mathcal{L}_n} D(\nu||\pi^*)$, we can write

$$\begin{aligned} P_\mu(\hat{p}_n \in D_\epsilon^c|\hat{p}_n \in L) &\leq C(n)e^{-n\inf_{\nu \in D_\epsilon^c \cap \mathcal{L}_n} D(\nu||\mu)} \\ &= C(n)e^{-n(D(\nu_n^{c*}||\pi^*)+D(\pi^*||\mu))} \end{aligned} \tag{3.14}$$

and

$$\begin{aligned} P_\mu(\hat{p}_n \in D_\epsilon|\hat{p}_n \in L) &\geq \frac{1}{C(n)}e^{-n\inf_{\nu \in D_\epsilon \cap \mathcal{L}_n} D(\nu||\mu)} \\ &= \frac{1}{C(n)}e^{-n(D(\nu_n^*||\pi^*)+D(\pi^*||\mu))} \end{aligned} \tag{3.15}$$

using Facts 3.1 and 3.2 from Section 3.4.3.

Since $L = D_\epsilon \cup D_\epsilon^c$ we can write the conditional probability as

$$\begin{aligned} P_\mu(D(\hat{p}_n||\pi^*) < \epsilon|\hat{p}_n \in L) &= P_\mu(\hat{p}_n \in D_\epsilon|\hat{p}_n \in L) \\ &= \frac{P_\mu(\hat{p}_n \in D_\epsilon)}{P_\mu(\hat{p}_n \in D_\epsilon) + P_\mu(\hat{p}_n \in D_\epsilon^c)} \\ &= \frac{1}{1 + \frac{P_\mu(\hat{p}_n \in D_\epsilon^c)}{P_\mu(\hat{p}_n \in D_\epsilon)}} \\ &\geq \frac{1}{1 + \frac{C(n)e^{-n(D(\nu_n^{c*}||\pi^*)+D(\pi^*||\mu))}}{\frac{1}{C(n)}e^{-n(D(\nu_n^*||\pi^*)+D(\pi^*||\mu))}}} \\ &= \frac{1}{1 + 2C(n)e^{-n(D(\nu_n^{c*}||\pi^*)+D(\pi^*||\mu)-D(\nu_n^*||\pi^*)-D(\pi^*||\mu))}} \\ &= \frac{1}{1 + 2C(n)e^{-n(D(\nu_n^{c*}||\pi^*)-D(\nu_n^*||\pi^*))}} \end{aligned}$$

where the bound is due to Equation 3.15 and Equation 3.14. By definition, we have that $D(\nu_n^{c*}||\pi^*) > \epsilon$ (in fact, $D(\nu_n^{c*}||\pi^*) \to \epsilon$ at rate $O(\frac{1}{n})$). The factor $-D(\nu_n^*||\pi^*) \to 0$, again at rate $O(\frac{1}{n})$. Therefore, for all $n$ sufficiently large, $D(\nu_n^{c*}||\pi^*) - D(\nu_n^*||\pi^*)) >$

0 and

$$P_\mu(\hat{p}_n \in D_\epsilon | \hat{p}_n \in L) \to 1$$

at an exponential rate of $\epsilon$.

Interpreting this result, we see that the probability mass for $\hat{p}_n$ conditioned on $\hat{p}_n \in L$ is concentrating in an arbitrarily small neighborhood around the point $\pi^*$. This justifies, to some extent, the interpretation of $\pi^*$ as an approximation of the assessor's subjective probability distribution $\hat{p}_n$.

One must be a bit cautious in interpreting this result, however. The statement applies to the *marginal* probability that each assessor's type is "close" to $\pi_i^*$; it doesn't imply that the *joint* probability that all types lie within an $\epsilon$-ball of their respective $\pi_i^*$.

## ■ 3.5 Properties of the IGCAP

In this section we analyze the IGCAP and demonstrate some of its properties and benefits, including:

1. The solution $P^*$ is (a) unique and (b) equal to $P$ if $P$ is coherent
2. The computation is particularly tractable for characteristic random variables, with a special case that agrees with the operator model suggested in Section 3.3.2
3. The mechanism can extend invariantly to assessments of non-characteristic random variables
4. The formulation can be expanded to allow for assessments given as ranges rather than points
5. Given a sequence of assessments $P_n$ such that $\lim_{n\to\infty} P_n = \bar{P}$ is coherent, the sequence of solutions $P_n^* \to \bar{P}$.

Each of these benefits will be discussed in greater detail in subsections below. An additional benefit is the easy extensibility of the theory to the case of conditional assessments (i.e. assessments of conditional random variables), but a discussion of that extension will be left until Chapter 4

### ■ 3.5.1 Solution Uniqueness and Coherence

**Proposition 3.3.** *If $P$ is incoherent then there exists a unique solution to Equation 3.9.*

Because each of the divergences under the summation are strictly convex in $Q$, the sum is strictly convex and therefore there is a unique solution. Furthermore the solution will lie in the interior of the probability simplex. □

**Proposition 3.4.** *If $P$ is coherent then $P = P^*$ where $P^* = \mathbf{X}\lambda^*$ and $\lambda^*$ satisfies Equation 3.9.*

For any $Q \in \cap_i L_{P_i}(X_i)$ will have (a)$\min_{\pi \in L_{P_i}}(X_i) D(\pi||Q) = 0$ for all $i$ since $Q \in L_{P_i}(X_i)$ and (b)$\mathbb{E}_Q[X] = P$ (by definition). Since divergence (for finite distributions) is non-negative, the minimum of Equation 3.9 is obtained for any $Q \in \cap_i L_{P_i}(X_i)$ (and, by strict convexity of the divergence, only for such $Q$). Thus $\bigcap_i L_{P_i}(X_i)$ is the set of all minimizers $\lambda^*$ of Equation 3.9 and by (b), $P^* = \mathbf{X}\lambda^* = \mathbb{E}_{\lambda^*}[X] = \mathbb{E}_Q[X] = P$.□

### ■ 3.5.2 Computation of IGCAP

One concern about moving from the CAP formulation to the IGCAP is computation. As stated earlier, the CAP has computational challenges, since it is a convex program with a potentially exponential (in the number of random variables) number of linear constraints. Some work has been done [78,98] to examine suboptimal approximations that are more readily computable. On it's face, the IGCAP seems likely to significantly increase the computational challenge of coherent approximation, as we now must solve a nested set of optimization problems. However, due to the special structure of the inner optimization the computational problem is not as great as might at first be feared.

The inner minimization problem in Equation 3.9 is commonly referred to as the I-projection. It is known [100] that, for a given $Q$, the I-projection of $Q$ onto a linear family defined by the statistic $X_i$ (i.e. $L_{P_i}(X_i)$) lies on the tilted exponential family of $Q$ defined by $\mathcal{Q}_i = \{Q' | Q' = \frac{Q}{Z(\theta)} e^{\theta x_i}\}$ where $Z(\theta)$ is commonly referred to as the partition function and $\theta$ as the natural parameter. It can be further shown that there is a unique point $P_i^* = L_{P_i}(X_i) \cap \mathcal{Q}_i$ that optimally solves the I-projection.

In the case that $X_i$ is a characteristic random variable, we can simplify Equation 3.9. Letting $\pi_i^*$ denote the solution to the I-projection for a given $Q$ we can write the elements of $\pi_i^*$ as

$$[\pi_i^*]_j = \begin{cases} \frac{Q_j}{Z_i} e^{\theta_i} & [x_i]_j = 1 \\ \frac{Q_j}{Z_i} & [x_i]_j = 0 \end{cases}$$

Since $\sum_j [\pi_i^*]_j = 1$ and $\mathbb{E}_{\pi_i^*}[X_i] = P_i$ we have

$$\begin{aligned} 1 - P_i &= \sum_j [\pi_i^*]_j - \mathbb{E}_{\pi_i^*}[X_i] = \sum_j [\pi_i^*]_j - \sum_{\{j|[x_i]_j=1\}} [\pi_i^*]_j \\ &= \sum_{\{j|[x_i]_j=0\}} [\pi_i^*]_j = \sum_{\{j|[x_i]_j=0\}} \frac{Q_j}{Z_i} \end{aligned}$$

Therefore the partition function for the optimizing distribution is $Z_i = \frac{\sum_{\{j|[x_i]_j=0\}} Q_j}{1-P_i}$.

Next, given this representation of $Z_i$ as a function of $Q$ and $x_i$, we can solve for the natural parameter of the optimizing distribution.

$$P_i = \mathbb{E}_{\pi_i^*}[X_i] = \sum_{\{j|[x_i]_j=1\}} [\pi_i^*]_j = \sum_{\{j|[x_i]_j=1\}} \frac{Q_j}{Z_i} e^{\theta_i}$$

and therefore

$$\theta_i = \log\left(\frac{P_i Z_i}{\sum_{\{j|[x_i]_j=1\}} Q_j}\right) = \log\left(\frac{P_i\left(\sum_{\{j|[x_i]_j=0\}} Q_j\right)}{(1-P_i)\left(\sum_{\{j|[x_i]_j=1\}} Q_j\right)}\right)$$

Therefore, we can write the inner minimization factor

$$\begin{aligned}
D(\pi_i^*||Q) &= \sum_j [\pi_i^*]_j \log \frac{[\pi_i^*]_j}{Q_j} \\
&= \sum_j \frac{Q_j}{Z_i} e^{\theta_i [x_i]_j} \log\left(\frac{\frac{Q_j}{Z_i} e^{\theta_i [x_i]_j}}{Q_j}\right) \\
&= \sum_{\{j|[x_i]_j=0\}} \frac{Q_j}{Z_i} \log \frac{1}{Z_i} + \sum_{\{j|[x_i]_j=1\}} \frac{Q_j P_i}{\sum_{\{j|[x_i]_j=1\}} Q_j} \log \frac{P_i}{\sum_{\{j|[x_i]_j=1\}} Q_j} \\
&= \frac{1}{Z_i} \log \frac{1}{Z_i} \sum_{\{j|[x_i]_j=0\}} Q_j + \frac{P_i}{\sum_{\{j|[x_i]_j=1\}} Q_j} \log \frac{P_i}{\sum_{\{j|[x_i]_j=1\}} Q_j} \sum_{\{j|[x_i]_j=1\}} Q_j \\
&= (1-P_i) \log \frac{1-P_i}{\sum_{\{j|[x_i]_j=0\}} Q_j} + P_i \log \frac{P_i}{\sum_{\{j|[x_i]_j=1\}} Q_j}
\end{aligned}$$

The IGCAP for characteristic random variables is therefore equivalent to the problem

$$\lambda^* = \arg\min_{Q \in \Delta} \sum_{i=1}^{N} (1-P_i) \log \frac{1-P_i}{\sum_{\{j|[x_i]_j=0\}} Q_j} + P_i \log \frac{P_i}{\sum_{\{j|[x_i]_j=1\}} Q_j} \tag{3.16}$$

### ■ 3.5.3 Comparisons of IGCAP to Other Coherent Approximation Formulations

In this subsection we compare the suggested method of coherent approximation to other suggested methods from the literature. We show specific cases in which each alternative formulation is deficient, and include an empirical experiment demonstrating the superiority of the IGCAP.

#### The CAP

The CAP, as formulated in [76, 77], is a Euclidean projection onto a convex set. However, for the purposes of comparing it to the IGCAP we introduce an equivalent view of the CAP.

Consider two linear families $L_{\alpha_1}(X)$ and $L_{\alpha_2}(X)$, and assume they are non-empty (i.e. $\alpha_1$ and $\alpha_2$ are valid expectations of $X$; this comports with the assumption, made in Section 2.3.2, that $P(X) \in [\min_i x_i, \max x_i]$ where $x_i$ is a realization of the random variable $X$).

**Proposition 3.5.** *$L_{\alpha_1}(X)$ and $L_{\alpha_2}(X)$ are parallel to each other on the simplex (in the standard, Euclidean way).*

The insight from Proposition 3.5 provides us a method from comparing the CAP with the IGCAP. Specifically, consider this equivalent formulation of the CAP:

$$\min_{\tilde{P}\in[0,1]^M} \sum_{i=1}^{M} d(L_{P_i}(X), L_{\tilde{P}_i}(X))$$
$$\text{s.t. } \cap_i L_{\tilde{P}_i}(X) \neq \emptyset \tag{3.17}$$

where $d(A, B)$ is the minimum Euclidean distance between sets $A$ and $B$. Viewed in information space, the CAP formulation in Equation 3.1 is equivalent to a minimal shift among the set of linear families until there is a non-empty intersection. However, the shift is performed along level sets of the linear families.

Other limitations of the CAP as formulated in Equation 3.1 will be suggested in Section 3.5.4.

#### Divergence-Based CAP

We now compare the IGCAP to the reformulation of the CAP to use a binary divergence-based metric in Section 3.3, and find that they coincide in a particular important case.

Suppose that $\mathbf{X} = I$ (hence $M = N$), per the empirical argument in Section 3.3.2. Since in this case the random variables are characteristic, Equation 3.16 applies in this scenario. Note that for any given $i$, $\sum_{\{j|[x_i]_j=1\}} Q_j = Q_i$ and $\sum_{\{j|[x_i]_j=0\}} Q_j = 1 - Q_i$. Furthermore $P^* = \chi\lambda^* = I\lambda^* = \lambda^*$. Therefore we can rewrite Equation 3.16 as

$$P^* = \arg\min_{Q\in\Delta} \sum_{i=1}^{N} (1 - P_i) \log \frac{1 - P_i}{1 - Q_i} + P_i \log \frac{P_i}{Q_i}$$

This is exactly the sum of binary divergences that was asserted as the proper objective function for the CAP in Section 3.3 based on an independent justification.

#### Maximum Entropy methods

Maximum entropy methods have been successfully applied to a diverse set of estimation problems. A general statement of the principle of maximum entropy is that when faced with a set of probability distributions with no prior information on which to select a single distribution, the distribution with maximum entropy should be chosen. This principle goes back to Jaynes [106], but has been employed by a great many researchers working on a diverse set of problems.

A possible alternative method for forming a coherent approximation, would be to select the maximum entropy distribution from each linear family $L_{P_i}(X_i)$ and then average (either linearly or log-linearly) the resulting distributions over the atoms of $\Omega$. While this method may often generate a good approximation, there are certain instances in which it may be quite bad.

Consider for example the situation when $\chi = I$, as in Section 3.3.2 and $P(X) = \begin{bmatrix} 0.8 & 0.1 & 0.14 \end{bmatrix}^T$. This is example is shown graphically in Figure 3-4. The three

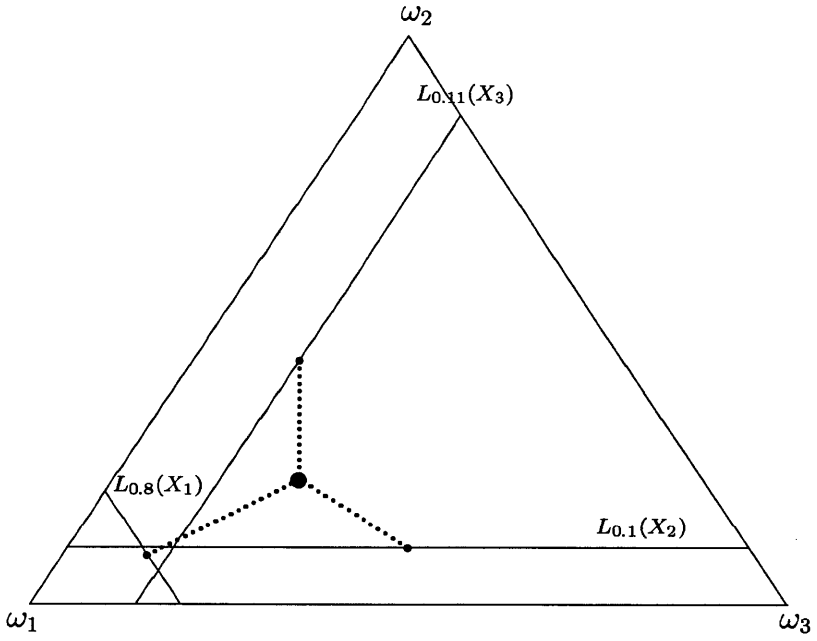



**Figure 3-4.** Failure of maximum entropy approximation

linear families, generated by three independent assessments, are shown in the figure. While their intersection is empty, the families seem to be "nearly" coherent (in the sense that a small perturbation of any one of the families would result in a non-empty intersection, and thus coherence).

However, the maximum entropy principle, if applied as suggested, would choose from each linear family its maximum entropy member. In this case, that means the point in each family that lies closest (in a divergence sense) to the middle of the simplex. Taking a linear average of those points will lead, in this example, to a coherent approximation that lies quite far away from the point at which the linear families nearly intersect. Mathematically, we define the maximum entropy approximation as $P^*_{ME} = \mathbf{X}\lambda^*_{ME}$ where

$$\lambda^*_{ME} = \frac{1}{M}\sum_i \underset{Q \in L_{P_i}(X_i)}{\arg\max}\, H(Q)$$

with $H(Q) = -\sum_j Q_j \log Q_j$. For the simple example above where $\mathbf{X} = I$, this results in

$$\begin{aligned} P^*_{ME} &= \frac{1}{3}\left(\underset{Q \in L_{0.8}(X_1)}{\arg\max}\, H(Q) + \underset{Q \in L_{0.1}(X_2)}{\arg\max}\, H(Q) + \underset{Q \in L_{0.14}(X_3)}{\arg\max}\, H(Q)\right) \\ &= \frac{1}{3}\left(\begin{bmatrix} 0.8 \\ 0.1 \\ 0.1 \end{bmatrix} + \begin{bmatrix} 0.45 \\ 0.1 \\ 0.45 \end{bmatrix} + \begin{bmatrix} 0.43 \\ 0.43 \\ 0.14 \end{bmatrix}\right) \\ &= \begin{bmatrix} 0.56 \\ 0.21 \\ 0.23 \end{bmatrix} \end{aligned}$$

Particularly troubling is that the method would be discontinuous in the limit. So, taking the above example, if we consider a sequence of assessments

$$P_n(X) = \begin{bmatrix} 0.8 \\ 0.1 \\ 0.1 + \frac{1}{n} \end{bmatrix}$$

we see that the sequence is 'approaching' coherence (in the sense that the limit is coherent). However, considering the limit of the maximum entropy approximations is

$$\lim_{n\to\infty} P^*_{ME_n} = \lim_{n\to\infty} \frac{1}{3}\begin{bmatrix} 1.7 - \frac{n}{2} \\ 0.65 - \frac{n}{2} \\ 0.65 + n \end{bmatrix} = \begin{bmatrix} 0.5\bar{6} \\ 0.2\bar{6} \\ 0.2\bar{6} \end{bmatrix} \neq \begin{bmatrix} 0.8 \\ 0.1 \\ 0.1 \end{bmatrix}$$

### ■ 3.5.4 Extension to Non-Characteristic Random Variables

The focus thus far has been predominantly on coherent assessments of characteristic random variables (i.e. probabilities). In de Finetti [26], probabilities are a special case

of the more general principle of prevision. As shown in Chapter 2, the philosophical justifications for coherent assessment are very similar between characteristic and non-characteristic random variables.

Because it is formulated in terms of subjective expectations, the IGCAP generalizes immediately to assessments of non-characteristic random variables. The notation of linear families $L_{P_i}(X_i)$ is identical, as is the interpretation. For a given assessment, the coherent approximation is given by $P^* = \mathbf{X}\lambda^*$ where $\lambda^*$ is defined in Equation 3.9. The only notational change is the replacement of the characteristic matrix $\chi$ with the more general outcome matrix $\mathbf{X}$.

One attractive property of the generalized IGCAP is that it is invariant under linear transforms of the random variables. The algorithm thus returns identical answers (appropriately scaled), regardless of the specific units in which experts choose to make assessments. Mathematically, let $\tilde{X}_i = a_i X_i + b_i$ and $\tilde{P}_i = a_i P_i + b_i$ $(a_i \neq 0)$. Then for $\tilde{P}^*$ calculated by the generalized IGCAP for assessment $\tilde{P}$ of random vector $\tilde{X}$ we have $\tilde{P}_i^* = a_i P_i^* + b_i$ where $P^*$ is the coherent approximation of assessment $P$ of random vector $X$. We state this fact in the following Proposition.

**Proposition 3.6.** *The IGCAP is invariant under affine random variable transformations.*

The critical point is that $L_{\tilde{P}_i}(\tilde{X}_i) = L_{P_i}(X_i)$. To see this, consider that

$$\begin{aligned} L_{\tilde{P}_i}(\tilde{X}_i) &= \{Q : E_Q[\tilde{X}_i] = \tilde{P}_i\} \\ &= \{Q : E_Q[a_i X_i + b_i] = \tilde{P}_i\} \\ &= \{Q : a_i E_Q[X_i] + b_i = \tilde{P}_i\} \\ &= \{Q : a_i E_Q[X_i] + b_i = a_i P_i + b_i\} \\ &= \{Q : E_Q[X_i] = P_i\} = L_{P_i}(X_i) \end{aligned}$$

Since the linear families are identical, the point $\tilde{\lambda}^*$ will be equal to $\lambda^*$ and therefore

$$\tilde{P}^* = \tilde{\mathbf{X}}\tilde{\lambda}^* = \tilde{\mathbf{X}}\lambda^* = a \times \mathbf{X}\lambda^* + b = a \times P^* + b$$

where $\times$ is taken component-wise. □

This stands in stark contrast to CAP. Returning to our earlier example of PMF estimation, assume that instead of an identity characteristic matrix we had the following outcome matrix:

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 100 \end{bmatrix}$$

and the assessment vector $P = \begin{bmatrix} 0.7 & 0.7 & 20 \end{bmatrix}^T$. By Proposition 3.6, since $\mathbf{X}$ is an affine transformation of the original identity matrix and $P$ is the analogously transformed assessment, the coherent approximation under IGCAP will be the transform of the original approximation. In this case, that means that $P^* = \begin{bmatrix} 0.46 & 0.46 & 8 \end{bmatrix}^T$.

By contrast, the solution to the (generalized) CAP is not invariant under affine transformation. Specifically, if we state the generalized CAP as

$$\lambda^* = \arg\min_{\{\lambda | \sum_i \lambda_i = 1,\ \lambda_i \geq 0\}} ||P - \mathbf{X}\lambda||_2^2 \tag{3.18}$$

with $P^* = \mathbf{X}\lambda^*$, and then apply Lagrangian analysis we get the following solution (note in the following that $\lambda$, which is often used as the dual variable, is actually the primal variable and we use $\nu$ as the dual variable).

We form the Lagrangian,

$$\begin{aligned} L(\lambda, \nu) &= \sum_i (P_i - \text{diag}(\mathbf{X})_i \lambda_i)^2 + \nu(1 - \sum_i \lambda_i) \\ &= (P_1 - \lambda_1)^2 + (P_2 - \lambda_2)^2 + (P_3 - 100\lambda_3)^2 + \nu(1 - \sum_i \lambda_i) \\ &= (0.7 - \lambda_1)^2 + (0.7 - \lambda_2)^2 + (20 - 100\lambda_3)^2 + \nu(1 - \sum_i \lambda_i) \end{aligned}$$

and then minimizing over $\lambda$ leads to

$$\lambda^*(\nu) = \left[\ 0.7 + \tfrac{\nu}{2} \quad 0.7 + \tfrac{\nu}{2} \quad 0.2 + \tfrac{\nu}{20{,}000}\ \right].$$

Enforcing the constraint that $\sum_i \lambda_i = 1$ we see that $\nu = -0.6\frac{20000}{20001}$ and therefore $P^* \simeq \left[\ 0.4001 \quad 0.4001 \quad 19.98\ \right]^T$. Compare this with the result under the identity characteristic matrix derived in Section 3.3.2, $P^* = \left[\ 0.5 \quad 0.5 \quad 0\ \right]^T$. Changing the units of assessment for one assessor, from 'probability' to 'percentage' results in significantly different optimization behavior.

### ■ 3.5.5 Market clearing and general utilities

One way of interpreting the IGCAP is that $\lambda^*$ represents a compromise solution among the inconsistent views of the various assessors. Because of the particular assessor model that we suggested in Section 3.4.3, we viewed the optimal compromise to the be the one detailed in the IGCAP, but as pointed out briefly in Section 2.5 alternative cost structures could be analyzed.

More generally, we could consider the sequence of actions as follows: each assessor makes assessment $P_i$ for $i = 1, 2, \ldots, M$. Then a social planner is responsible for selecting a distribution $\lambda^*$ over the atomic events that maximizes social welfare among the set of assessors. Each individual assessor's welfare is defined by a utility function $u_i(\pi_i, \lambda^*)$ where $\pi_i$ is the (implicit) distribution assessor $i$ holds over the atomic events. The social welfare problem is then

$$\min_{\lambda \in \Delta} \sum_i \max_{\pi \in L_{P_i}} u_i(\pi, \lambda).$$

It is evident that the IGCAP is a special case of this formulation, where $u_i(\pi, \lambda) =$

$-D(\pi||\lambda)$ for all $i$. However, other cost functions could be employed. For example, as was shown earlier, if the utility for each agent is quadratic, under an identity output matrix the problem reverts to the original CAP formulation. Another possible cost function, so far unexplored, is the earth-mover's distance between distributions, or other forms of the divergence function. We can also represent non-expectation maximizing agents, allowing us to model risk aversion or risk seeking behavior among assessors.

The structure of this problem bears a striking resemblance to the formulation of Arrow-Debreu model of competitive markets, with assessments representing resource (or endowment) constraints on individuals, $\lambda^*$ representing the equilibrium price vector, and the individual utility functions defining the cost of allocating weight $\pi_i(\omega_j)$. Mathematically, agent $i$ chooses $\pi_i^*$ to maximize $u_i(\pi_i, \lambda^*)$ subject to the constraint $P_i = \sum_j \pi_i(\omega_j) X_i(\omega_j)$. Meanwhile the social planner chooses the price vector $\lambda^*$ to maximize $\sum_i u_i(\pi_i^*, \lambda^*)$ subject to the constraint that $\sum_j \lambda_j^* = 1$.

While we don't further develop this connection here, we point the reader to references [107,108] for development of pricing policies in Arrow-Debreu models, particularly with regards to incomplete markets.

## ■ 3.6 Exchangeability and Coherence

The final two substantive sections of this chapter analyze the impact of introducing additional constraints on the set of coherent assessments. Specifically, in this section we will analyze constraining the random variables under assessment to be exchangeable and in the next we will analyze what happens to the set of coherent assessments when the random variables are constrained to be Markov with respect to some graph. The results in these sections are proven specifically for characteristic random variables, but the major results extend to non-characteristic random variables through analogous arguments.

In addition to the concept of coherence, de Finetti also introduced the concept of *exchangeability* of random variables. In one sense, exchangeability is a generalization of the concept of independent and identically distributed. However, more relevant to the material in this thesis is the constraint that it implies with respect to the joint distribution of the random variables.

A set of random variable is said to be *exchangeable* if it is invariant under permutation. Let $\tau$ be a one-to-one mapping $\tau : \{1, 2, \ldots, N\} \to \{1, 2, \ldots, N\}$. Then, by definition, a set of random variables is exchangeable if $P(X_{\{1,2,\ldots,N\}}) = P(X_{\tau(\{1,2,\ldots,N\})})$.

If the only structural information about a set of binary random variables is that they are exchangeable, then the only constraint on the joint probability distribution is that it is symmetric (in the sense of invariance to permutation). As a consequence of this symmetry, it is simple to show that the marginal probabilities of a set of exchangeable random variables must be equal. Indeed, this is the only constraint levied on the marginal probabilities by exchangeability. In other words, given a set of binary random variables $X_1, X_2, \ldots, X_n$, exchangeability implies $P(X_i) = P(X_j)$, $\forall i, j$. There are no other implications with regards to the marginal probabilities. So, con-

sistent with the geometric interpretation of coherence one can consider the set of coherent marginals with respect to exchangeability to be those that lie on the line connecting the $[0]^n$ and $[1]^n$ vertices of the hypercube.

### ■ 3.6.1 Characteristic Matrices with Matched Exchangeability Constraints

When multiple forms of structural information are known about a set of random variables, a consistent estimate of the probabilities must be in agreement with all the constraints. When a set of random variables are known to be 1) exchangeable and 2) coherent w.r.t. a characteristic matrix, the set of consistent probabilities is restricted more than by taking either structural requirement individually. There are situations in which the characteristic matrix can be "matched" to the exchangeability constraint, i.e. where coherence levies no constraints above and beyond exchangeability.

Consider a characteristic matrix $\chi$ with columns $\chi(\omega_i)$, $i = 1, 2, \ldots, |\Omega|$. Assume w.l.o.g. that all columns are unique and that $|\Omega| \leq 2^N$. For any column $\chi(\omega_i)$ let $n_i = \sum_j \mathbb{1}_{A_j}\omega_i)$, and consider sets of columns $J_k = \{\chi(\omega_i)|n_i = k\}$. Note that $|J_k| \leq \binom{N}{k}$. The following lemma is a direct consequence of the invariance under permutation property of exchangeable random variables.

**Lemma 3.1.** *If $\mathcal{A}$ is an exchangeable set of binary random variables with characteristic matrix $\chi$ then $\forall i, j,\ n_i = n_j \Rightarrow \lambda_i = \lambda_j$. Furthermore, for all $k$, $|J_k| < \binom{N}{k} \Rightarrow \lambda_i = 0\ \forall i$ s.t. $n_i = k$.*

An immediate corollary of Lemma 3.6.1 gives necessary and sufficient conditions under which a characteristic matrix is matched with exchangeability.

**Corollary 3.2.** *A characteristic matrix $\chi$ is matched to the exchangeability constraint iff $|J_k| = 0$ for all $k \notin \{0, N\}$*

Another corollary gives the conditions on a characteristic matrix such that the set of feasible marginals under both exchangeability and a characteristic matrix is exactly the intersection of the feasible sets under each structural condition alone.

**Corollary 3.3.** *For event set $\mathcal{A}$ let $\mathcal{P}_1$ be the set of marginals consistent with exchangeability, $\mathcal{P}_2$ be the set of marginals consistent with a characteristic matrix and $\mathcal{P}$ be the set of marginals consistent with both exchangeability and a characteristic matrix. Then $\mathcal{P} = \mathcal{P}_1 \cap \mathcal{P}_2$ iff $|J_k| \in \{0, \binom{N}{k}\}$ for all $k \in \{0, 1, ..., N\}$*

### ■ 3.6.2 Non-additivity of Marginal Constraints: A Counterexample

Explaining Figure 3-5, (a) are the feasible marginals based only on characteristic matrix and (b) are the feasible marginals after including exchangeability constraints (omitting the equality constraint for expositional clarity).

As stated earlier, if the members of a set of random variables are exchangeable, the feasible set of marginal probabilities are exactly those which lie on the line segment

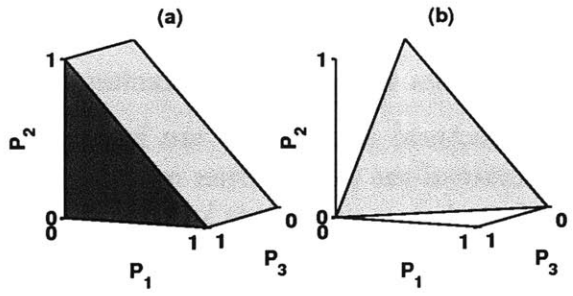



**Figure 3-5.** Feasible marginals under a combination of coherence and exchangeability constraints

between the origin and unity. To demonstrate the non-additivity of the marginal constraints, consider the following example:

$$\begin{aligned} \Omega &= \{\omega_0, \omega_1, \ldots, \omega_6\} \\ A_1 &= \{\omega_1, \omega_3, \omega_5\} \\ A_2 &= \{\omega_2, \omega_3\} \\ A_3 &= \{\omega_4, \omega_5\} \end{aligned}$$

Its characteristic matrix is

$$\chi = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

and the set of coherent marginals is shown in Figure 3-5. Note from the assessment space that

$$A_1^c \cap A_2 \cap A_3 = \emptyset \Rightarrow P(A_1^c \cap A_2 \cap A_3) = 0$$

Now, if we assume additionally that the random variables are exchangeable, we have the symmetry constraints

$$P(A_2^c \cap A_3 \cap A_1) = P(A_3^c \cap A_1 \cap A_2) = 0$$

This is equivalent to the requirement that $\lambda_3 = \lambda_5 = 0$. In addition, as stated earlier, exchangeability requires that the marginals all be equal. These two constraints together imply that only marginals of the form $P = [\alpha, \alpha, \alpha]$ where $0 \leq \alpha \leq \frac{1}{3}$ are feasible, or consistent with the structural constraints. Note, however, that the intersection of the feasibility sets for the two types of structural constraints taken

independently is $P = [\alpha, \alpha, \alpha]$, $0 \leq \alpha \leq \frac{\sqrt{2}}{2}$. This demonstrates how the set of feasible marginals given the combination of two types of structural constraints can be a strict subset of the intersection of the sets of feasible marginals of each type of structural constraint taken individually.

## 3.7 Markovianity and Coherence

Just as in Section 3.6 we examined the combination of a characteristic matrix with exchangeability among the random variables, in this section we will investigate the combination of a characteristic matrix with Markov relationships among the random variables. To motivate the problem, consider the following situation

Consider a group of individuals under suspicion of terrorist activities. These activities include involvement in one or more of a set of previously carried out terrorist events. Based on a variety of information sources, a group of experts estimate the probability that each actor in the network is, in fact, involved in terrorist activity. Now, assume there is a known social network connecting this group of individuals. Social ties are a central element in fomenting terrorist activity, so much so that it can be assumed that the probability an individual is involved in terrorist activity is conditionally independent of the activities of all other individuals in the network, given his group of immediate social connections (his neighborhood). This is exactly equivalent to saying the participation of an individual in terrorist activities is Markov with respect to the graph representing the social network.

Let $G = (V, E)$ be a graph such that for every $i$, node $v_i$ corresponds to $A_i$. Constrain the joint probability over $\mathcal{A}$ to be Markov w.r.t. $G$ and let the set of realizable marginal probabilities be denoted $\hat{\mathcal{P}}$.

### 3.7.1 Another Example of Non-Additivity of Combining Types of Structural Constraints

Consider the following example of how constraining the joint probability to be Markov w.r.t. $G$ can affect the set of realizable marginal probabilities. Consider an assessment space with the set of coherent marginals shown in Figure 3-6(a). Now, consider the graph shown in Figure 3-7(a). The interpretation of Figure 3-6 and Figure 3-7 is: (a)Top left: $\mathcal{P}$, no graph constraints (or, equivalently, a complete graph); (b) Top-Right: $\hat{\mathcal{P}}$ for graph in Figure 3-7(a); (c) Bottom-Left: $\hat{\mathcal{P}}$ for graph in Figure 3-7(b); (d) Bottom-Right: $\hat{\mathcal{P}}$ for graph in Figure 3-7(c) The additional constraint implied by this graph is $P(A_1 \cap A_3|A_2) = P(A_1|A_2)P(A_3|A_2)$. Since there is no $\omega$ s.t. $A_1, A_2$, and $A_3$ all occur, $P(A_1 \cap A_3|A_2) = 0$. Rewriting the constraint in terms of $\lambda$, we see that

$$P(A_1 \cap A_3|A_2) = 0 = \frac{\lambda_3}{\lambda_2 + \lambda_3 + \lambda_6} \frac{\lambda_6}{\lambda_2 + \lambda_3 + \lambda_6}$$

The additional constraints introduced by the graph imply that either $\lambda_3 = 0$ or $\lambda_6 = 0$ (or both). This results $\hat{\mathcal{P}}$ as shown in Figure 3-6(b). Resultant $\hat{\mathcal{P}}$ for the graphs shown in Figure 3-7(b)-(c) are shown in Figure 3-6(c)-(d), respectively.

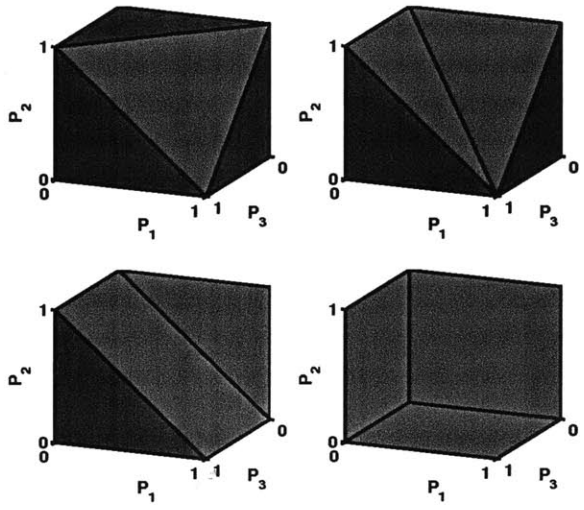
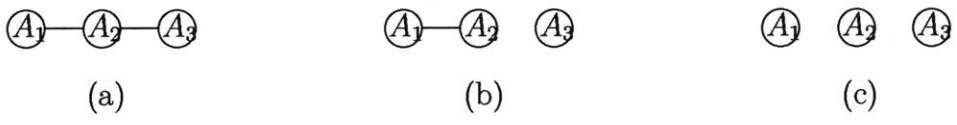



**Figure 3-6.** Feasible Sets under joint coherence and Markov constraints for four graphical structures




(a) (b) (c)

**Figure 3-7.** Markov graphs for Figure 3-6

### 3.7.2 Complete Event Sets

**Definition 3.2.** *The set of events $\mathcal{A}$, and its corresponding characteristic matrix $\chi$, is said to be* ***complete*** *if for every $a \in \{0,1\}^N$ $\exists j$ s.t. $a = \chi(\omega_j)$.*

For any complete $\mathcal{A}$, $\mathcal{P}$ is exactly the $N-$dimensional hypercube.

**Theorem 3.2.** *For any graph $G$, if $\mathcal{A}$ is complete then $\hat{\mathcal{P}} = \mathcal{P}$*

*Proof*: wlog, assume $\omega_0, \ldots, \omega_{2^N-1}$ correspond to the binary sequence ordering for $\chi(\omega_j)$ (i.e. $\chi(\omega_j) = \text{dec2bin}(j)$). For any $P \in \mathcal{P}$, let

$$\lambda_j = \begin{cases} \left(\prod_{i:\omega_j \in A_i} P_i\right)\left(\prod_{i:\omega_j \notin A_i}(1-P_i)\right) & j = 1,2,\ldots,2^N-1 \\ 0 & j \geq 2^N \end{cases} \tag{3.19}$$

**Lemma 3.2.** *Equation 3.19 satisfies the constraints $\sum_{i=0}^{|\Omega|} \lambda_i = 1$ and $P_i = \sum_{j:\omega_j \in A_i} \lambda_j$*

Thus, for any point $P \in \mathcal{P}$, where $\mathcal{P}$ is the $N$-dimensional hypercube, $\lambda$ as defined in Equation 3.19 will satisfy $P = \chi\lambda$.

**Lemma 3.3.** *Equation 3.19 satisfies the constraints for an independent joint distribution over $\mathcal{A}$*

Since the independence graph $G = (V, \emptyset)$ introduces the maximum number of constraints, and since Lemmas 3.2-3.3 imply $\hat{\mathcal{P}} = \mathcal{P}$ under the independence graph, then for any graph $G$, $\lambda$ defined by Equation 3.19 will satisfy all constraints, and thus for any graph $G$, if $\mathcal{A}$ is complete then $\hat{\mathcal{P}} = \mathcal{P}$.

**Definition 3.3.** *The set of events $\mathcal{A}$ and its corresponding characteristic matrix $\chi$, is said to be* ***degenerately complete*** *if $\mathcal{A}$ can be decomposed into two disjoint sets $(\mathcal{A}_c, \mathcal{A}_d)$ where $\mathcal{A}_c$ is complete and $\mathcal{A}_d$ is deterministic (i.e. if $A_d \in \mathcal{A}_d$ then $\chi_{dj}$ is uniformly 0 or 1 for all $j$)*

**Proposition 3.7.** *For a general graph $G = (V, E)$, $\hat{\mathcal{P}} = \mathcal{P}$ iff, for each possible decomposition of the graph into $(\mathbb{P}, \mathbb{C}, \mathbb{F})$, where $\mathbb{C}$ is a cut set on $G$, the characteristic submatrices corresponding to each realization $C$ of $\mathbb{C}$ are degenerately complete*

With the insight gained from Proposition 3.7, we can turn our attention to the central question of what effect constraining the joint probability space by graph $G$ has on the coherence set $\mathcal{P}$. From de Finetti's theorem, we have $\mathcal{P} = \texttt{convhull}(\chi)$.

Suppose $(\mathbb{P}, \mathbb{C}, \mathbb{F})$ is a partition of graph $G$ s.t. $\mathbb{C}$ is a cut set (and is not a superset of some other cut set). The set of instantiations of $\mathbb{C}$ form a partition on $\Omega$. Let the subset of atomic events in an instantiation $C$ of $\mathbb{C}$ be denoted $\Omega_C$.

Now, suppose that for some instantiation $C$ of cut set $\mathbb{C}$ of graph $G$, the conditions of Proposition 3.7 do not hold. Then, for the joint distribution to support the constraint implied by the graph, $\lambda_I = 0$ for some set $I \subset \{0, 1, \ldots, N\}$. Specifically, those atomic events which cause the degenerate completeness condition to be violated must receive zero weight.

Let $\mathcal{I}_C = \{I : \chi_{\mathcal{A}\backslash\mathbb{C}}(\Omega_C \backslash \omega_I)$ is degenerately complete$\}$ and let the elements of $\mathcal{I}_C$ be denoted $I_C$. Let $\mathcal{I}_{\mathbb{C}} = \left\{I : I = \bigcup_{C\in\mathbb{C}} I_C\right\}$ and let the elements of $\mathcal{I}_{\mathbb{C}}$ be denoted $I_{\mathbb{C}}$. Let $\mathcal{I} = \{I : I = \bigcup_{\mathbb{C}} I_{\mathbb{C}}\}$ and let $\hat{\mathcal{P}}_I = \chi\lambda$ s.t. $\lambda_I = 0$. Then

$$\hat{\mathcal{P}} = \bigcup_{I\in\mathcal{I}} \hat{\mathcal{P}}_I$$

Returning to the example, consider the graph shown in Figure 3-7(a). This can be decomposed into $\mathbb{P} = A_1$, $\mathbb{C} = A_2$, $\mathbb{F} = A_3$. First, conditioning on the event $\bar{A}_2 = \{\omega_0, \omega_1, \omega_4, \omega_5\}$, we get the characteristic submatrix

$$\chi_{\mathcal{A}\backslash\{A_2\}}(\bar{A}_2) = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

This submatrix is complete (and hence degenerately complete), and so there are no conditions on $\lambda$ due to conditioning on the event $\bar{A}_2$ (i.e. $\mathcal{I}_{A_2=0} = \emptyset$).

Next, conditioning that event $A_2 = \{\omega_2, \omega_3, \omega_6\}$ results in the characteristic submatrix

$$\chi_{\mathcal{A}\backslash\{A_2\}}(A_2) = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The characteristic submatrix is not degenerately complete. Therefore $\hat{\mathcal{P}} \neq \mathcal{P}$ (compare Figure 3-6(a) with Figure 3-6(b)). Since

$$\chi_{\{A_1,A_3\}}(A_2\backslash\omega_3) = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

and

$$\chi_{\{A_1,A_3\}}(A_2\backslash\omega_6) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

are degenerately complete, $\mathcal{I}_{A_2=1} = \{3, 6\}$. The set $I_{A_2}$ is therefore equal to $\{I : I = \bigcup_{i\in\{0,1\}} I_{A_2=i} = \{\{3, \emptyset\}, \{6, \emptyset\}\} = \{3, 6\}$. Since $A_2$ is the only cut set of $G$ that is not a superset of another cut set, we have $\mathcal{I} = I_{A_2} = \{3, 6\}$. Therefore the set $\hat{\mathcal{P}}$ can be described by

$$\hat{\mathcal{P}} = \bigcup_{i\in\{3,6\}} \{\chi\lambda | \sum_j \lambda_j = 1;\ \lambda_j \geq 0; \lambda_i = 0\}$$

## ■ 3.8 Conclusion

In this Section we have analyzed the problem of approximating an assessment of a random vector. Incoherent assessments, i.e. those that *could not* have been generated by a probability distribution are fundamentally flawed for the reasons outlined in Chapter 2. Therefore by approximating an incoherent assessment with a coherent one should improve the quality of assessment for decision making. However, care needs to be taken in the method of approximation.

One consideration is the computability of a coherent solution. We have developed

a computable approximation method, based on the Coherent Approximation Principle (CAP), that provides optimal solutions in closed-form for specially structured random vectors. It provides a boundedly suboptimal coherent approximation for all other characteristic random vectors.

We then developed, based on pragmatic and philosophical concerns with the CAP, an alternative approximation method based in information geometry. The Information Geometric CAP (IGCAP) has several attractive properties, including natural barriers as approximations move toward the edges of the simplex, the ability to handle imprecise assessments, and a unit invariant generalization to non-characteristic random vectors. We compared the IGCAP to previous suggested solution methodologies including the CAP, a divergence-based CAP, and a suggested maximum entropy method.

Finally, we considered the combination of the structural information given by an outcome matrix $\mathbf{X}$ with other types of structural constraints. Specifically, in the case of $\mathbf{X} = \chi$ we analyzed the impact on the feasible sets of joint distributions under both exchangeability and Markovianity constraints.

### 3.8.1 Future Work

We have assumed an assessment model based in the Dutch Book/Arbitrage arguments from Chapter 2 that assessments are subjective expectations of random variables. However, a large body of economic literature has analyzed human behavior and found that pricing and other assessments are often not made in an expected value manner. In general, particularly when dealing with randomized variables that represent monetized gains, human behavior seems to tend toward risk aversion. This can be mathematically modeled by replacing $\mathbb{E}[X]$ with $\mathbb{E}[u(X)]$ where $u$ is a concave function. In general, the utility function may differ between assessors. It might be fruitful to analyze such a case. As a starting point, Dutch Book arguments are analyzed for non-expected value models in [109]. The personal utility functions of individual assessors could be learned over time; such information could theoretically improve the approximation.

Another commonly employed method in expert opinion fusion is the use of individual *reputations*. Such reputations could lend more or less weight to individual assessor's assessments, and could be based on past predictive performance (which would require the imposition of a scoring rule, as discussed in [78]). In the IGCAP model, reputation could be added quite naturally by adding a multiplicative factor $\alpha_i$ to each of the inner minimization functions. Thus, if an assessor's reputation is bad, his $\alpha$ is near zero, and the overall cost is relatively immune to large divergences from his opinion. Similarly, if an assessor's reputation is good, his $\alpha$ is large and the overall cost is much more reactive to approximation deviations from his opinion.

The divergence-based cost model we've used in this development has been contrasted with a few other suggested methodologies. However, there are many other potential cost models that have yet to be explored. In particular, it would be interesting to employ the theory of optimal transport which uses the Wasserstein metric as its distance function. In terms of coherent approximation, one optimal transport-like

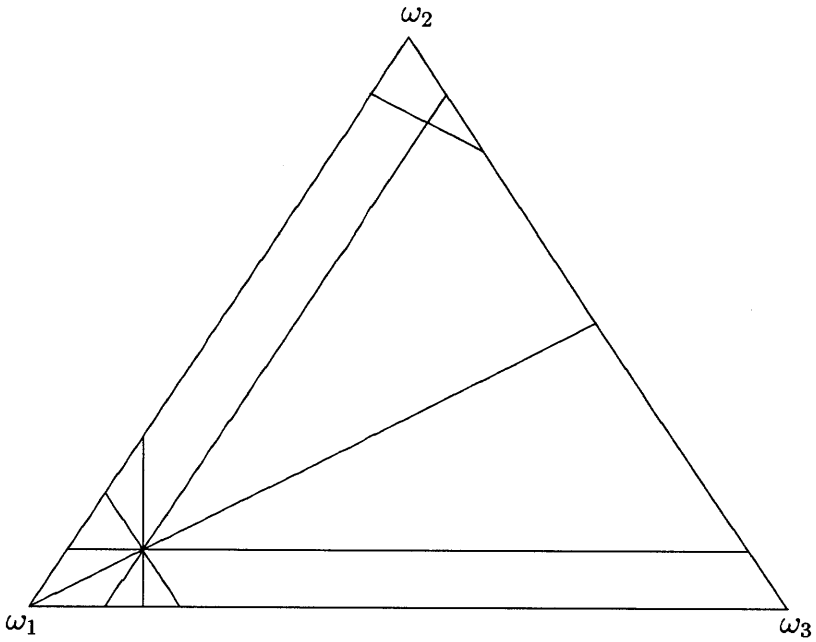



**Figure 3-8.** An example of outlier identification for coherent approximation

formulation would minimize the Wasserstein metric for moving between each of the linear families induced by the assessments.

Finally, there is a potentially interesting problem of outlier detection, demonstrated in Figure 3-8. In the figure a scenario is shown where a large number of assessors are coherent, or nearly coherent with respect to each other. But a single assessor has a significantly divergent opinion which, in this case, could 'pull' the coherent approximation significantly away from the assessment agreed upon (or nearly agreed upon) by all other assessors.

# Chapter 4

# Dynamic Coherent Approximation

## 4.1 Introduction

In the previous two chapters we developed a theory for using coherence as a mechanism for approximating mutually inconsistent assessments of multiple experts. In the development thus far the focus has been on static assessments, in which experts make a one-time assessment of a set of random variables and no further information is considered. This neglects the important role that sequential information may play and the manner with experts update their assessments with respect to that information, or the way in which the decision maker who is coherently approximating the assessments may integrate further information. As an example of the critical importance this update function may have on coherence, consider the following real life example of collective incoherence.

### Prediction Market Example

An increasingly popular method of obtaining probability assessments is through the use of so-called "prediction markets," also sometimes termed opinion markets or information markets [110, 111]. In these futures markets contracts are sold on future events. In general a contract is worth one dollar if the event obtains and zero otherwise. It is fairly simple to see that such markets (if frictionless) would induce a purchase price equal to the participants' subjective probability of the occurrence of the event [111, 112] (although see [113] for a critique of this view). These opinion markets have most popularly been applied to predicting political outcomes [114].

Shown in Figure 4-1 are the prices of two contracts from the popular on-line prediction market InTrade$^{TM}$. The lower (red) line in the figure encodes the contract price of the future contract "Democrats control 50 seats or fewer in the 2010 US Senate". The middle (blue) line encodes the contract price of the complementary contract "Democrats control 51 seats or more in the 2010 US Senate". The key point to notice is that at no point during the month and a half prior to the election were the assessments coherent. Indeed, more than half the time the contracts were sufficiently incoherent that, even given the trading costs (frictions), there existed the potential for guaranteed profit. Specifically, the upper (magenta) line shows the cost of purchasing both contracts. Since the two are complementary, one or the other will pay out \$1. So purchasing both results in a guaranteed outcome of \$1, and therefore the combined price should always equal \$1. However, as is shown in the figure, the

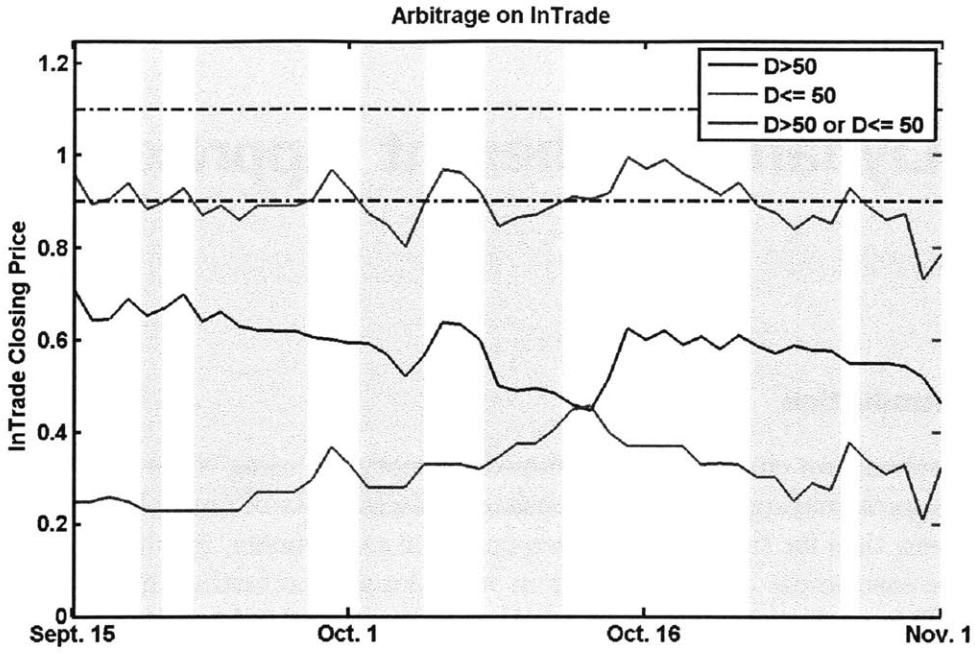



**Figure 4-1.** Incoherence in futures prices for two InTrade contracts

combined price is often significantly less that \$1; so much so that even accounting for the \$0.05 fee per contract levied by InTrade, there remained frequent arbitrage opportunities (highlighted by yellow stripes in the figure).

The relationship between arbitrage in prediction markets (such as that shown in the figure) and probabilistic coherence was previously developed in Chapter 2. We return to it now in order to reinforce the idea that progressive revelation of information does not necessarily lead to more coherent assessments.

### ■ 4.1.1 Information Integration

In the opinion market example the information provided to the decision maker is a sequence of assessments. Presumably some information integration process is driving the change in opinion within the market, but that information is obscured from the decision maker. As such, when formulating a dynamic coherent approximation, the decision maker must depend solely on the observed sequences of assessments.

A different model of information integration would consist of each assessor providing the decision maker a likelihood function that will govern the "probability kinematics" [115,116] of their individual assessment update. Then, assuming the sequence of observations is visible to the decision maker, the temporal path of each individual assessment can be determined. In this case the information integration is performed by the decision maker himself, but according to predefined rules given by the assessors. In this chapter we will formulate a concept of *dynamic coherence*. The central question we wish to answer is how the Bayesian belief revision process can introduce

incoherence into a set of assessments and what can be done about it. We first extend the mathematical model to include observational data and analyze the convergence properties of Bayesian belief revision. Then we formulate two classes of coherent likelihood models: step-wise coherent models and weakly, asymptotically coherent models. We show that step-wise coherent models are a strict subset of weakly, asymptotically coherent models. We further suggest an algorithm by which the dynamic assessments due to asymptotically coherent models can be made step-wise coherent. And finally we suggest some possible extensions to the case of non-characteristic random variables.

In this chapter we will analyze three models of information integration and suggest methods for dynamic coherent approximation based in the principles of the IGCAP introduced in Chapter 3.

#### Subjective Likelihoods

In the first model we analyze we assume a sequence of IID globally observable random variables. Each assessor has specified a subjective likelihood model over the random variables conditioned on the occurrence or non-occurrence of his personal event under assessment (the development is in terms of characteristic random variables, but generalizes in the obvious way to non-characteristic random variables). We assume that, after making an observation, each assessor updates his assessment using his subjective likelihood model via Bayes' rule.

We suggest two potential definitions of coherence for the likelihood models of the group, taken collectively, one provably strictly weaker than the other. We then employ the weaker concept of coherence to formulate a dynamic approximation mechanism, based on a principle of preserving predictive uncertainty and demonstrate the algorithm's effectiveness relative to other dynamic approximation methods.

#### Coherent Approximation of Conditional Assessments

A second model involves the coherent approximation of conditional assessments. In this model, assessments are given sequentially, but no *a priori* likelihood model is specified. The fundamental question we wish to ask is how to formulate and then utilize the concept of coherence when assessments are conditioned on different underlying events.

We show that the linear family framework, developed in Chapter 3 is sufficiently flexible to be adapted to this problem. Specifically, we show that any conditional assessment defines a linear family on the simplex of probability distributions over the atoms of $\Omega$. Thus, coherent approximation can proceed just as before, by finding a distribution with minimum average divergence to a set of linear families.

#### Coherent Markov Filtering

In this model we assume a finite number of possible states and an ergodic transition matrix between states. The 'truth' state is dynamically varying, following a Markov path determined by the transition matrix. At each time step assessors provide as-

sessments of their individual random variables given their (incomplete) knowledge of the current 'truth' state.

We analyze this problem in the same framework as the previous ones, with assessments at each time step being coherently approximated, then serving as an observation for the correction of the current predicted state. The key difficulty to be overcome is the lack of an equivalent of the observation model. However, we are able to employ the IGCAP to overcome this challenge in an innovative way.

### ■ 4.1.2 Previous Work

Previous authors have analyzed coherence with respect to contingent (or conditional) probability assessments [117–119]. These developments attempt to determine conditions characterizing coherent subjective posteriors. While likelihood models are a form of contingent probability assessment, this paper goes further in analyzing the impact of these assessments on coherent belief dynamics.

In [120, 121] a different form of conditional coherence is suggested which derives from coherence of a joint probability distribution over observations and states of nature. It is shown that for this stronger form of conditional coherence, certain specially structured event sets and likelihood functions will produce coherent posterior assessments.

The work of Skyrms takes still another, more philosophical, approach to the concept of sequential coherence [71, 116, 122]. In [116] the concept of probability kinematics (following Jeffreys [63] and Diaconis and Zabell [115]) is exploited to demonstrate a probabilistic epistemological system in which the only coherent update rule is Bayesian conditionalization (or it's generalization under Jeffreys' belief kinematics model). In [122] a brief survey is taken of sequential coherence-like epistemological arguments, and it is concluded that a variety of Dutch Book arguments are invoked.

These treatments of conditional and sequential coherence have focused primarily on the definitional question of what conditional coherence is, and how it should be detected and understood. In this section we take a different approach, focusing instead on the practical question of how to revise sequences of incoherent assessments in an appropriate manner. In purpose, this work is more similar to recent work in propositional logic, where the question of how to revise databases in the face of contradicting facts is analyzed. Recent work largely derives from a set of axioms proposed by Alchourron, Gardenfors and Makinson [123]. They're concern is primarily with revising a corpus of "facts" to be consistent with a new observation. They define certain minimality conditions with respect to the corpus, and derive a method that satisfies the conditions.

## ■ 4.2 Subjective Likelihood Functions

In this Section we begin the analysis of the first type of information integration structure: subjective likelihood functions. In this case the approximator knows, for each assessor, the likelihood function he will use to update his assessment given an observation. The assessor also knows the sequence of observations (we assume all

information is public). We will formulate a method for approximating the sequence of assessments by another sequence that is 1) coherent at each step and 2) conserves uncertainty in a specific way.

### 4.2.1 Motivating Example

Concerned with their network security, BigCorps wants to purchase an Intrusion Detection and Prevention System (IDPS). They have two options, $\texttt{IDPS}_1$ and $\texttt{IDPS}_2$. $\texttt{IDPS}_1$ detects both distributed denial of service (DDoS) attacks and port scan (PS) attacks, while $\texttt{IDPS}_2$ detects only DDoS attacks. While studying the NIST guide to IDPSs [124], BigCorps' CTO notes the recommendation that "organizations should consider using multiple types of IDPS technologies to achieve more comprehensive and accurate detection and prevention of malicious activity." Following the NIST recommendation, BigCorps purchases both IDPSs and sets them to work monitoring network traffic.

One morning while reading the output reports of the two detectors, an intrepid security analyst witnesses an interesting behavior. $\texttt{IDPS}_2$ is registering an attack probability of 0.1 while detector $\texttt{IDPS}_1$ is reading an attack probability of 0.05. Since the threats detected by $\texttt{IDPS}_1$ are a superset of those detected by $\texttt{IDPS}_2$, the probability assigned by $\texttt{IDPS}_1$ should always be larger than that assigned by $\texttt{IDPS}_2$. The dilemma faced by our analyst is how to reconcile the logically incoherent outputs of the two detectors. Particularly, how to ascribe probabilities in a way that is logically consistent, but still retains as much as possible the expert assessments of the detectors.

### 4.2.2 Mathematical Model

Let $\Omega = \{\omega_1, \omega_2, \dots\}$ be an event space and $(\Omega, \mathcal{F})$ a measurable space. Let $\theta : \Omega \to \Theta$ be a measurable random variable; consider $\Theta = \{\theta^1, \theta^j, \dots, \theta^J\}$ to be the set of all possible "states of the world." Also, let $Y_i : \Omega \to \mathcal{Y}$ be a sequence of measurable random variables; consider $Y_i$ to be the sequence of observations, with $\mathcal{Y} = \{y^1, y^2, \dots, y^K\}$ and $K < \infty$. Generally we can assume $\mathcal{Y} = \mathbb{R}^K$. Let $\Omega_\theta$ (resp. $\Omega_{Y_i}$) be the pre-image of $\theta$ (resp. $Y_i$). Since the random variables are assumed measurable, $\Omega_\theta$ and $\Omega_{Y_i}$ are measurable sets (i.e. elements of $\mathcal{F}$), as are their countable intersections and unions.

For $i = 1, 2, \dots, N$, let $A_i^\theta$ be a subset of $\Theta$, let $A_i = \cup_{\{\theta \in A_i^\theta\}} \Omega_\theta$ and let $\mathcal{A} = \{A_i\}$. We call elements of $\mathcal{A}$ **events under assessment**. The characteristic matrix $\chi$ for the events under assessment is defined as

$$\chi_{ij} = \begin{cases} 1 & \theta^j \in A_i^\theta \\ 0 & \text{o.w.} \end{cases}.$$

An individual probability assessment $P : \mathcal{A} \to [0, 1]$ maps each event under assessment to the unit interval. We write $P(\cdot)$ to denote the function and $P$ to denote the vector $P(A_i)_{i=1}^N$ and refer to $P$ as a (joint) assessment. A **coherent** assessment (i.e. one that is logically consistent) can be described geometrically as lying in the convex hull of the columns of $\chi$, meaning $\exists \lambda \in [0, 1]^J$ s.t. $\sum_i \lambda_i = 1$ and $P = \chi\lambda$.

While we specialize the development to events $A$, the development can be generalized to any set of random variables $X_i$, with the special case that $X_i = \mathbb{1}_{A_i}$ being the one here under consideration.

We now consider a sequence of probability assessments $P_n$ defined as follows: $P_n$ is the result of a belief revision process based on an initial probability assessment $P_0$, a likelihood model $p_n(y|A)$, and a sequence of observations $Y_1, Y_2, \ldots, Y_n$.

A likelihood model $p_n(y|\mathcal{A})$ is a pair of probability mass functions over the observations: one conditioned on $A$ and the other conditioned on $\bar{A}$ (where $\bar{A}$ denotes the complement of $A$). We will make the simplifying assumption that the likelihood model is *static*, i.e. $p_n(y|A) = p(y|A)$ and $p_n(y|\bar{A}) = p(y|\bar{A})$ for all $n$.

In this section we assume belief revision dynamics governed by Bayes' rule, i.e.

$$P_{n+1} = \frac{p(y_{n+1}|A) * P_n}{p(y_{n+1}|A) * P_n + p(y_{n+1}|\bar{A}) * (1 - P_n)} = \frac{1}{1 + \frac{p(y_{n+1}|\bar{A})}{p(y_{n+1}|A)} \frac{1-P_n}{P_n}}$$

To simplify development, denote $p(y = y^i|A_j) = \alpha_{ij}$ and $p(y = y^i|\bar{A}_j) = \beta_{ij}$ and assume $\forall j$, $\exists i$ s.t. $\alpha_{ij} \neq \beta_{ij}$ (i.e. each event has at least one informative observation) and $\alpha_{ij} \in (0,1)$, $\beta_{ij} \in (0,1)$ for all $i, j$ (i.e. no observation determines absolutely whether any event obtains). Then by induction the posterior probability of event $A$ after $n$ observations is:

$$P_n(A_j) = \frac{1}{1 + \frac{1-P_0}{P_0} \prod_{i=1}^{K} \left(\frac{\beta_{ij}}{\alpha_i}\right)^{n_i}} \tag{4.1}$$

when $n_i$ is the number of observations $y^i$.

## ■ 4.3 Probability Convergence for Single Assessors

For a single assessor revising his estimate of the likelihood of event $A$, let the probability model be given by $p(y = y^i|A) = \alpha_i$ and $p(y = y^i|\bar{A}) = \beta_i$. It is convenient to rewrite (4.1) in terms of the ratio $\rho_i = \frac{n_i}{n}$ and for simplicity assuming $P_0 = 0.5$ (although the analysis holds for general $P_0 \in (0,1)$). Substituting yields

$$P_n = \frac{1}{1 + \left[\prod_{i=1}^{K} \left(\frac{\beta_i}{\alpha_i}\right)^{\rho_i}\right]^n} \tag{4.2}$$

Note that 1) $\rho$ is the empirical distribution over the observations, and so converges almost surely (a.s.) to the true generating distribution, and 2) the convergence properties of $P_n$ are determined by the quantity between the square brackets in (4.2). Specifically, let

$$L_\infty = \lim_{n\to\infty} \prod_{i=1}^{K} \left(\frac{\beta_i}{\alpha_i}\right)^{\rho_i}$$

$L_\infty$ is commonly referred to as the likelihood ratio, familiar from classical binary hypothesis testing. Since $\rho$ converges a.s. and the function is continuous, $L_\infty$ exists

a.s. If $L_\infty < 1$ then $P_n \to 1$; if $L_\infty > 1$ then $P_n \to 0$; if $L_\infty = 1$ then $P_n \to \frac{1}{2}$.

### 4.3.1 Matched Likelihood Functions

Assume that the likelihood model is both infinitely precise and infinitely accurate, meaning that when $A$ (resp. $\bar{A}$) obtains observations are generated i.i.d. according to $\alpha$ (resp. $\beta$).

Assume that $A$ obtains; then $L_\infty = \prod_{i=1}^{K} \left(\frac{\beta_i}{\alpha_i}\right)^{\alpha_i}$ a.s. Let $\mathcal{L}_\infty = \log L_\infty$ which in this case yields

$$\mathcal{L}_\infty = \log \prod_{i=1}^{K} \left(\frac{\beta_i}{\alpha_i}\right)^{\alpha_i} = \sum_{i=1}^{K} \alpha_i \log \frac{\beta_i}{\alpha_i} = -D(\alpha||\beta) < 0$$

where all relations hold a.s., $D(\cdot||\cdot)$ is the relative entropy [125], and the last inequality follows since by assumption $\alpha \neq \beta$. Since $\mathcal{L}_\infty < 0 \Leftrightarrow L_\infty < 1$, this implies that when the true generating distribution is $\alpha$, $P_n \to 1$ a.s.

Similarly, when $\bar{A}$ obtains, we have

$$\mathcal{L}_\infty = \log \prod_{i=1}^{K} \left(\frac{\beta_i}{\alpha_i}\right)^{\beta_i} = \sum_{i=1}^{K} \beta_i \log \frac{\beta_i}{\alpha_i} = D(\beta||\alpha) > 0$$

and $P_n \to 0$ a.s.

### 4.3.2 Mismatched Likelihood Functions

Now consider the situation when the expert assessed likelihood model is incorrect. Assume the observation generating distribution is $\gamma = \mathbb{P}(Y_i = y)$ where $\gamma \neq \alpha$ and $\gamma \neq \beta$. In this case, $\mathcal{L}_\infty = \sum \gamma_i \log \frac{\beta_i}{\alpha_i}$. We define

$$T(\gamma) = -\mathcal{L}_\infty = \sum_i \gamma_i \log \frac{\alpha_i}{\beta_i} \tag{4.3}$$

Then the probability simplex over the observation space $\mathcal{Y}$ can be partitioned into two sets: $\mathcal{P}_0 = \{\gamma | T(\gamma) < 0\}$ and $\mathcal{P}_1 = \{\gamma | T(\gamma) > 0\}$. By the a.s. convergence of the empirical distribution, $\gamma \in \mathcal{P}_i \Rightarrow P_n \to i$. (The boundary set $\{\gamma | T(\gamma) = 0\}$ represents an unstable equilibrium in which $P_n$ a.s. converges to $\frac{1}{2}$).

The problem of mismatched likelihood functions is similar to composite hypothesis testing (c.f. [126] and references therein). Composite hypothesis testing attempts to design tests to determine the truth or falsity of a hypothesis with some ambiguity in the underlying parameter space. Because of this ambiguity, each hypothesis $\mathcal{H}_i$ corresponds not to a single distribution, but to a set of possible distributions. In the mismatched likelihood function problem, composite spaces are formed due to the properties of Bayes' rule for a specific likelihood model. A corollary of the above result is that if $\mathcal{H}_i \subseteq \mathcal{P}_i$ then Bayes' rule (under the specific likelihood model) is an asymptotically perfect detector.

## ■ 4.4 Multiple Assessors with Structural Constraints

In Section 4.3 we analyzed convergence properties of a single event under assessment. Considering multiple events introduces the challenge of defining a dynamic concept of coherence for the assessment revision process. In this section we suggest two possible definitions of dynamic coherence and consider some of the implications of these definitions.

### ■ 4.4.1 Step-wise Coherence

We first introduce a step-wise definition of coherence, and derive equivalency conditions for the special class of 2-expert likelihood models.

**Definition 4.1.** *Under the Bayes' rule revision process, a likelihood model $p(y|\mathcal{A})$ is* ***step-wise coherent*** *(SWC) if $P_n \in \texttt{convhull}(\chi) \Rightarrow P_{n+1} \in \texttt{convhull}(\chi)$ for all $y \in \mathcal{Y}$.*

Essentially this definition says that if the posterior assessment process is coherent at any time, it will remain coherent perpetually, independent of observation sequence. We derive necessary and sufficient conditions for SWC for the characteristic matrix given by

$$\chi = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix} \tag{4.4}$$

Generalizations of this development are possible for any $\chi \in \{0,1\}^{2\times|\Theta|}$.

Note that under the characteristic matrix given by (4.4) a model is SWC iff $P_n(A_1) \geq P_n(A_2)$ for all $n$ and all coherent $P_0$. Proceeding inductively, assume $P_n$ is *marginally* SWC, i.e. $P_n(A_1) = P_n(A_2) = \pi$. Due to the continuity of the update rule, a model will be SWC iff it is coherent at the margins. For coherence, for any $i$ we must have $P_{n+1}(A_1) \geq P_{n+1}(A_2)$. By substitution into (4.1)
$\frac{\alpha_{i1}\pi}{\alpha_{i1}\pi+\beta_{i1}(1-\pi)} \geq \frac{\alpha_{i2}\pi}{\alpha_{i2}\pi+\beta_{i2}(1-\pi)}$ or, equivalently, $\frac{\alpha_{i1}}{\alpha_{i2}} \geq \frac{\alpha_{i1}\pi+\beta_{i1}(1-\pi)}{\alpha_{i2}\pi+\beta_{i2}(1-\pi)}$.
By monotonicity, $\frac{\alpha_{i1}\pi+\beta_{i1}(1-\pi)}{\alpha_{i2}\pi+\beta_{i2}(1-\pi)} \in \left[\min\left\{\frac{\alpha_{i1}}{\alpha_{i2}}, \frac{\beta_{i1}}{\beta_{i2}}\right\}, \max\left\{\frac{\alpha_{i1}}{\alpha_{i2}}, \frac{\beta_{i1}}{\beta_{i2}}\right\}\right]$. Since $\frac{\alpha_{i1}}{\alpha_{i2}} \geq \frac{\alpha_{i1}}{\alpha_{i2}}$ degenerately, for $\chi$ given by (4.4), the model will be SWC iff $\frac{\alpha_{i1}}{\alpha_{i2}} \geq \frac{\beta_{i1}}{\beta_{i2}}$ $\forall i$, or (rearranging)

$$\forall i,\ \frac{\alpha_{i1}}{\beta_{i1}} \geq \frac{\alpha_{i2}}{\beta_{i2}} \tag{4.5}$$

### ■ 4.4.2 Asymptotic Coherence

While it is relatively simple to characterize coherent models in the two assessor case, in general SWC is difficult to check. As such, we introduce a simpler condition:

**Definition 4.2.** *A likelihood model $p(y|A)$ is* ***weakly asymptotically coherent*** *(WAC) if for all observation generating distributions $\gamma$ s.t. $\lim_{n\to\infty} P_n \in \{0,1\}^N$, $\exists i$ s.t. $\lim_{n\to\infty} P_n = \chi e_i$ a.s., where $e_i$ is the $i^{th}$ unit vector.*

**Lemma 4.1.** *Step-wise coherence implies weakly asymptotic coherence.*

Assume that a model is SWC but not WAC. Since it's not WAC, there exists a $\gamma$ s.t. $Y_i$ drawn IID from $\gamma$ a.s. results in $P_n \to \hat{P}$ where $\hat{P} \in \{0,1\}^N$ is not a column of $\chi$ and is therefore not coherent. Since this holds regardless of initial conditions, assume the process is initialized coherently. Then, by a separating hyperplane argument, there must exist some $n$ (and therefore some $y_n$) s.t. $P_n \in \texttt{convhull}(\chi)$ and $P_{n+1} \notin \texttt{convhull}(\chi)$. This contradicts the assumption that the likelihood model is SWC. Therefore any SWC model is also WAC. We demonstrate that the converse is not true by counterexample in Section 4.4.2.

#### WAC for Static Models

Analogous to (4.3), we define

$$T_j(\gamma) = \sum_i \gamma_i \log \frac{\alpha_{ij}}{\beta_{ij}}. \tag{4.6}$$

For a given $\gamma$, define the logical vector $r(\gamma)$ as

$$r_j(\gamma) = \begin{cases} 0 & T_j(\gamma) < 0 \\ 1 & T_j(\gamma) > 0 \\ \text{undet} & T_j(\gamma) = 0 \end{cases} \tag{4.7}$$

**Lemma 4.2.** *A likelihood model is WAC if* $\forall \gamma$ *s.t.* $\lim_{n\to\infty} P_n \in \{0,1\}^N$, $\exists i$ *s.t.* $r(\gamma) = \chi e_i$.

Define the sets $\mathcal{P}_i = \{\gamma | r(\gamma) = \chi e_i\}$. Lemma 4.2 states that for a WAC likelihood model, $\{\mathcal{P}_i\}$ partitions the simplex (excluding unstable edge events) into sets of distributions s.t. $\gamma \in \mathcal{P}_i \Rightarrow P_n \to \chi e_i$. It is simple to show that the sets $\mathcal{P}_i$ are convex, and by definition the boundaries between sets are linear.

#### Motivating Example Revisited

Consider again the motivating example of the two IDPSs from Section 4.2.1. Recall that $\texttt{IDPS}_1$ detects a superset of the attacks detected by $\texttt{IDPS}_2$, and so this scenario conforms to the characteristic matrix analyzed in Section 4.4.1. Therefore (4.5) gives necessary and sufficient conditions for SWC, while (4.7) gives necessary and sufficient conditions for WAC.

Suppose that both the IDPSs use the interval between packet arrivals as their observation and assume the learned likelihood models for the two IDPSs happen to be geometrically distributed with parameters $a_1, a_2$ (when an attack is occurring) and $b_1, b_2$ (when no attack is occurring), with the index denoting the IDPS. We will analyze SWC and WAC for this class of models.

Plugging the given likelihood model into (4.5) implies that the model is SWC iff, for $y = 0, 1, 2, \ldots$

$$\left(\frac{1-a_1}{1-b_1}\right)^y \frac{a_1}{b_1} \geq \left(\frac{1-a_2}{1-b_2}\right)^y \frac{a_2}{b_2} \tag{4.8}$$

Equation (4.8) will be satisfied iff $\frac{a_1}{b_1} \geq \frac{a_2}{b_2}$ and $\frac{1-a_1}{1-b_1} \geq \frac{1-a_2}{1-b_2}$, which is therefore a necessary and sufficient condition for SWC.

Now, we turn to WAC. Forming $T$ as defined in (4.6), we see that

$$T_j(\gamma) = \sum_y \gamma_y y \log \frac{1-a_j}{1-b_j} + \log \frac{a_j}{b_j} = \mu \log \frac{1-a_j}{1-b_j} + \log \frac{a_j}{b_j} \tag{4.9}$$

where $\mu = E_\gamma[Y]$. By the structure of the characteristic matrix, the model will be WAC iff $T_2(\gamma) > 0 \Rightarrow T_1(\gamma) > 0$ for all $\mu \geq 0$. Assume for convenience that $a_i > b_i$. Then $\{\gamma | T_i(\gamma) < 0\} = \{\gamma | \mu < \frac{\log b_i/a_i}{\log(1-a_i)/(1-b_i)}\}$ and therefore the model is WAC iff

$$\frac{\log \frac{a_2}{b_2}}{\log \frac{1-a_2}{1-b_2}} \geq \frac{\log \frac{a_1}{b_1}}{\log \frac{1-a_1}{1-b_1}} \tag{4.10}$$

Comparing the conditions for SWC (4.8) to those for WAC (4.10), we see that any parameters satisfying (4.8) also satisfy (4.10) but not vice versa. For example $a_1 = 0.3$, $a_2 = 0.5$, $b_1 = 0.2$, $b_2 = 0.25$ don't satisfy (4.8), but do satisfy (4.10). Thus WAC is truly a weaker sense of convergence than SWC.

## ■ 4.5 Coherence with Only Finitely Many Observations

As shown in Sections 4.3 and 4.4, a WAC likelihood model generates a partition $\{\mathcal{P}_i\}$ over the observation probability simplex such that $\gamma \in \mathcal{P}_i \Rightarrow P_n \to \chi e_i$. The question we now address is, given a WAC likelihood model and finitely many observations (with empirical distribution $\hat{\gamma}_n$), how to revise an incoherent posterior probability assessment $P_n$ so that it is both coherent and consistent with the observed data.

> **Principle of Conserving Predictive Uncertainty**: Given $\hat{\gamma}_n$, choose $\lambda$ such that $\lambda_i = \Pr[\lim_{n\to\infty} \hat{\gamma}_n \in \mathcal{P}_i]$ for each $i$ (where $\gamma \in \mathcal{P}_i$ iff $P_n \to \chi e_i$).

The principle of conserving predictive uncertainty states that in revising an incoherent assessment $P_n$ to a coherent one $\tilde{P}_n$, the weight vectors over the columns of $\chi$ should reflect the uncertainty in whether the observations are being generated by a distribution in the corresponding element of the partition $\{\mathcal{P}_i\}$ (and therefore whether $P_n$ is converging to $\chi e_i$).

Given a uniform prior over generating distributions $\gamma$ and assuming Lebesgue measure $\mu$ over the parameters of the generating distribution, we can write

$$\begin{aligned} P(\gamma \in \mathcal{P}_i | \hat{\gamma}_n) &= \int_{\gamma \in \mathcal{P}_i} P(\gamma | \hat{\gamma}_n) d\mu \qquad = \int_{\gamma \in \mathcal{P}_i} \frac{P(\hat{\gamma}_n | \gamma) P(\gamma)}{\int_{\mathcal{P}} P(\hat{\gamma}_n | \gamma') P(\gamma') d\mu'} d\mu \\ &= \int_{\gamma \in \mathcal{P}_i} \frac{P(\hat{\gamma}_n | \gamma)}{\int_{\mathcal{P}} P(\hat{\gamma}_n | \gamma') d\mu'} d\mu \qquad = \frac{1}{\int_{\mathcal{P}} P(\hat{\gamma}_n | \gamma') d\mu'} \int_{\gamma \in \mathcal{P}_i} P(\hat{\gamma}_n | \gamma) d\mu \end{aligned}$$

In the limit of large $n$ $P(\hat{\gamma}_n | \gamma) \doteq e^{-nD(\hat{\gamma} || \gamma)}$ (where $\doteq$ denotes equality to the first degree in the exponent; c.f. [125]). This implies that as $n$ gets large, $\Pr[\lim_{n\to\infty} \hat{\gamma}_n \in$

$\mathcal{P}_i]$ is dominated by the point $\gamma_i^* = \arg\min_{\gamma \in \mathcal{P}_i} D(\hat{\gamma}_n || \gamma)$ (i.e. the reverse i-projection, or Maximum Likelihood estimate). This suggests the following approximation method for determining a coherent projection of $P_n$:

$$\lambda_j = \frac{P(\hat{\gamma}|\gamma_i^*)}{\sum_{j \leq |\{\mathcal{P}_i\}|} P(\hat{\gamma}|\gamma_j^*)} \tag{4.11}$$

The relationship between the ML estimates ($\gamma_i^*$) and the probability over the columns of the characteristic matrix is represented graphically in Figure 4-2. As will be shown in Section 4.6, the principle of conserving predictive uncertainty can even be effectively applied to non-WAC models.

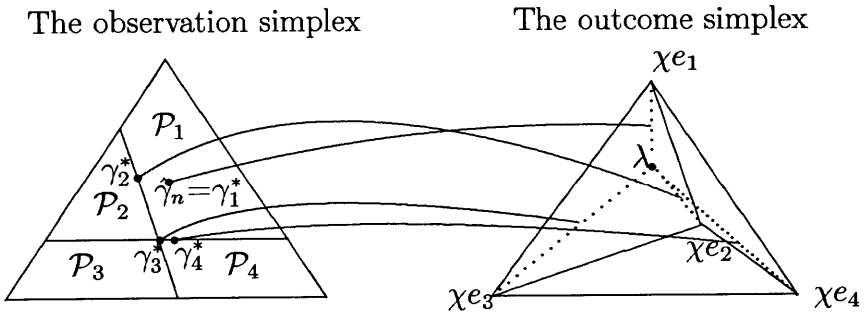



**Figure 4-2.** The relationship between observation and outcome simplices

### ■ 4.5.1 Sparse Coherent Approximation

In general $|\Theta|$ (the length of the vector $\lambda$) can be of order $2^N$ (where $N$ is the number of assessors), so solving for $\lambda$ directly using (4.11) may be computationally infeasible. The following result suggests that to generate the optimal (in the sense of capturing to most possible weight) $O(N)$ sparse approximation of $\lambda$ we need only calculate the $O(N^2)$ reverse i-projections.

Let $\lambda$ be determined according to (4.11) and let $\{\mathcal{P}_i\}$ be as defined in Section 4.4. Assume wlog that $\lambda_i \geq \lambda_j$ for all $i > j$. Define the neighborhood of $\mathcal{P}_i$ as $\mathcal{N}(\mathcal{P}_i) = \{\mathcal{P}_j : |r(\mathcal{P}_i) - r(\mathcal{P}_j)| = 1\}$ where $r(\mathcal{P}_i)$ is defined as in (4.7). The neighborhood of $\mathcal{P}_i$ is the set of partition elements such that the limit of one (and only one) assessor's probability assessment has changed. The size of the neighborhood is thus less than or equal to $N$.

By the assumed ordering of $\lambda$ and (4.11), it is immediately evident that $\hat{\gamma} = \gamma_1^*$, i.e. the maximally weighted partition element is the one that contains the empirical distribution. It can be shown that $\gamma_2^* \in \mathcal{N}(\mathcal{P}_1)$, and thus recursively that $\gamma_i^* \in \bigcup_{j<i} \mathcal{N}(\mathcal{P}_j)$. Therefore the total number of projections in calculating the $i = N$ largest weights is bounded by

$$\left| \bigcup_{j<i} \mathcal{N}(\mathcal{P}_j) \right| \leq \sum_{j<i} |\mathcal{N}(\mathcal{P}_j)| \leq \sum_{j<i} \max_j |\mathcal{N}(\mathcal{P}_j)| \leq \sum_{j<i} N = N^2.$$

Assume that each $\lambda_i$ is unique (the following argument can be generalized, if necessary). Consider $\lambda_2$ (the second largest weight) determined by Equation 4.11. The corresponding set $\mathcal{P}_2$ must be in $\mathcal{N}(\mathcal{P}_1)$. To see why, assume the contrary. Let $\delta = \{i : r(\mathcal{P}_1) \neq r(\mathcal{P}_2)\}$; by assumption $|\delta| > 1$. Let $\gamma_2^*$ be the reverse i-projection of $\hat{\gamma}$ onto the linear family $\mathcal{Q} = \{q : E_q[T_\delta] = 0\}$ and let $\tilde{\mathcal{Q}}_i = \{q : E_q[T_{\delta_i}] = 0\}$ be the linear families generated by considering each element of $\delta$ independently, with associated reverse i-projections $\tilde{\gamma}_i^*$. By construction, $\mathcal{Q} \subset \tilde{\mathcal{Q}}_i$ for all $i$ and therefore $D(\hat{\gamma}||\gamma_2^*) \geq D(\hat{\gamma}||\tilde{\gamma}_i^*)$. But this implies for large $N$ that $P(\hat{\gamma}|\tilde{\gamma}_i^*) \geq P(\hat{\gamma}|\gamma_2^*)$. And since $\exists i$ s.t. $\tilde{\gamma}_i^*$ corresponds to the reverse i-projection of $\hat{\gamma}$ onto $\mathcal{P}_i \in \mathcal{N}(\mathcal{P}_1)$, that implies that some $\lambda_i \geq \lambda_2$, which contradicts our assumption. Therefore $\mathcal{P}_2 \in \mathcal{N}(\mathcal{P}_1)$ for $n$ sufficiently large.

The foregoing analysis can be applied recursively to demonstrate that, for sufficiently large $n$, the optimal $O(N)$ sparse approximation of $\lambda$ need only compute $O(N^2)$ reverse i-projections.

## ■ 4.6 Asymptotic Coherence Simulation

Consider a three-assessor situation with an identity characteristic matrix, i.e. each of three assessors estimates the probability that his unique outcome has occurred knowing exactly one has occurred. Suppose each event is *a priori* equally likely, and a sequence of iid observations is generated with conditional probability $p(y^i|A^i) = 0.4$ and $p(y^{\bar{i}}|A^i) = 0.3$ (thus observation $y^i$ is somewhat weak evidence that event $A^i$ has occurred). Optimal joint estimation results in the posterior distribution convergence regions shown in Figure 4-3(a). Marginal estimation introduces incoherent convergence regions (4-3(b)); but for well-calibrated models, the empirical distribution is exponentially unlikely to lie in an incoherent region. However, miscalibrated models (4-3(c)) may lead to the true distribution lying in an incoherence region. WAC-approximation can ameliorate such miscalibration. The results of a Monte Carlo

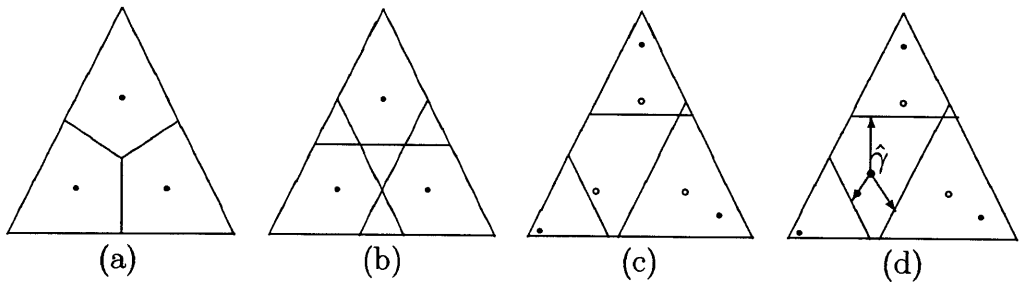



**Figure 4-3.** Equivalent decision boundaries under various likelihood model frameworks

implementation of this miscalibrated estimation is shown in Figure 4-4. The top line (blue) shows the average error for accepting the posterior assessments generated by the miscalibrated observation models. The next line (green) corresponds to renormalization at each time step, equivalent to projecting the posterior into the coherent set with a divergence-based objective function. Next (red) shows the error generated by

standard (L2) projection of the miscalibrated posterior into the coherent set. Finally, in cyan is shown the WAC approximation.

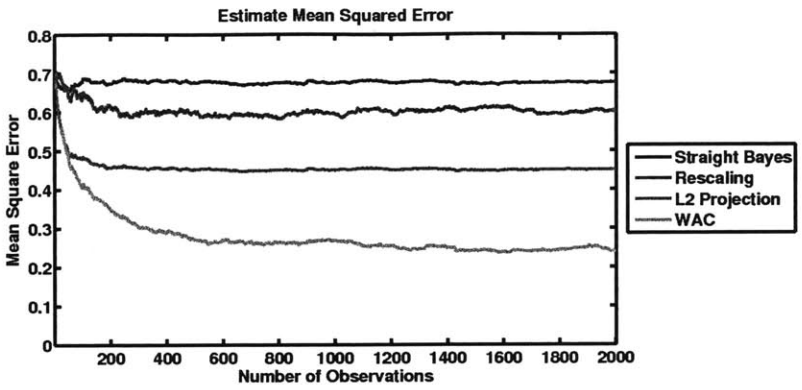



**Figure 4-4.** Comparison of mean-square errors as a function of the number of observations under four different estimation techniques

### ■ 4.6.1 Conclusions about Asymptotic Coherence

In the previous three Sections we have analyzed the problem of generating sequences of coherent assessments given subjective likelihood functions. It was shown that any collection of subjective likelihood functions forms a partition of the observational simplex, and that this partition can be used to approximate the (potentially incoherent) belief updates with a coherent approximation that preserves the amount of predictive uncertainty at each point in the sequence. While the development has been entirely in terms of characteristic random variables, the same principles can be applied directly to any finite alphabet random variables.

## ■ 4.7 Fusing Conditional and Unconditional Assessments via IGCAP

We turn now to the second model of information integration suggested in the introduction to the chapter: coherent approximation of conditional assessments with known conditioning events. We demonstrate that such conditional assessments define linear families, just as in the case of unconditional assessments. Therefore, the IGCAP formulation can be immediately adapted to this case.

First, we define what is meant by a conditional assessment. If $P$ is a conditional assessment of random variable $X$ given event $A \subseteq \Omega$, it means that $P$ is the subjective conditional expectation of an assessor, i.e. an assessor only chooses to distribute subjective probability mass over the subset of atoms in set $A$. Such a conditional assessment may be cognitively or algorithmically simpler for the assessor, since it requires forming an opinion about the relative merits of a smaller set of events. Or it may be that the assessor has additional side information that justifies making a conditional assessment.

### 4.7.1 Conditional Assessments Generate Linear Families

Consider an arbitrary random variable $X$ defined on event set $\Omega$ and an assessment $P$ of $X$ conditioned on $A \subseteq \Omega$. By a conditional assessment we mean that $P = \mathbb{E}_{Q|A}[X|A]$, i.e. that there exists a subjective conditional distribution over the elements of $A$ such that the conditional expectation is $P$.

Consider the set of all unconditional distributions consistent with conditional assessment $P$, i.e.

$$U_P(X, A) \triangleq \{Q | \mathbb{E}_{Q|A}[X|A] = P\}$$

Define random variable $\tilde{X}$ s.t.

$$\tilde{X}(\omega) = \begin{cases} X(\omega) & \omega \in A \\ P & \omega \notin A \end{cases}$$

and consider its linear family $L_P(\tilde{X}) = \{Q | \mathbb{E}_Q[\tilde{X}] = P\}$. Let $Q \in L_P(\tilde{X})$ and define $\alpha = \sum_{\omega \in A} Q(\omega)$.

$$\begin{aligned} P &= \mathbb{E}_Q[\tilde{X}] = \alpha \mathbb{E}_{Q|A}[\tilde{X}|A] + (1-\alpha)\mathbb{E}_{Q|\bar{A}}[\tilde{X}|\bar{A}] \\ &= \alpha \mathbb{E}_{Q|A}[X|A] + (1-\alpha)\mathbb{E}_{Q|\bar{A}}[P|\bar{A}] \\ &= \alpha \mathbb{E}_{Q|A}[X|A] + (1-\alpha)P \end{aligned}$$

Therefore $\mathbb{E}_{Q|A}[X|A] = P$, $Q \in U_P(X, A)$ and $L_{P_2}(\tilde{X}) \subseteq U_P(X, A)$.

Similarly, for $Q \in U_P(X, A)$,

$$\begin{aligned} \mathbb{E}_Q[\tilde{X}] &= \alpha \mathbb{E}_{Q|A}[\tilde{X}|A] + (1-\alpha)\mathbb{E}_{Q|\bar{A}}[\tilde{X}|\bar{A}] \\ &= \alpha \mathbb{E}_{Q|A}[X|A] + (1-\alpha)\mathbb{E}_{Q|\bar{A}}[P|\bar{A}] \\ &= \alpha P + (1-\alpha)P \end{aligned}$$

Therefore $Q \in L_P(\tilde{X})$, $U_P(X, A) \subseteq L_P(\tilde{X}$ and therefore $U_P(X, A) = L_P(\tilde{X})$. This demonstrates that the set of unconditional distributions consistent with a given conditional assessment form a linear family over the (unconditional) simplex. This is shown graphically in Figure 4-5 for $X = \mathbb{1}_{\omega_1}$ and $A = \{\omega_1, \omega_2\}$.

### 4.7.2 IGCAP with Conditional Assessments

Since conditional assessments define linear families the IGCAP formulation can combine conditional assessments of any known conditioning and produce a coherent approximation. Specifically, for assessments $P_1^M$ of random variables $X_1^M$ conditioned on events $A_1^M$, we define the IGCAP for conditional assessments as

$$Q^* = \arg\min_Q \sum_{i=1}^{M} D(\pi_i^* || Q) \tag{4.12}$$

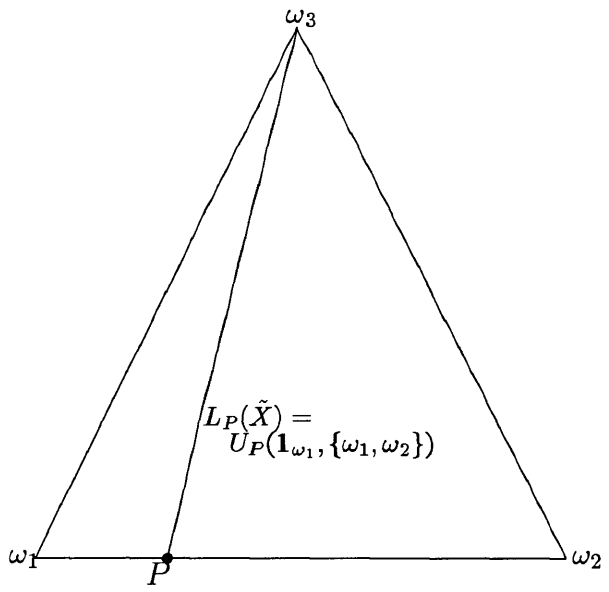



**Figure 4-5.** Linear family associated with a conditional assessment

where, for a given $Q$,

$$\pi_i^* = \underset{\pi \in U_{P_i}(X_i)}{\arg\min}\, D(\pi||Q) = \underset{\pi \in L_{P_i}(\tilde{X}_i)}{\arg\min}\, D(\pi||Q).$$

Note that this definition includes the standard IGCAP introduced in Chapter 3 if $A_i = \Omega$ for all $i$. The coherent approximation is then given as $P_i^* = \mathbb{E}_{Q^*|A_i}[X_i|A_i]$.

We now state a few facts about the IGCAP for conditional assessments. First, let $L_P(X|A)$ be the linear family of *conditional* distributions generated by assessment $P$, which lives on the simplex of $A$.

**Fact 4.1.** *If* $\bigcap L_{P_i}(X_i|A_i) \neq \emptyset$ *then* $\bigcap L_{P_i}(\tilde{X}_i) \neq \emptyset$ *and thus* $P^* = P$.

The intuition is that if the conditional linear families have a non-empty intersection, then their extension to the unconditional simplex will of necessity have a non-empty intersection. Therefore the optimal $Q^*$ will be in the non-empty intersection and $P^*$ will equal $P$.

The converse, however is not true. If $\bigcap L_{P_i}(X_i|A_i) = \emptyset$ it is possible that $\bigcap L_{P_i}(\tilde{X}_i) \neq \emptyset$. In fact, if $A_i = A$ for all $i$, $\bigcap L_{P_i}(X_i|A) \supseteq \{Q|\forall \omega \in A, Q(\omega) = 0\}$ which is non-empty for $A \neq \Omega$. In this case the IGCAP as stated will result in $Q^* \in \{Q|\forall \omega \in A, Q(\omega) = 0\}$ which is a degenerate solution (if $A \neq \Omega$). In fact, this will be the case for any conditional assessments such that $\bar{A} \triangleq \bigcap \bar{A}_i \neq \emptyset$.

We therefore propose the following slight modification to the IGCAP for conditional assessments

$$Q^* = \underset{Q \setminus D_\epsilon}{\arg\min} \sum_{i=1}^{M} D(\pi_i^*||Q) \tag{4.13}$$

where $D \triangleq \{Q|Q(\bar{A}) < 1 - \epsilon\}$ and $\epsilon > 0$ is a chosen parameter.

If $M$ assessments are given conditioned on the same event (i.e. for all $i, j$, $A_i = A_j = A$), it is reasonable to question whether the IGCAP operating on the reduced space of the simplex over $A$ results in the same solution as the IGCAP working on the unconditional space of the simplex over $\Omega$. Let $Q_u^*$ be the solution to (4.13) and $Q_c^*$ be the solution to IGCAP on the simplex of $A$, and let $P_u^*$ and $P_c^*$ be the associated coherent approximations.

**Proposition 4.1.** *$Q_c^* = Q_u^*|A$ and therefore $P_c^* = P_u^*$.*

This states that the solution to the IGCAP on the simplex of $A$ is exactly the conditional distribution of the solution to Equation 4.13 on the simplex of $\Omega$.

Before proving the Proposition, we state and prove the following Lemma. Let $L$ be a linear family on $\Delta(A)$ ($A \subset \Omega$) and define $\mathcal{L}^\alpha = \{Q \in \Delta(\Omega) : Q|A \in L, Q(A) \geq \alpha\}$, $\alpha > 0$. Thus, for every $P \in \mathcal{L}^\alpha$ there is an associated $\pi \in L$ s.t. $P_i = \alpha' \pi_i$ for all $\omega_i \in A$ and some $\alpha' \in [\alpha, 1]$. Let $Q \in \Delta(A)$ and let $\pi^* = \arg\min_{\pi \in L} D(\pi||Q)$. Define $Q^\alpha$ ($\alpha \in (0, 1]$) to be any distribution in $\Delta(\Omega)$ s.t.

$$Q_i^\alpha = \begin{cases} \alpha Q_i & i : \omega_i \in A \\ (1-\alpha)\rho_i & i : \omega_i \in \bar{A} \end{cases} \tag{4.14}$$

where $\rho$ is any valid distribution over $\bar{A}$, and define

$$\pi^{\alpha^*} = \begin{cases} Q_i^\alpha & \omega_i \notin A \\ \alpha \pi_i^* & \omega_i \in A \end{cases}.$$

**Lemma 4.3.** *If $\pi^* = \arg\min_{\pi \in L} D(P||Q)$ then $\pi^\alpha = \arg\min_{\pi \in \mathcal{L}^\alpha} D(\pi||Q^\alpha)$.*

$$\begin{aligned} D(\pi^{\alpha^*}||Q^\alpha) &= \sum \pi_i^{\alpha^*} \log \frac{\pi_i^{\alpha^*}}{Q_i^\alpha} \\ &= \sum_{\{i|\omega_i \in A\}} \pi_i^{\alpha^*} \log \frac{\pi_i^{\alpha^*}}{Q_i^\alpha} + \sum_{\{i|\omega_i \notin A\}} \pi_i^{\alpha^*} \log \frac{\pi_i^{\alpha^*}}{Q_i^\alpha} \\ &= \sum_{\{i|\omega_i \in A\}} \alpha \pi_i^* \log \frac{\alpha \pi_i^*}{\alpha Q_i} + \sum_{\{i|\omega_i \notin A\}} Q_i^\alpha \log \frac{Q_i^\alpha}{Q_i^\alpha} \\ &= \alpha D(\pi^*||Q) \end{aligned}$$

Therefore $D(\pi^{\alpha^*}||Q^\alpha) \leq \alpha D(\pi'||Q)$ for all $\pi' \in L$, and, by the positivity of divergence,

$$D(\pi^{\alpha^*}||Q^\alpha) \leq \alpha \sum_{\{i|\omega_i \in A\}} \pi_i' \log \frac{\alpha \pi_i'}{\alpha Q_i} + \sum_{\{i|\omega_i \notin A\}} \pi_i' \log \frac{\pi_i'}{Q_i^\alpha} = D(\pi'||Q)$$

for any $\pi' \in \mathcal{L}^\alpha$. □

Next, note the fact that there must exist $Q \in \Delta(A)$ such that the solution to Equation 4.13 $Q^* = Q^\epsilon$ ($Q^\epsilon$ defined as in Equation 4.14). This simply says that when all conditional assessments are conditioned on the same event $A$, the optimal solution $Q^*$ will place $1 - \epsilon$ weight on the event $\bar{A}$. This is due to the convexity of the cost function and the fact that it is zero for any mixture over the events in $\bar{A}$.

*Proof of Proposition 4.1*

If $P$ is coherent then the Proposition is trivially true. Assume $P$ is incoherent and there is therefore a unique $Q_c^*$. Let $\pi_{ic}^*$ denote its I-projections onto the linear families $L_{P_i}(X_i|A)$. Consider $Q_c^{*\epsilon}$ (defined as in Equation 4.14). By Lemma 4.3 its I-projections are $\pi_{ic}^{*\epsilon}$ and the cost associated with it is

$$\sum_i D(\pi_{ic}^{*\epsilon}||Q_c^{*\epsilon}) = \epsilon \sum_i D(\pi_{ic}^*||Q_c^*) < \epsilon \sum_i D(\pi_i^*||Q)$$

for any $Q \in \Delta(A)$ with associated I-projections $\pi_i^*$. Therefore, $Q_c^{*\epsilon}$ has minimum cost among all parameterized distributions $Q^\epsilon$. The proof of the Proposition follows from the fact that $Q_u^* = Q^\epsilon$ for some $Q \in \Delta(A)$. □

## ■ 4.8 Coherent Approximation on Markov-varying States

In this Section we address a third and final information integration structure. In previous structures we assumed that the approximator knew either 1) the likelihood functions of the individual assessors and the global sequence of observations or 2) the events on which the assessors are conditioning. We also assumed that, while the information state was dynamic, the state about which information was accruing was static. In this Section we remove these assumptions. We wish to analyze the case when an underlying state is dynamically varying, and assessors are providing assessments of unknown (to the approximator) conditioning.

Of the three dynamic models we have considered so far, this model is the most similar to the opinion market example with which we began the section. In that case, one could say the true state at any given time was whether Democrats or Republicans would eventually win the majority. The assessments, which took the form of prices in the opinion market, were conditioned on observations unavailable to the approximator. As was seen, the assessments were seldom coherent, and often were wildly incoherent, so the job of the approximator would be to take the sequences of assessments with unknown conditioning and derive an approximation that captured, to the greatest possible degree, the current wisdom of the market.

### ■ 4.8.1 Mathematical Notation

Because of the several distinct aspects of this model versus those that have gone before, we will need to introduce several new mathematical notations. Most notably, we denote by $\Theta$ a set of possible "world" states. At each discrete time $t$, a random variable $\theta_t : \Omega \to \Theta$ represents the current state of the world.

We assume the sequence of world states forms a Markov Chain, governed by the transition matrix $T$. Thus, if the probability mass over the states of the world at time $t$ is $P(\theta_t)$ then $P(\theta_{t+1}) = TP(\theta_t)$. We assume that $T$ is ergodic and thus it has a stationary distribution.

At each time $t$, assessor $i$ provides an assessment of a random variable $X_t^i$ conditioned on a private signal received at time $t$, denoted $y_t^i$, with

$$y_t = \begin{bmatrix} y_t^1 \\ y_t^2 \\ \vdots \\ y_t^M \end{bmatrix}.$$

As before, we assume the assessment is a subjective expectation, due to Dutch book arguments. In this case, the subjective distribution underlying the assessment is a distribution over the *states* given the assessors private signal. We assume that $X_t^i : \Theta \to \mathcal{X}$, i.e. that $X_t^i$ is just a function of $\theta_t$. Furthermore, we assume it is a time-invariant function, and thus will drop the time index and refer to it simply as $X_i(\theta_t)$ or simply $X_i$ (although the value will vary with time due to the underlying random state $\theta_t$).

We assume now that observations are not public knowledge, but viewed privately by each assessor. We denote the observation at time $t$ by assessor $i$ as $y_t^i$, and the set of observations by all assessors at time $t$ as $y_t$. The set of all observations by a single assessor from time $t_0$ up to and including time $t_1$ is denoted $[y_{t_0}^{t_1}]^i$ and the set of all such ranges of observations is denoted $y_{t_0}^{t_1}$.

In this case, the atoms $\Omega$ should be considered to be trajectories of $\theta_t$ over time, along with the random observations $y_t$. We assume, however, that any observation occurs with non-zero probability from any state. This assumption means that all coherence conditions of the probability space $\Omega$ are encoded in the functions $X_i$ and in the transition matrix $T$.

### ■ 4.8.2 Bayesian Filtering

Bayesian filtering is a recursive algorithm for probabilistically tracking randomly varying values over time. The fundamental structure underlying Bayesian filtering is the Hidden Markov Model (HMM) depicted in Figure 4-6. Standard assumptions for

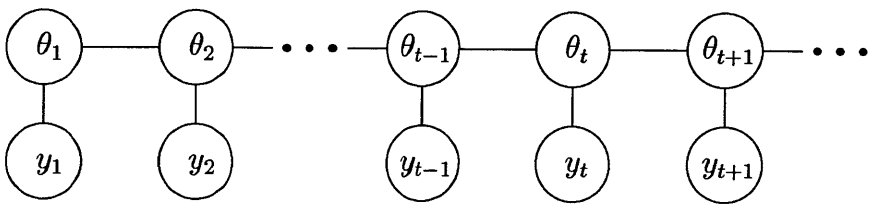



**Figure 4-6.** Graphical depiction of a Hidden Markov Model

Bayesian filtering are that the observations are independent of each other and all states future and past conditioned on the current state, and that the states evolve according to a Markov process. It is generally assumed that there exists a time-homogeneous transition matrix $T$ that governs the state dynamics, and a time-homogeneous observation model $P(y|\theta)$ that governs the observations.

The goal of Bayesian filtering is to calculate, at any given time step $t$, the probability density of $\theta_t$ given all past observations. This can be accomplished via a recursive algorithm in the following way. First, assume an initial density of the states, $P(\theta_0)$. The unconditional state at time 1 can then be calculated as

$$P(\theta_1) = \sum_{\theta_0} P(\theta_1|\theta_0)P(\theta_0) = TP(\theta_0)$$

This is generally called the 'prediction' step.

Next, we wish to integrate the observational information, forming the posterior distribution $P(\theta_1|y_1)$. This is done using Bayes' rule. Specifically,

$$P(\theta_1|y_1) = \frac{P(y_1|\theta_1)P(\theta_1)}{\sum_{\theta_1} P(y_1|\theta_1)P(\theta_1)}$$

This is called the 'update' step.

We can repeat these two steps indefinitely, calculating at each time step $t$ the distribution over the states at the current time, conditioned on all previous information. Note that the two critical elements to the Bayesian filter as developed here are the transition matrix $T$ and the observation model $P(y|\theta)$.

## ■ 4.9 Filtering Without an Observation Model

We wish to adapt the Bayesian filtering perspective to the assessment framework we have developed thus far. We have assumed knowledge of the state transition model $T$, but we do not have access, except indirectly, to any observations or even observation models. In the assessment model, what is provided at each time step is not an observation which can then be optimally combined with prior information through an observation model to form a posterior distribution. Instead, each individual assessor integrates his personal observation into his personal probability distribution, and then announces a statistic based on this personal conditional probability distribution of his personal random variable $X_i$. The information integration step is wholly hidden from the approximator.

As was demonstrated in Chapter 3, due to the subjective expectation assumption these assessments define a linear family on the simplex over the atomic states. In this case, since $X_i$ is a function of $\theta_t$, the assessment $P_t^i$ will define a set of distributions over $\theta_t$. Specifically, the linear family is a set of posterior distributions given assessor i's private information $y_i^t$ over the elements of $\Theta$. Thus each element of the linear family is a candidate distribution $p_i(\theta_t|y_i^t)$. We will denote this linear family as $L_{P_t^i}(X_i|Y_i^t)$ (where the conditioning notation is used simply to denote that the

underlying distribution is conditioned on some unobserved, to the approximator, set of information).

Suppose we had a method to select a single distribution from each linear family to use as the 'true' posterior distribution of each assessor. A method for estimating such a distribution will be discussed shortly, but for now assume that a satisfactory method exists. We wish to use this distribution to perform the update step of the Bayesian filtering algorithm.

The posterior distribution $p(\theta_t|[y_1^t]^i)$ can be rewritten using Bayes' rule as

$$p(\theta_t|[y_1^t]^i) = \frac{p(y_t^i|\theta_t)p(\theta_t|[y_1^{t-1}]^i}{p(y_t^i|[y_1^{t-1}]^i)} \tag{4.15}$$

where we've assumed that, given the current state, the current observation is independent of all past observations. We have assumed knowledge of $p(\theta_t|[y_1^t]^i)$, and we further assume that the assessors are aware of the Markov dynamics underlying the state evolution. Therefore we can rewrite Equation 4.15 as

$$p(\theta_t|[y_1^t]^i) = \frac{p(y_t^i|\theta_t)Tp(\theta_{t-1}|[y_1^{t-1}]^i}{p(y_t^i|[y_1^{t-1}]^i)} \tag{4.16}$$

We wish to isolate $p(y_t^i|\theta_t)$. Doing so results in the equation

$$p(y_t^i|\theta_t) = \alpha_i \frac{p(\theta_t|[y_1^t]^i)}{Tp(\theta_{t-1}|[y_1^{t-1}]^i)} \tag{4.17}$$

where we've denoted $\alpha_i = p(t_t^i|[y_1^{t01}]^i)$. Note that if there exists $\theta_t$ s.t. $p(\theta_t|[y_1^{t-1}]^i) = 0$ then so must $p(\theta_t|[y_1^t]^i) = 0$, and the likelihood model $p(y_t^i|\theta_t)$ would be undefined for such a $\theta_t$. This would complicate, but not significantly alter, the following analysis. As such we assume $Tp(\theta_{t-1}|[y_1^{t-1}]^i) = p(\theta_t|[y_1^{t-1}]^i) \neq 0$ for all values of $\theta_t$.

We now make an assumption about the private information available to each assessor at time $t$.

**Assumption 4.1.** *For each assessor $i$, $y_t^i$ is conditionally independent given the underlying state, i.e. $p(y_t^i, y_t^j|\theta_t) = p(y_t^i|\theta_t)p(y_t^j|\theta_t)$.*

This is a fairly standard assumption in distributed estimation problems. Essentially it says that private observations are corrupted independently. Given this assumption, we can formulate the joint probability of all private information given the underlying state. Let $y_t = \begin{bmatrix} y_t^1 & y_t^2 & \cdots & y_t^M \end{bmatrix}$.

$$\begin{aligned} p(y_t|\theta_t) &= \prod_i p_i(y_t^i|\theta_t) = \prod_i \alpha_i \frac{p(\theta_t|[y_1^t]^i)}{Tp(\theta_{t-1}|[y_1^{t-1}]^i)} \\ &= A \prod_i \frac{p(\theta_t|[y_1^t]^i)}{Tp(\theta_t|[y_1^{t-1}]^i)} \end{aligned}$$

where $A = \prod_i \alpha_i$. Note that $\alpha_i$ are still, as yet, undetermined. The vector $\alpha$ lies in

some constrained set (linear constrained set if we assume the set of possible observations is finite), but to determine it would require knowledge of the observation model, which is exactly what we don't have. However, we will see that exact knowledge of $\alpha$ is unnecessary for computing the update step of the Bayesian filter.

We now have an expression for the conditional likelihood of the set of observations at time $t$, despite the fact that we don't know what those observations or the probabilistic model that generated them were. Plugging this expression into Bayes' rule, we get the following:

$$\begin{aligned} P(\theta_t|y_1^t) &= \frac{P(y_t|\theta_t)P(\theta_t|y_1^{t-1})}{\sum_{\theta_t} P(y_t|\theta_t)P(\theta_t|y_1^{t-1})} \\ &= \frac{\left(A\prod \frac{p_i(\theta_t|y_i^i)}{Tp_i(\theta_{t-1}|y_{t-1}^i)}\right) TP(\theta_{t-1}|y_1^{t-1})}{\sum_{\theta_t}\left(A\prod_i \frac{p_i(\theta_t|y_t^i)}{Tp_i(\theta_{t-1}|y_{t-1}^i)}\right)} \\ &= \frac{\prod p_i(\theta_t|y_t^i)TP(\theta_{t-1}|y_1^{t-1})}{\sum_{\theta_t}\prod_i p_i(\theta_t|y_t^i)TP(\theta_{t-1}|y_1^{t-1})} \end{aligned} \tag{4.18}$$

So, given our assumptions that we have access to the 'true' posterior distribution $p_i(\theta_t|y_t^i)$ and assessors' observations are conditionally independent, we now have a complete method for performing Bayesian filtering in the absence of a likelihood model. Specifically, the update step consists of taking the product over all the personal posterior distributions with the prior distribution, and then appropriately normalizing.

### ■ 4.9.1 Distributed Posterior State Estimate

One assumption made in the above development was that the approximator had access to a posterior assessment, $p(\theta_t|y_i^t)$ for each assessor $i$. But for pragmatic reasons previously outlined, expecting assessors to generate a full subjective belief distribution is unreasonable. Instead, we assume that assessors offer a subjective expectation $P_i = \mathbb{E}_{p_{\theta|Y_t^i}}(\theta|y_t^i)$. The question then arises of how to go from such assessments to posterior probability distributions.

As was seen in our development in Chapter 3, each subjective assessment defines a linear family over the atomic space (in this case, $\Theta$). The IGCAP chooses as the optimal approximation the point that lies closest to those linear families in terms of summed I-divergence. A critical element of this algorithm is the calculation of the I-projection onto the linear families, which point was denoted $\pi_i^*$. The suggestion was that these points reflected the most reasonable assumption of the distribution underlying each assessor's subjective expectation because they were the 'closest' (in one sense) points that agreed with each assessor's assessment. A limited justification, in the case of characteristic random variables, was offered in terms of ML estimates from a specific assessor model, using large deviation approximations.

In this case, given a set of posterior assessments with unknown conditioning $P_i$, we can employ the IGCAP to approximate the subjective posterior probabilities

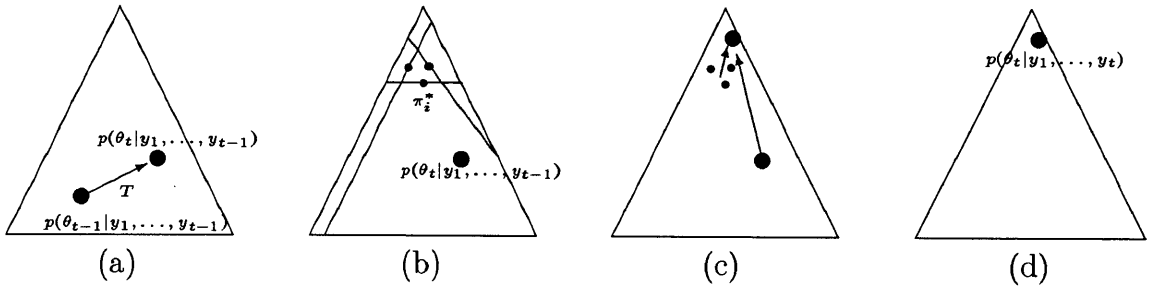



**Figure 4-7.** Graphical depiction of Bayesian filtering with assessments of unknown conditioning

$p_{\theta|Y_t^i}(\theta_t|y_t^i)$ by the points $\pi_i^*$. The rest of the algorithm proceeds according to the process outlined above and depicted graphically in Figure 4-7. In prediction step (a) the current state estimate is transitioned to a new state estimate. Then in (b) assessments are received and made coherent via IGCAP. In (c) the projection points $\pi_i^*$ are used to approximate the subjective posterior probability distributions, and are then used via Equation 4.18 to update the state estimate, as shown in (d).

### ■ 4.9.2 Max of Divergences Cost Structure

As we pointed out in Chapter 2 the specific cost structure used in the approximation problem is not unique. We have argued that divergence-based cost functions better encode intuition about how to approximate probabilities, and have advanced an assessor model in support of the sum of divergences cost function. Specifically, it was shown that, given a set of assessors, each with $n_i$ conditionally independent observations, for $n$ sufficiently large the maximum likelihood estimate of the observation generating distribution was

$$Q^* = \arg\min_Q \sum_i n_i D(\pi_i^* || Q)$$

where $\pi_i^*$ is the I-projection of $Q$ onto the linear family generated by assessor $i$'s assessment $P_i$.

Since the amount of experience of each assessor ($n_i$) is unknown, we assumed all assessors had equal experience, leading to the cost function we've employed throughout the past two chapters. As mentioned in Section 3.4.3 however, if assessor experience is treated adversarially (meaning an adversarial nature chooses each assessor's relative experience $n_i$ in order to maximize the approximation error), then a slightly different

cost function is induced, specifically

$$Q^* = \arg\min_{Q \in \mathcal{Q}} \left( \max_{\pi_i^*} D(\pi_i^* || Q) \right) \tag{4.19}$$

which will be optimized for the $Q$ that is equidistant (in terms of minimum divergence) from all the linear families. In the current context, this minimax cost structure has an interesting interpretation in terms of channel capacity.

Channel capacity is generally defined as

$$\max_{P_x} I(P_{XY})$$

where $P_x$ is an input distribution to be chosen, $P_{XY}$ is the joint distribution between inputs and outputs to the channel and $I$ is the mutual information (c.f. [125]). It can be shown that, for a given (discrete, memoryless) channel $P_{Y|X}(y|x)$, the channel capacity problem is equivalent to inducting an output distribution $P_Y$ such that

$$D(P_{Y|X}(\cdot|x)||P_Y(\cdot)) = K$$

for some constant $K$. Thus the channel capacity optimization is an attempt to choose a distribution equidistant from a set of distributions parameterized by input parameter $x$. Reasoning by analogy, we hypothesize but leave for future work to analyze, that the solution to the IGCAP as formulated in (4.19) would maximize the mutual information between local, private observations and the distribution over the states at time $t$.

## ■ 4.10 Bayesian Filtering Simulation

To verify the efficacy of the IGCAP method for filtering without (known) likelihoods, we implemented a simple simulation. As in Section 4.6 a three-state model was assumed, with three assessors, each 'responsible' for one of the three states. Also as before, the assessors have access to observations generated according to a state-dependent probability distribution. However, unlike before there is no public observation; each assessor receives a single, private observation at each time step. Also, unlike before, the state varies Markovianly according to transition matrix $T$.

Assessor $i$ performs local Bayesian filtering, generating a sequence of assessments $P_i(n)$. We assume that the assessors' local observation models are well-calibrated. If the state were not varying, this assumption would mean that the collective assessment process almost surely converges to a coherent point. However, given both the state dynamics and the independent, private observations, the sequence of assessments will generally be incoherent.

For the results reported below, observations were generated according to the following observation model. Observations were drawn from $\Theta$ according to the condi-

tional distribution, parameterized by $a \in (0,1)$, given by

$$P(y|\theta) = \begin{cases} a & y = \theta \\ \frac{1-a}{2} & y \neq \theta \end{cases} \tag{4.20}$$

In like fashion, the set of transition probabilities parameterized by $b \in (0,1)$ were given as

$$P(\theta_{n+1}|\theta_n) = \begin{cases} b & \theta_{n+1} = \theta_n \\ \frac{1-b}{2} & \theta_{n+1} \neq \theta_n \end{cases} \tag{4.21}$$

It is easy to see that the associated Markov process is ergodic with uniform stationary distribution.

Each Monte Carlo sample of the experiment consisted of 200 time steps of the Markov chain with a given set of observation and transition probabilities. New observation and transition models were generated for each Monte Carlo sample, with $a$ drawn uniformly from $[0.4, 0.8]$ and $b$ drawn uniformly from $[0.5, 0.7]$. In general, a larger $a$ means more informative observations and a larger $b$ means slower state dynamics, which favors filter accuracy.

Shown in Figure 4-8 are the combined results of 100 Monte Carlo trials. The graph represents the empirical distribution of the estimation error, i.e. $1 - P_{alg}(\theta_t)$, where $P_{alg}$ is the reported probability of state $\theta_t$ under each of several algorithms. Each line represents the aggregation of the 20,000 samples generated over the 100 Monte Carlo trials. A perfect algorithm would be uniformly equal to 1 on the plot, meaning the estimation error was zero for every sample in every trial. The blue line

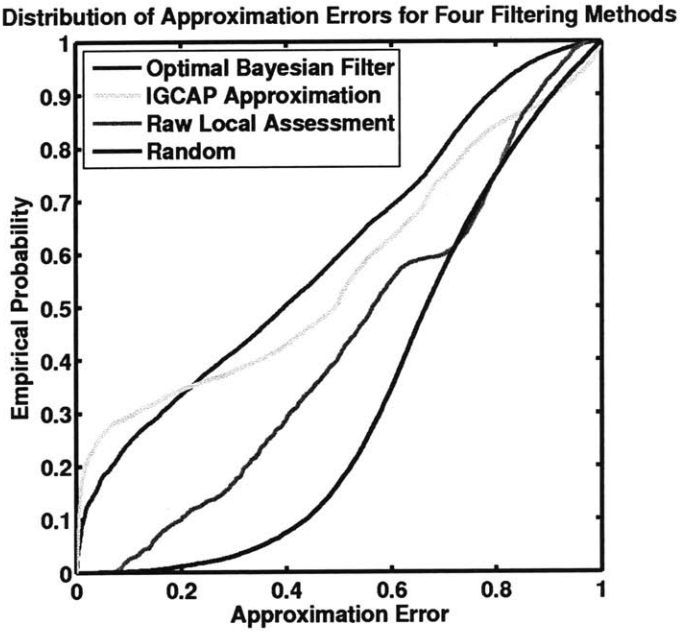



**Figure 4-8.** Performance of IGCAP for Bayesian filtering

indicates the performance of standard, centralized Bayesian filtering. In this case,

| | Bayes Filt | IGCAP Approx | Raw Assess | Random |
|---|---|---|---|---|
| mean | 0.40 | 0.44 | 0.57 | 0.67 |
| median | 0.40 | 0.49 | 0.56 | 0.67 |
| std | 0.29 | 0.34 | 0.25 | 0.18 |

**Table 4.1.** Estimation error statistics for four algorithms

three independent observations are incorporated at each time step, according to a known likelihood model. A full state probability vector is maintained and propagated according to Bayes' rule.

The red line corresponds to the performance of the local assessors. Each assessor receives only one observation per time step, and the local Bayesian filters are of a reduced form, with assessor $i$ tracking only the probability the state is $\theta_i$ or not $\theta_i$ (rather than the full state vector). Similarly, observations are of the reduced form $y == \theta_i$ or $y \neq \theta_i$.

The green line corresponds to the approximate Bayesian filtering algorithm formulated in Section 4.9, which relies on the fusion of the local assessments through IGCAP to provide an approximate likelihood model. In this case the observations are never visible to the approximator, only through the associated assessments of the local assessors.

The black line, offered as a reference, is the result of choosing $P_{rand}(\theta_t)$ uniformly at random from the interval $[0, 1]$. As can be seen from the figure, and from the statistics summarized in Table 4.1, the IGCAP filter performed quite well relative to the optimal Bayesian filter with full state and observational information. Interestingly, the IGCAP filter 'outperforms' the optimal Bayesian filter in the low estimation-error regime, although this is compensated by correspondingly poor performance in the high estimation-error regime. This indicates the algorithm is systematically overconfident, and thus that its performance could potentially be improved by introducing a correction factor into the update step of the algorithm. This result seems to be robust to Monte Carlo approximation error; it persists regardless of initial seed and number of samples.

## ■ 4.11 Conclusions

In this chapter we have analyzed the implications of coherent approximation for sequences of assessments. We considered several different information integration architectures, each varied according to what information was private and what was public, and what distributions were subjectively determined.

First we analyzed the situation where the underlying state is static but assessments are varying due to the introduction of a globally-observable random sequence of observations. While the observation process was globally observable, the *interpretation* of the data was subjective in the form of subjective likelihood functions. Two classifications of coherent likelihood models, step-wise coherence (SWC) and weak, asymptotic coherence (WAC) were suggested, and WAC was shown to be provably

weaker than SWC. Based on the structure of WAC likelihood models, a method was introduced for approximating incoherent sequences of assessments with coherent sequences that preserve the inherent uncertainty in the assessments, and this method was shown to outperform ad-hoc methods for inducing sequential coherence in a simulated example.

Next, we turned our attention to the question of coherent approximation based on both conditional assessments and unconditional assessments. It was shown that, just as with unconditional assessments, conditional assessments induce a linear family of potential unconditional probability distributions on the simplex. This means that the same principles of coherent approximation that were developed in Chapter 3 can be directly applied to the problem of coherent approximation of conditional assessments.

Finally, we introduced a model of sequential assessment with states varying over time according to a Markov process and sequences of wholly private observations. The IGCAP introduced in Chapter 3 was used to approximate the unknown likelihoods of the unknown observations. A simulation was developed in which it was shown that the IGCAP-based approximation significantly improved filtering performance over the raw local assessments.

# Chapter 5

# Distributed Coherent Risk Assessment

## 5.1 Introduction

In this chapter we turn from developing a framework for coherent approximation of distributed assessments to a particular application area. Specifically, we will employ the coherence framework developed in the foregoing chapters to analyze the problem of distributed risk assessment.

Risk assessment, as the name implies, is the process of quantifying the amount of risk in taking a particular financial position. For instance, if my financial planner suggests that I should invest my retirement savings in lottery tickets, I might like to have a quantification of how likely it is that I'll lose all my money. Individual investors are not the only parties interested in risk assessments; banks, investment firms, market regulators and governments are all impacted by risks incurred by their own and others' choices of financial positions.

Several methods of risk assessment have been suggested in the literature; we will give an incomplete overview in section 5.1.1. The critical point is that there is no standard method for assessing the risk of a financial position. As such, the risk estimates of various market participants may be inconsistent with one another, or at least provide an inconsistent picture of the understood risks of a position. We suggest methods based on the developments in Chapters 3 and 4 for coherently approximating these distributed, heterogeneous risk assessments.

### 5.1.1 Background

Risky activities were arguably the basis for the development of probability theory. The early work of Bernoulli, Pascal, de Moivre and Fermat was often motivated by gambling behavior. An entertaining account of the development of probability in terms of risk measurement and mitigation is given in [54]; a more academic treatment can be found in [42].

In terms of financial risk, the earliest modern suggested risk measure was arguably the standard deviation of the return, as per Markowitz [127][1]. Building on

[1]It has recently come to light [128] that the mean-variance portfolio selection model suggested by Markowitz in 1952 [127] had been previously published by de Finetti [129], but was not translated into English until recently. While the models are not identical [130], it is interesting to note de

Markowitz's pricing model, Treynor [131], Sharpe [132], and Lintner [133] independently developed the Capital Asset Pricing Model (CAPM). The CAPM includes a risk parameter ($\beta$) that is meant to model the non-diversifiable risk of the asset; essentially the $\beta$ parameter is the ratio of the covariance between the asset and the market and the variance of the market as a whole.

The variance-based notions of risk are intuitively appealing, but fail to adequately model realistic asset risks within a market. As such, several other risk measures have been suggested. One particularly influential risk measure is the Value at Risk (VaR) [134,135]. VaR is defined mathematically in the following section, but speaking informally the VaR measures the risk of an asset as the minimum amount of risk-free investment (i.e. 'cash') that would be needed such that the probability of combined loss between the risky and risk-free asset is less than some parameter $\gamma$. Essentially, it reports the $\gamma$-quantile of the probability distribution of the risky asset.

VaR has been an influential risk measure over the past fifteen years, but it also has significant shortcomings. This has led to the development of several suggested variants of VaR, such as the TailVaR [136], conditional VaR [137, 138], expected shortfall [139], and others. These variations are largely based on an axiomatization of risk measurement, which led to the definition of coherent risk measures [136, 140] and their generalization to convex risk measures [141–143]. More will be said about these classes of risk measures in the following sections.

### ■ 5.1.2 Coherent Approximation of Risk Measures

As was pointed out in Section 5.1.1, there is no standard risk measure; different regulatory entities depend on different measures of risk. One question that arises in such a situation is when do independent regulators, applying heterogeneous risk measures, behave consistently. Another is whether the multiple risk measures employed by a group of regulators can be coherently approximated and/or fused into a single, governing risk measure.

In this chapter we investigate several aspects of these questions. We focus primarily on *coherent* risk measures, as introduced in [136,140] and their generalization to *convex* risk measures [141]. The two questions we wish to ask with regards to heterogeneous risk measures are 1) when do risk measures behave 'well' (or, complementarily, when do they behave 'badly') under diversity of opinion among risk assessors and 2) how should multiple risk measures on different but related positions be aggregated into a global risk measure.

First in Section 5.2, we introduce our mathematical model and review some of the previous literature both on opinion divergence in financial markets and in risk measurement. Then, in Section 5.3 we use the equivalence between risk measures and acceptance sets to define a minimal convex extension of the VaR risk measure. Next, in Section 5.4 we define a robustness principle for risk measures and analyze its implications. In Section 5.5 we formulate a mechanism for fusing coherent risk assessments of individual positions into an aggregate risk assessment, and in Section 5.6

Finetti's early contribution to the field of financial risk management

we formulate the sister problem of fusing *incoherent* risk assessments into a coherent aggregate assessment. Finally, in Section 5.7 we review the main results and suggest future research directions.

## ■ 5.2 Coherence and Risk Measures

Risk measures are mappings from the space of investment outcomes to the real numbers. As the name implies, risk measures are used to quantify how much uncertainty there is in the outcome of the investment. They are particularly useful for regulators, both internal and external, in determining whether a particular position or investment should be made.

### ■ 5.2.1 Mathematical Model

A financial position $X$ is a random variable on the space $\Omega$ with $X : \Omega \to \mathbf{X}$. As per our general assumption, $|\Omega| < \infty$. A risk measure $\rho$ is a mapping from some set $\mathcal{X}$ of (absolutely bounded) random variables into the real numbers, i.e. $\rho : \mathcal{X} \to \mathbb{R}$.

In [136] risk measures are classified as 'model-dependent' if $\rho(X)$ depends on the distribution of $X$ and 'model-free' if it does not. The concept of model-dependent risk measures will be important in the development in Section 5.4.

### ■ 5.2.2 Coherent Risk Measures

In [136,140] an axiomatization is presented for risk measures. Specifically, the authors suggest four axioms relating to risk measures:

(A1) Monotonicity: If $\forall\omega,\ X(\omega) \leq Y(\omega)$, then $\rho(X) \geq \rho(Y)$.

(A2) Translation Invariance: If $m \in \mathbb{R}$, then $\rho(X + m) = \rho(X) - m$

(A3) Subadditivity: $\rho(X + Y) \leq \rho(X) + \rho(Y)$

(A4) Positive Homogeneity: If $\lambda \geq 0$, then $\rho(\lambda X) = \lambda\rho(X)$.

(A1) says that if position $X$ is uniformly worse than position $Y$ it's risk measure can be no greater than that of $Y$. (A2) says that holding an amount $m$ of risk-free asset decreases the risk of position $X$ by exactly the amount of risk-free asset held. (A3) says that diversifying positions doesn't increase the risk. And (A4) states that the risk scales linearly with the amount invested in the position.

Any risk measure that obeys these four axioms is called *coherent*. As we will see, there is a correspondence between coherent risk measures and coherent assessments as we have defined them throughout this thesis.

Any risk measure has an associated 'acceptance set.' The acceptance set $\mathcal{A}_\rho$ of risk measure $\rho$ is the set of all positions such that their risk is non-positive.

$$\mathcal{A}_\rho = \{X \in \mathcal{X} | \rho(X) \leq 0\}.$$

Alternatively, a risk measure can be defined in terms of an arbitrary acceptance set $\mathcal{A}$ as $\rho_{\mathcal{A}} = \inf\{m|X+m \in \mathcal{A}\}$. In [136] it was shown that the acceptance set associated with a coherent risk measure is closed and satisfies the following set of properties

(P1) $A_\rho \supseteq \{X \in \mathcal{X}|\forall\omega,\ X(\omega) \geq 0)$.

(P2) $A_\rho \cap \{X \in \mathcal{X}|\forall\omega,\ X(\omega) < 0) = \emptyset$.

(P3) $A_\rho$ is convex.

(P4) $A_\rho$ is a positively homogeneous cone.

In fact, the authors show a basic equivalence between the two sets (A1)-(A4) and (P1)-(P4) which, while interesting, is extraneous to our current development.

One of the central results in [136] is a representation theorem for coherent risk measures, which we reproduce here. For details on the proof, see [136].

**Proposition 5.1.** *A risk measure $\rho$ is coherent if an only if there exists a family of probability measures $\mathcal{P}$ such that*

$$\rho(X) = \sup_{P\in\mathcal{P}} \mathbb{E}_P[-X] \tag{5.1}$$

In the words of Artzner et al.,

> Any coherent risk measure appears therefore as given by a "worst case method" in a framework of generalized scenarios.

The authors immediately follow this observation with the assertion that the set of generalized scenarios should be broadly announced in order to maintain consistency within a firm or between firms on the level of cash reserves necessary to hedge against potential losses. In this chapter we suggest a different approach, in which individuals employ whichever coherent risk measure (or, equivalently, set of scenarios) they prefer, and then a decision maker interested in capturing a consistent picture of the sets of risks employs the coherent approximation techniques developed in Chapters 3 and 4 to find an equivalent joint risk measurement.

### ■ 5.2.3 Convex Risk Measures

In [142,143] the axiomatic treatment of risk measures is significantly expanded upon and generalized. A class of risk measures, termed *convex* risk measures, is developed as a relaxation of coherent risk measures. Specifically, the authors replace Artzner et al.'s (A3) and (A4) with a single convexity axiom:

(A3') Convexity: $\rho(\lambda X + (1-\lambda)Y) \leq \lambda\rho(X) + (1-\lambda)\rho(Y)$, for $0 \leq \lambda \leq 1$.

Risk measures that obey axioms (A1), (A2), and (A3') are defined as convex risk measures. It is easy to see that convexity is equivalent to subadditivity if positive homogeneity is assumed, and so any coherent risk measure is convex, but not vice

versa. In Section 5.4 we will analyze the entropic risk measure which is convex but not coherent.

As with coherent risk measures, an equivalence can be shown between the axioms of convex risk measures and properties of their acceptance sets.

(P1') $\mathcal{A}_\rho$ is convex and non-empty.

(P2') If $X \in \mathcal{A}$ and $Y \in \mathcal{X}$ are such that $Y(\omega) \geq X(\omega)$ for all $\omega$, then $Y \in \mathcal{A}$.

(P3') If $X \in \mathcal{A}$ and $Y \in \mathcal{X}$, then $\{\lambda \in [0,1] | \lambda X + (1-\lambda)Y \in A\}$ is closed in $[0,1]$.

The authors also derive a representation theorem for convex risk measures, presented here without proof. Details can be found in [143].

Define $\mathcal{M}_{1,f}$ to be the class of all finitely additive and non-negative set functions $Q$ on $\mathcal{F}$ which are normalized to $Q[\Omega] = 1$. For $|\Omega| < \infty$ this is equivalent to the set of probability distributions over the atoms of $\Omega$.

**Proposition 5.2.** *Any convex measure of risk $\rho$ on $\mathcal{X}$ is of the form*

$$\rho(X) = \max_{Q \in \mathcal{M}_{1,f}} (\mathbb{E}_Q[-X] - \alpha_{min}(Q)) \tag{5.2}$$

*where the penalty functional $\alpha_{min}$ is given by*

$$\alpha_{min}(Q) \triangleq \sup_{X \in \mathcal{A}_\rho} \mathbb{E}_Q[-X]$$

## ■ 5.3 Minimum Convex Extensions of VaR

In [136] two coherent, VaR-like risk measures are introduced and analyzed: tail conditional expectation (TCE) and worst conditional expectation (WCE). In this section we seek to answer the question of what is the coherent (or, more generally, convex) risk measure that is 'closest' (in a meaningful way) to VaR.

We begin by defining the VaR risk measure.

$$\text{VaR}_\gamma(X) = \inf\{m \in \mathbb{R} | P(X + m \leq 0) \leq \gamma\} \tag{5.3}$$

Read directly, for a given $\gamma$ VaR measures how much extra risk-free capital is necessary to bring the total probability of loss down (or up) to level $\gamma$. Perhaps the easiest way to understand it is as the negative of the value of $X$ at the $\gamma$-quantile.

We first suggest a general method to define the closest convex risk measure for any risk measure. As set out earlier, there is a relationship between risk measures and acceptance sets:

$$\mathcal{A}_\rho = \{X \in \mathcal{X} \mid \rho(X) \leq 0\}. \tag{5.4}$$

and

$$\rho_{\mathcal{A}}(X) = \inf\{m \in \mathbb{R} \mid m + X \in \mathcal{A}\}. \tag{5.5}$$

Proposition 2 in [141] shows that if $\rho$ is a convex risk measure, then $\rho_{\mathcal{A}_\rho} = \rho$. We will call any acceptance set which induces a convex risk measure under Equation 5.5 a

**convex acceptance set.** As was shown in [141], this is equivalent to an acceptance set with properties (P1')-(P3').

The equivalence between convex risk measures and convex acceptance sets provides a definition for the minimal convex extension of a given risk measure.

**Definition 5.1.** *For a given risk measure $\rho$ and it's acceptance region $\mathcal{A}_\rho$, the* **minimal convex extension of** $\rho$ *is $\rho_{\mathcal{A}_\rho^*}$ defined by Equation 5.5 where $\mathcal{A}_\rho^*$ is the smallest convex acceptance set that contains $\mathcal{A}_\rho$.*

This definition assumes that a smallest convex acceptance set exists that contains $\mathcal{A}_\rho$. We briefly justify this assumption.

Since $|\Omega| = N < \infty$, $\mathbf{X} \in \mathbb{R}^N$ and there exists a convex acceptance set containing $\mathcal{A}_\rho$ for all $\rho$ (specifically, $\mathbb{R}^N$). Analyzing properties (P1')-(P3') it is evident that the intersection of any finite number of convex acceptance sets is itself a convex acceptance set. Some care needs to be taken with an infinite number of intersections and the closedness property, but aside that technical detail it is possible to identify the smallest convex acceptance set containing $\mathcal{A}_\rho$ as the intersection of all the convex acceptance sets containing $\mathcal{A}_\rho$. Specifically, let $\mathbb{A}$ be the set of all convex acceptance sets. Then

$$\mathcal{A}_\rho^* \triangleq \bigcap_{\{\mathcal{A} | \mathcal{A} \in \mathbb{A}, \mathcal{A} \supseteq \mathcal{A}_\rho\}} \mathcal{A}. \tag{5.6}$$

In the remainder of this section we assume that $\rho = \text{VaR}_\gamma$ and we seek to define $\rho_{\mathcal{A}_\rho^*}$ for $\rho = \text{VaR}_\gamma$. Recall the definition of $\text{VaR}_\gamma$ given in Equation 5.3. Since $|\Omega| = N < \infty$, $\mathbf{X} \subseteq \mathbb{R}^N$. Denote the $2^N$ orthants of $\mathbb{R}^N$ as $O_i$ for $i \in [1 : 2^N]$, ordered according to binary expansion (i.e. $O_1$ is the positive orthant, $O_2$ is the orthant with negative first component and positive components 2 to $N$, etc). We say random variable $X$ is 'in' orthant $O_i$ if the vector

$$X = \begin{bmatrix} X(\omega_1) \\ X(\omega_2) \\ \vdots \\ X(\omega_N) \end{bmatrix}$$

is in $O_i$.

We begin by stating the following Lemma.

**Lemma 5.1.** *If $X, Y \in O_i$ then $X \in \mathcal{A}_\rho \Leftrightarrow Y \in \mathcal{A}_\rho$.*

Define the set $\Omega_{X^-} = \{\omega \in \Omega | X(\omega) \leq 0\}$. If $X, Y \in O_i$ then $X(\omega) \leq 0 \Leftrightarrow Y(\omega) \leq 0$. Therefore $P(X \leq 0) = \sum_{\omega \in \Omega_{X^-}} P(\omega) = P(Y \leq 0)$. By definition, $\mathcal{A}_\rho = \{X | P(X \leq 0) < \gamma)\}$. Therefore, $X \in \mathcal{A}_\rho$ iff $Y \in \mathcal{A}_\rho$. □

A direct result of Lemma 5.1 is that the acceptance set for $\rho = \text{VaR}_\gamma$ can be represented as a finite union of orthants, intersected with the set of positions. Formally,

for all $\gamma$, there exists $\mathcal{I}(\gamma) \subseteq [1 : 2^N]$ s.t.

$$\mathcal{A}_\rho = \left( \bigcup_{i \in \mathcal{I}(\gamma)} O_i \right) \bigcap \mathcal{X}. \tag{5.7}$$

Note that such a set need not be convex, as the following modification of the first example in Section 3.3 of [136] demonstrates.

#### Non-convex Acceptance Set Example

Given $\Omega = \{\omega_1, \omega_2, \omega_3\}$ and $P(\omega_1) = P(\omega_3) = 0.008$, consider the following pair of positions (where the $i^{th}$ vector component corresponds to $X(\omega_i)$):

$$X_1 = \begin{bmatrix} -1 \\ 0.5 \\ 0.5 \end{bmatrix} \qquad X_2 = \begin{bmatrix} 0.5 \\ 0.5 \\ -1 \end{bmatrix}$$

For $\gamma = 0.01$,

$$\rho(X_1) = \inf\{m | P(X + m \leq 0) < 0.01\} = -0.5$$

since $P(X_1(\omega_1)) = 0.008 < 0.01$. Similarly $\rho(X_2) = -0.5$ since $P(X_2(\omega_3)) = 0.008 < 0.01$ and therefore both risks are deemed acceptable under $\rho$. However, for

$$\frac{1}{2}X_1 + \frac{1}{2}X_2 = \begin{bmatrix} -0.25 \\ 0.5 \\ -0.25 \end{bmatrix}$$

we see that $\rho(\frac{1}{2}X_1 + \frac{1}{2}X_2) = 0.25$ since $P(\omega_1) + P(\omega_3) = 0.016 > 0.01$, meaning the risk of the combined position is unacceptable.

In this case it can be seen that, $\mathcal{A}_\rho = (O_1 \cup O_2 \cup O_5)$, which is a non-convex set. Specifically for $\lambda \in [1/3, 2/3]$, the set of points $\lambda X_1 + (1 - \lambda)X_2 \in O_6$ and therefore do not belong to $\mathcal{A}_\rho$.

### ■ 5.3.1 Minimal Convex Extension of VaR

To determine the minimal convex extension of VaR, we first categorize the convex risk measures associated with *convex* sets representable by Equation 5.7.

**Lemma 5.2.** *If $\mathcal{A}_\rho$ is a convex acceptance set and can be represented as in Equation 5.7, then $\exists \gamma \in [0, 1]$ s.t. $\rho_{\mathcal{A}_\rho} = \text{VaR}_\gamma$.*

Lemma 5.2 essentially says that any convex acceptance set that is equal to the union of a finite number of orthants has $\text{VaR}_\gamma$ as its associated risk measure with some parameter $\gamma$.

Next, note the fact that the smallest convex set containing a union of orthants is also a union of orthants. This, combined with Lemmas 5.1 and 5.2 gives us the following theorem.

**Theorem 5.1.** *For $\rho = \text{VaR}_\gamma$, for any given $\gamma \in [0,1]$ $\exists \gamma^*$ such that $\rho_{\mathcal{A}_\rho^*} = \text{VaR}_{\gamma^*}$.*

By Lemma 5.1, the acceptance set associated with $\rho_{\mathcal{A}_\rho^*}$ is representable by Equation 5.7. The smallest convex acceptance set containing such a set is itself a union of orthants and therefore, by Lemma 5.2, has associated risk measure $\text{VaR}_{\gamma^*}$ for some $\gamma^* \in [0,1]$. □

The import of Theorem 5.1 is that if VaR is deemed an insufficient risk measure due to its lack of convexity, and if the desire is to replace it with a minimally different risk measure that is convex, then the chosen risk measure (for a particular definition of minimal) is itself a VaR risk measure. On the one hand, this implies that VaR may not be deserving of the amount of criticism it has received. Or, viewed another way, it suggests that VaR, depending as it does on quantiles to define acceptable risk, is so rough an estimate of risk that it can't reasonably be approximated convexly with a more elegant measure. Personally, I feel that VaR may be useful in appropriate situations, but care should be taken in selecting $\gamma$ so that the resulting measure is convex.

## 5.4 Risk Measures under Divergence of Opinion

In the axiomatic developments of coherent and convex risk measures, one aspect that is not considered is the impact of opinion divergence on risk measure. In a multiple assessor environment it is natural to suppose that there may be a divergence of opinion among risk assessors as to the probability distribution over a set of outcomes. This opinion divergence can have significant effect, as the following example demonstrates.

#### Example: Divergence of opinion in VaR

Given $X = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$ and risk measure $\rho(X) = \text{VaR}_\gamma$, suppose two assessors $(A_1, A_2)$ independently evaluate the risk of position $X$. $A_1$ has private probability parameterized by $P(X = -1) = p - \delta$ and $A_2$ has private probability $P(X = -1) = p + \delta$. For $\gamma = p$ and any $\delta > 0$, $\rho_1(X) = -1$ and $\rho_2(X) = 1$. Thus, although the two opinions diverge by an arbitrarily small amount (in terms of total variation), the subjective risk measures for the position remain far apart.

### 5.4.1 Risk Measure Continuity

As was noted in Section 5.2.1, risk measures can be divided into two types: model-free, in which the risk measure is only a function of the *support* of the outcomes of a position, and model dependent, in which the risk measure is a function of the value of the probability distribution over the outcomes of the position. By their nature, model free risk measures will be immune to divergent opinions among regulators (up to the assumption that probability distributions are absolutely continuous with respect to each other). But model dependent risk measures can generally exhibit a significant effect, as the VaR example above demonstrates.

In this section (and this section only) we will break with convention and write a risk measure as a function of both the random variable $X$ and an associated probability distribution $P$, e.g. $\rho(X, P)$. Given the focus on divergence of opinion, this will simplify the exposition significantly.

Suppose we have a model dependent risk measure $\rho$ and a set of positions $\mathcal{X}$. How robust is $\rho$ to small changes in the assumed probability distribution $P$? First we will quantify what constitutes a 'small change' in $P$, and then we will define a robustness property.

For two probability distributions $P$ and $Q$ ($Q << P$), we reiterate the definition of I-divergence, previously covered in both Chapters 3 and 4, written here in its more general form.

$$D(P||Q) = \int \log \frac{dQ}{dP} dP \tag{5.8}$$

where $\frac{dQ}{dP}$ is the Radon-Nikodym derivative. Recall that the information divergence is non-negative and equal to zero if and only if $P = Q$.

The information divergence is not a metric; it is not symmetric, nor does it satisfy the triangle inequality. However, when $P \approx Q$ the divergence behaves roughly quadratically. For more details on the use of information divergence, see references [100, 101, 125]. We will define the distance *from* distribution $Q$ to distribution $P$ as $D(P||Q)$. We thus consider a small change from distribution $P$ to be

$$D_\epsilon \triangleq \{Q : D(P||Q) < \epsilon\} \tag{5.9}$$

Given this understanding of what constitutes a 'small-change' to a distribution $P$, we now define a robustness property for risk measures.

**Definition 5.2.** *A risk measure is* **robust** *if*

$$D(P||Q) < \delta \Rightarrow |\rho(X, P) - \rho(X, Q)| < C\delta \tag{5.10}$$

*for some* $C > 0$.

Essentially, this definition of robustness enforces a uniform bound on how rapidly the risk measure can vary with small changes in subjective probability.

Returning to the VaR example above, we see that VaR is not a robust risk measure. Specifically, letting $P$ be the distribution parameterized by $P(X = -1) = p - \delta$ and $Q$ be the distribution parameterized by $Q(X = -1) = p + \delta$, we see that

$$D(P||Q) = (p - \delta) \log \left( \frac{p - \delta}{p + \delta} \right) + (1 - p + \delta) \log \left( \frac{1 - p + \delta}{1 - p - \delta} \right)$$

Notice that in the above equation, $\lim_{\delta \to 0} D(P||Q) = 0$. Since $\text{VaR}_\gamma(X, P) = -1$ and $\text{VaR}_\gamma(X, Q) = 1$ for all $\delta > 0$ and therefore $|\text{VaR}_\gamma(X, P) - \text{VaR}_\gamma(X, Q)| = 2$, given any $C$ it is possible to choose $\delta$ sufficiently small such that

$$|\text{VaR}_\gamma(X, P) - \text{VaR}_\gamma(X, Q)| > C\delta.$$

### 5.4.2 Robustness of some VaR-like Risk Measures

Other $\text{VaR}_\gamma$-like risk measures are suggested in the literature, such as $TailVaR$ and $WCE$. The same example can be used to demonstrate that these risk measures are similarly fragile to small changes in assessed probability. Recall the following definitions:

$$\text{TCE}_\gamma(X) = -\mathbb{E}_{\mathbb{P}}[X|X \leq -\text{VaR}_\gamma(X)] \tag{5.11}$$

$$\text{WCE}_\gamma(X) = -\inf\{\mathbb{E}_{\mathbb{P}}[X|A] \mid \mathbb{P}[A] \geq \gamma\} \tag{5.12}$$

The TCE (also TailVaR) computes the expectation of the $\gamma$-quantile tail event while the WCE computes the smallest expected return given a non-tail event has occurred.

Using the above definitions, we can consider the robustness of TCE and WCE to small changes in probability using the same example as above. In the case of $\text{TCE}_\gamma$, we see that

$$\begin{aligned} |\text{TCE}_\gamma(X,P) - \text{TCE}_\gamma(X,Q)| &= |\mathbb{E}_P[X|X \leq 1] - \mathbb{E}_Q[X|X \leq -1]| \\ &= |\mathbb{E}_P[X] - (-1)| \\ &= -1(\gamma - \delta) + (1 - (\gamma - \delta)) + 1 \\ &= 2(1 - \gamma + \delta) \end{aligned}$$

So, for general $\gamma$, we see that as $\delta \to 0$, the difference in measure does not go to zero and therefore $\text{TCE}_\gamma$ is not robust.

Similarly, for $\text{WCE}_\gamma$ (assuming $\gamma < 0.5$), we have

$$\begin{aligned} |\text{WCE}_\gamma(X,P) - \text{WCE}_\gamma(X,Q)| &= |\min(\mathbb{E}_P[X], \mathbb{E}_P[X|X=1]) - \\ &\quad \min(\mathbb{E}_Q[X], \mathbb{E}_Q[X|X=1], \mathbb{E}_Q[X|X=-1])| \\ &= |\min(1 - 2\gamma + 2\delta, 1) - \min(\mathbb{E}_Q[X], 1, -1)| \\ &= |1 - 2\gamma + 2\delta - (-1)| = 2(1 - \gamma + \delta) \end{aligned}$$

Since this is identical to the result under $\text{TCE}_\gamma$ by the same reasoning $\text{WCE}_\gamma$ is not a robust risk measure.

### 5.4.3 Robustness of other Convex Risk Measures

In addition to $\text{TCE}_\gamma$ and $\text{WCE}_\gamma$ we here investigate the robustness of two other convex risk measures suggested in the literature: expected shortfall (or CVaR or Average VaR) [138] and entropic risk measure (which is convex but not coherent) [141].

#### Expected Shortfall

Given an $\alpha \in (0,1)$, the expected shortfall is defined as

$$ES_\alpha(X,P) = \frac{1}{\alpha}\int_0^\alpha \text{VaR}_\gamma(X,P)d\gamma$$

Employing the same example as before, and letting $\alpha = \mathbb{P}(X = -1) + \delta = \mathbb{Q}(X = -1) - \delta$ we see that for all $\gamma \leq \alpha$, $\text{VaR}_\gamma(X) = -1$. Therefore $ES_\alpha(X, Q) = -1$ while

$$\begin{aligned} ES_\alpha(X, P) &= \frac{1}{\alpha}\left(\int_0^{\alpha-\delta} -d\gamma + \int_{\alpha-\delta}^{\alpha} d\gamma\right) \\ &= \frac{1}{\alpha}\left(-(\alpha - \delta) + \delta\right) = -1 + \frac{2\delta}{\alpha} \end{aligned}$$

Therefore $|ES_\alpha(X, P) - ES_\alpha(X, Q)| = \frac{2\delta}{\alpha}$, so unlike the other $\text{VaR}_\gamma$-like risk measures, expected shortfall does not exhibit fragility in this simple example. In fact, given the assumption of a finite sample space, the expected shortfall can be shown to be robust to small model variations.

**Proposition 5.3.** *The expected shortfall risk measure is robust in the sense of Equation 5.10*

*Proof:*

$$\begin{aligned} |ES_\alpha(X, P) - ES_\alpha(X, Q)| &= \left|\frac{1}{\alpha}\int_0^\alpha \text{VaR}_\gamma(X, P)d\gamma - \frac{1}{\alpha}\int_0^\alpha \text{VaR}_\gamma(X, Q)d\gamma\right| \\ &= \frac{1}{\alpha}\left|\int_0^\alpha \left(\text{VaR}_\gamma(X, P) - \text{VaR}_\gamma(X, Q)\right) d\gamma\right| \\ &\leq \frac{1}{\alpha}\int_0^\alpha \left|\text{VaR}_\gamma(X, P) - \text{VaR}_\gamma(X, Q)\right| d\gamma \\ &= \frac{1}{\alpha}\sum_{i=1}^{N}\sum_{j\neq i} |x_i - x_j|\,\lambda_{ij} \\ &\leq \frac{c}{\alpha}\sum_{i=1}^{N}\sum_{j\neq i} \lambda_{ij} \\ &\leq \left(\frac{c}{\alpha}\right)\delta \end{aligned}$$

where $\{x_i\}_{i=1}^N$ is the ordered alphabet of $X$, $\lambda_{ij}$ is the Lebesgue measure of the interval $\{\gamma | \text{VaR}_\gamma(X, P) = x_i, \text{VaR}_\gamma(X, Q) = x_j\}$ and $c = \max_{i,j} |x_i - x_j|$. The key insight is that $\sum_{i=1}^N \sum_{j\neq i} \lambda_{ij} \leq ||P - Q||_1 \leq \delta$.

#### Entropic Risk Measure

Given a 'risk aversion' parameter $\theta$, the entropic risk measure is defined as

$$ERM_\theta(X, P) = \frac{1}{\theta}\log \mathbb{E}_P\left[e^{-\theta X}\right]$$

**Proposition 5.4.** *The entropic risk measure is robust in the sense of Equation 5.10*

*Proof:*

$$\begin{aligned}
|ERM_\theta(X,P) - ERM_\theta(X,Q)| &= \left|\frac{1}{\theta}\log \mathbb{E}_P\left[e^{-\theta X}\right] - \frac{1}{\theta}\log \mathbb{E}_Q\left[e^{-\theta X}\right]\right| \\
&= \frac{1}{\theta}\left|\log \sum_i P_i e^{-\theta x_i} - \log \sum_i Q_i e^{-\theta x_i}\right| \\
&= \frac{1}{\theta}\left|\log \frac{\sum_i P_i e^{-\theta x_i}}{\sum_i Q_i e^{-\theta x_i}}\right|
\end{aligned}$$

If $\frac{\sum_i P_i e^{-\theta x_i}}{\sum_i Q_i e^{-\theta x_i}} \geq 1$ then we have the following:

$$\begin{aligned}
|ERM_\theta(X,P) - ERM_\theta(X,Q)| &= \frac{1}{\theta}\left|\log \frac{\sum_i P_i e^{-\theta x_i}}{\sum_i Q_i e^{-\theta x_i}}\right| = \frac{1}{\theta}\log \frac{\sum_i P_i e^{-\theta x_i}}{\sum_i Q_i e^{-\theta x_i}} \\
&= \frac{1}{\theta}\log\left(1 + \frac{\sum_i \delta_i e^{-\theta x_i}}{\sum_i Q_i e^{-\theta x_i}}\right) \\
&\leq \frac{1}{\theta}\log\left(1 + \frac{\sum_i |\delta_i| e^{-\theta x_i}}{\sum_i Q_i e^{-\theta x_i}}\right) \\
&\leq \frac{1}{\theta}\log\left(1 + c_1\delta\right) \\
&\leq \left(\frac{c_1}{\theta}\right)\delta
\end{aligned}$$

where $\delta_i = P_i - Q_i$, $\sum_i |\delta_i| = ||P - Q||_1 \leq \delta$ and $c_1 = \frac{e^{-\theta \min x_i}}{N\left(\min_j Q_j e^{-\theta x_j}\right)}$.

If $\frac{\sum_i P_i e^{-\theta x_i}}{\sum_i Q_i e^{-\theta x_i}} < 1$ then we have the following:

$$\begin{aligned}
|ERM_\theta(X,P) - ERM_\theta(X,Q)| &= \frac{1}{\theta}\left|\log \frac{\sum_i P_i e^{-\theta x_i}}{\sum_i Q_i e^{-\theta x_i}}\right| = -\frac{1}{\theta}\log \frac{\sum_i P_i e^{-\theta x_i}}{\sum_i Q_i e^{-\theta x_i}} \\
&= \frac{1}{\theta}\log \frac{\sum_i Q_i e^{-\theta x_i}}{\sum_i P_i e^{-\theta x_i}} \\
&= \frac{1}{\theta}\log\left(1 + \frac{\sum_i -\delta_i e^{-\theta x_i}}{\sum_i P_i e^{-\theta x_i}}\right) \\
&\leq \frac{1}{\theta}\log\left(1 + \frac{\sum_i |\delta_i| e^{-\theta x_i}}{\sum_i Q_i e^{-\theta x_i}}\right) \\
&\leq \frac{1}{\theta}\log\left(1 + c_2\delta\right) \\
&\leq \left(\frac{c_2}{\theta}\right)\delta
\end{aligned}$$

where $c_2 = \frac{e^{-\theta \min x_i}}{N\left(\min_j P_j e^{-\theta x_j}\right)}$. Therefore $|ERM_\theta(X,P) - ERM_\theta(X,Q)| \leq \left(\frac{c}{\theta}\right)\delta$ where

$c = \max(c_1, c_2)$. □

In this section we have examined the robustness properties of several convex (and one non-convex) risk measures. We find that while some convex risk measures (Expected Shortfall, Entropic Risk) are robust to small divergences of opinion, other popular risk measures (VaR, TailVaR, WCE) are not. In a distributed environment where subjective probabilities are likely play a role, caution should be exercised in choosing risk measures that are robust to variations in the assumed probability distribution underlying a financial position.

## ■ 5.5 Fusing Risk Assessments

In the next two sections we will address the problem of fusing the risk assessments of several independent assessors. By a risk assessment we mean a specific instance of a risk measure.

The problem of fusing risk assessments occurs in several situations. For instance, a risk manager at an investment firm must take the individual risk measures of the several analysts and generate a single assessment of the overall risk held by the firm. Alternatively, a market or exchange regulator may receive risk assessments from individual firms and need to generate an overall measure of the comprehensive risk in the market. The challenge of the fusion problem is to generate an overall assessment that captures as much as possible the expert opinion of the individual assessors.

We analyze two related problems with regards to the fusion of risk assessments. The first assumes that all individual assessments are the result of a single coherent risk measure, operating on a set of different positions, but that the approximator does not know which specific coherent risk measure is being used. This may arise, for instance, in the case of an external entity who is regulating the activities of an investment firm. Perhaps the specific risk measure used within the firm is proprietary, but the assessments of individual positions are not. In this case, it may be reasonable for a regulator to attempt to approximate the underlying risk measure for the purpose of generating an estimate of the overall risk held by the firm.

In Section 5.6 we address the related problem of generating a fused risk measure based on the output of several independent, coherent risk measures. Specifically, assuming each assessor uses an arbitrary, coherent risk measure, how should the approximator generate a joint measure based on the individual measures. In this case we employ the principles of the IGCAP developed in Chapter 3 to approximate the (potentially) contradictory individual risk measures with a single fused measure.

### ■ 5.5.1 Mathematical Notation

By Proposition 5.1, any coherent risk measure can be represented as the supremum over a family of distributions of the negative expectation of the position. The representation theorem suggests that, in effect, any coherent risk measure is a 'worst-case' analysis of a set of exemplar scenarios.

We assume throughout the following two sections that any risk measure $\rho$ is coherent, and has an associated 'scenario' set under the representation theorem of $\mathcal{P}$.

We assume that each of $M$ assessors is using risk measure $\rho_i$ to assess the risk in position $X_i$. The value $\rho_i(X_i)$ will be denoted $r_i$.

**Definition 5.3.** *We say a set of assessments* $\{r_i\}$ *is* ***generated coherently*** *if* $\rho_i = \rho$ *for all* $i$.

While we would like assessments to be generated coherently, it is usually not possible to determine whether a set of assessments was generated coherently without further information. Therefore we define the following weaker condition of coherent-equivalent assessments.

**Definition 5.4.** *A set of assessments* $\{r_i\}$ *is* ***coherent-equivalent*** *if* $\exists \mathcal{P}$ *s.t.* $r_i = \sup_{Q \in \mathcal{P}} \mathbb{E}_Q[-X_i]$ *for all* $i$. *A set of assessments* $\{r_i\}$ *that are not coherent-equivalent are termed* ***incoherent.***

A coherent-equivalent set of assessments *may* have been coherently generated, or they may have been generated by a divergent set of risk measures. The important point is that there exists *some* coherent risk measure that agrees with every assessment.

For any risk assessment $r_i$ such that $r_i = \rho_i(X_i)$ we can define a linear family of possible worst case scenarios. We denote as

$$L_{r_i}(-X_i) = \{Q | \mathbb{E}_Q[-X] = r_i\}$$

the linear family of distributions which are candidates as the 'worst-case' scenario for a given risk assessment $r_i$ (with the negative in the argument reminding us that risk is an expectation over the negative of a position).

**Fact 5.1.** $L_{r_i}(-X_i)$ *is a supporting hyperplane of* $\mathcal{P}_i$

This view of $L_{r_i}(-X_i)$ as supporting hyperplanes will be used extensively throughout the following two sections.

### ■ 5.5.2 Coherent Risk Assessments

We assume in this section that the set of risk assessments were generated coherently; thus there is a single scenario set $\mathcal{P}$ used by all assessors. We show in Figure 5-1 a set of scenarios and two linear families generated by risk assessments made using the associated coherent risk measure. The arrow indicates the direction of increasing risk; it points normal to the set $\mathcal{P}$ at the point where $\mathcal{P}$ is supported by $L_{r_i}(-X_i)$.

For each $L_{r_i}(-X_i)$ define $K_i$ to be

$$K_i \triangleq \{Q | \mathbb{E}_Q[-X_i] \leq r_i\} \tag{5.13}$$

and let $K = \cap_i K_i$. For each $i$, $K_i$ is the set of points on the simplex supported by the hyperplane $L_{r_i}(-X_i)$. As such, $\mathcal{P} \subseteq K_i$ for all $K_i$ and therefore $\mathcal{P} \subseteq K$.

We now repeat the representational form of any coherent risk measure:

$$\rho(X) = \sup_{Q \in \mathcal{P}} \mathbb{E}_Q[-X]$$

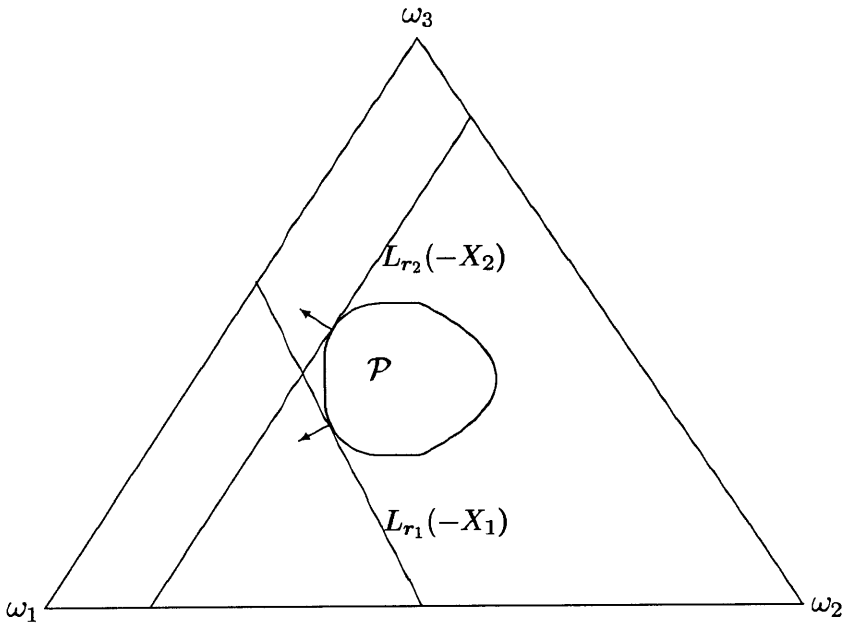



**Figure 5-1.** Linear families generated by risk assessments

Note from this definition that the aggregate risk of positions $X_1, X_2, \ldots, X_M$ is

$$\rho\left(\sum_i X_i\right) = \sup_{Q\in\mathcal{P}} \mathbb{E}_Q\left[-\sum_i X_i\right] \leq \sum_i \sup_{Q\in\mathcal{P}} \mathbb{E}_Q[-X_i] = \sum_i \rho(X_i)$$

Thus we could naively fuse the individual risk assessments into an aggregate risk simply by summing them up. Such a fusion method would be extremely conservative, however, as the following example demonstrates.

#### Risk Measure Aggregation Example

Consider an extremely simple system with $\Omega = \{\omega_1, \omega_2\}$ and two financial positions,

$$X_1 = \begin{bmatrix} -5 \\ 6 \end{bmatrix} \qquad X_2 = \begin{bmatrix} 6 \\ -5 \end{bmatrix}$$

Suppose that $\mathcal{P} = \{P | P(\omega_1) \in [0.25, 0.75]\}$. Then the individual risk of each position is

$$\rho(X_i) = \sup_{Q\in\mathcal{P}} \mathbb{E}_Q[-X_i] = 2.5$$

where the supremum is obtained for $\rho(X_i)$ at $P^* = 0.75 * \delta[\omega_i] + 0.25 * \delta[\omega_{\bar{i}}]$. If we wanted to approximate the joint risk measure by the upper bound of the sum of the individual risk measures, we would get that $\hat{\rho}(X_1 + X_2) = 2.5 + 2.5 = 5$.

However, considering the combined position $X_1 + X_2$ we see that

$$X_1 + X_2 = \begin{bmatrix} -5 \\ 6 \end{bmatrix} + \begin{bmatrix} 6 \\ -5 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

and therefore the coherent risk measure of $X_1 + X_2$ is

$$\rho(X_1 + X_2) = \sup_{Q \in \mathcal{P}} \mathbb{E}_Q[-X_1 - X_2] = -1$$

Given the positive homogeneity of coherent risk measures, by appropriately scaling the above example we can show that the sum of individual risk assessments, even when derived coherently, can be an arbitrarily bad bound of the risk of the joint position.

### ■ 5.5.3 Improved Fusion using Assessment Bounds

As suggested above, the naive aggregation method produces overly conservative assessments of the aggregate risk of a set of financial positions. An alternative method for estimating the aggregate risk would be to use $K$ as a surrogate for $\mathcal{P}$, i.e.

$$\hat{r} = \max_{Q \in K} \mathbb{E}_Q \left[ -\sum_i X_i \right] \tag{5.14}$$

We state a few facts about Equation 5.14:

**Fact 5.2.** $|\hat{r} - \rho(\sum_i X_i)| \leq |\sum_i \rho(X_i) - \rho(\sum_i X_i)|$

This states that the estimate obtained by using $K$ as a surrogate for $\mathcal{P}$ is uniformly closer to the true risk under $\rho$ than the sum of the individual risk measures is. To see this, simply note that any distribution $P$ such that

$$\mathbb{E}_P \left[ -\sum_i X_i \right] = \sum_i \mathbb{E}_P[-X_i] > \sum_i r_i$$

must have some $i'$ such that $\mathbb{E}_P[-X_{i'}] > r_{i'}$ and therefore $P \notin K_{i'}$ and therefore $P \notin K$.

**Fact 5.3.** *If* $\bigcap_i L_{r_i}(-X_i) = \emptyset$ *then* $\rho(\sum_i X_i) \neq \sum_i \rho(X_i)$.

This says that if there is no point in the intersection of all the linear families generated by the individual risk assessments, then the aggregate risk assessment is *strictly* less than the sum of the individual risk assessments. This is shown similarly to the previous fact; if the intersection is empty then there can be no distribution $P \in \text{cl}(\mathcal{P})$ that has $\mathbb{E}_P[-X_i] = r_i$ for all $i$ and therefore the inequality is strict.

**Fact 5.4.** *If* $\bigcap_i L_{r_i}(-X_i) \neq \emptyset$ *and if* $X_i \neq X_j$ *for all* $i, j$, *and* $|\Omega| = N \leq M$ *then* $\rho\left(\sum_i X_i\right) = \sum_i \rho(X_i)$.

Letting $P = \bigcap_i L_{r_i}(-X_i)$, this fact is a consequence of the fact that if $N$ intersecting hyperplanes support an $N-1$ dimensional set, then at least one supports the set at (and only at) $P$. Let $L_{r_{i'}}(-X_{i'})$ be the hyperplane. Then, since $r_{i'} = \sup_{Q \in \mathcal{P}} \mathbb{E}_Q[-X_{i'}]$ and the only point from $L_{r_{i'}}(-X_{i'}) \in K$ is $P$. And since $\mathcal{P} \subseteq K$ therefore $P \in \text{cl}(\mathcal{P})$. Then

$$\mathbb{E}_P\left[-\sum_i X_i\right] = \sum_i \mathbb{E}_P[-X_i] = \sum_i r_i.$$

which, combined with the fact that $\rho\left(\sum_i X_i\right) \leq \sum_i \rho(X_i)$, gives the result.

**Fact 5.5.** *Since* $K$ *is a polyhedron, the optimization in Equation 5.14 is a linear program.*

We have thus suggested a mechanism for estimating the joint risk of a set of positions, given a coherent-equivalent set of distributed risk assessments. We have identified the set $K$ as a superset of $\mathcal{P}$ and therefore the risk measure associated with the set of scenarios $K$ is an upperbound on the true coherent risk measure. This bound can only improve the estimation accuracy over the naive summation bound, and it is easily computable via linear programming.

## ■ 5.6 Coherent Fusion of Mutually Incoherent Risk Assessments

In this section we abandon the idea that the set of risk assessments were generated by a *single* coherent risk measure. Instead, we assume that each assessor's assessment was the result of an individual risk measure $\rho_i$ with associated set of scenarios $\mathcal{P}_i$. We first suggest, based on the development in Section 5.5.2, a method for detecting when a group of assessments were not generated coherently. We then suggest possible fusion methods, including a mechanism based in the IGCAP developed in Chapter 3.

### ■ 5.6.1 Detecting Incoherence

In Section 5.5.2 it was shown that a set of assessments from a coherent risk measure define a polyhedron that bounds the set $\mathcal{P}$ of scenarios considered by the coherent risk measure. Specifically, for each risk assessment $r_i$, Equation 5.13 defines a closed set $K_i$ of distributions that must contain $\mathcal{P}$. Defining $K = \bigcap_i K_i$ provides a closed and bounded polytope which contains $\mathcal{P}$ (presuming all risk assessments were generated coherently). This leads to the following fact.

**Fact 5.6.** *If* $K = \emptyset$ *than the set of risk assessments* $\{r_i\}$ *could not have been generated coherently.*

Figure 5-2 graphically depicts a set of incoherent risk measures. Risk assessment $r_1$ generates set $K_1$ that lies below the line $L_{r_1}(-X_1)$ while $r_2$ generates the set $K_2$ that lies 'below' (in terms of increasing level sets of risk) the line $L_{r_2}(-X_2)$. Since $K_1 \cap K_2 = \emptyset$ it is evident that the scenario sets of the two assessors are disjoint.

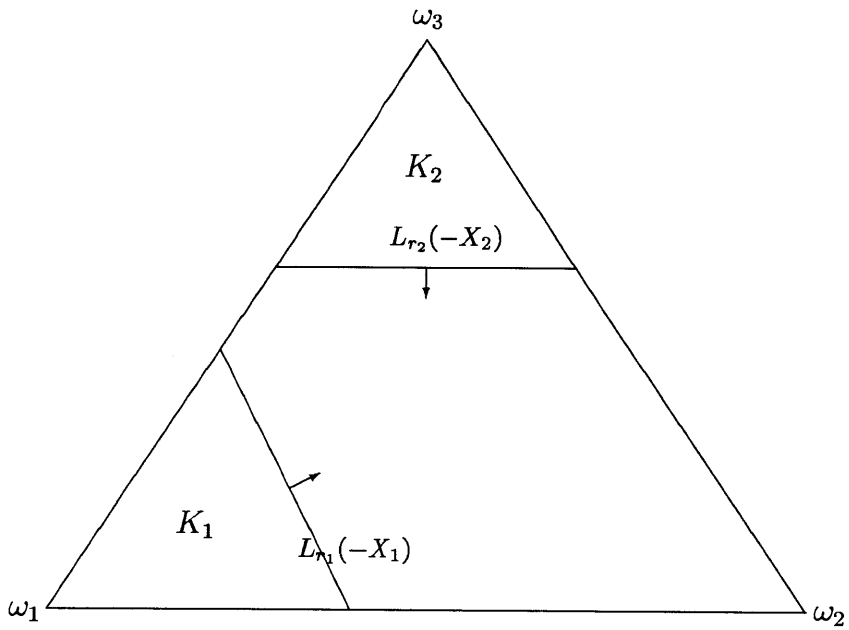



**Figure 5-2.** Incoherent risk assessments

Note that the condition $K \neq \emptyset$ is a necessary, but not a sufficient condition for the set of risk assessments to be generated coherently. Referring back to Figure 5-1, the linear families are shown as having been generated by a single set $\mathcal{P}$. But they could equivalently have been generated by any two coherent risk measures, $\rho_1$ and $\rho_2$ such that $\mathcal{P}_i \subset K$ and $\mathcal{P}_i \cap L_{r_i}(-X_i) \neq \emptyset$.

As another example, suppose $X_1 = X_2$ and $r_1 < r_2$. In this case $K = K_1 \cap K_2 = K_1 \neq \emptyset$, but it is obvious that $r_1$ and $r_2$ could not have been generated coherently. More generally, we state the following definition and fact:

**Definition 5.5.** *A linear family* $L$ dominates $L_{r_i}(-X_i)$ *if for all* $P \in L$, $\mathbb{E}_P[-X_i] > r_i$.

**Fact 5.7.** *For any* $i, j$, *if* $L_{r_i}(-X_i)$ *dominates* $L_{r_j}(-X_j)$ *then* $r_i, r_j$ *were not generated coherently.*

The reasoning is as follows: because some point in $L_{r_i}(-X_i)$ is in the closure of $\mathcal{P}$ and because *every* point $P \in L_{r_i}(-X_i)$ results in $\mathbb{E}_P[-X_j] > r_j$, there must be some point $P'$ in the closure of $\mathcal{P}$ s.t. $\mathbb{E}_{P'}[-X_j] > r_j$. Therefore $r_j \neq \sup_{Q \in \mathcal{P}} \mathbb{E}_Q[-X_j]$ and the assessment could not have been coherently generated.

In Figure 5-2, $L_{r_1}(-X_i)$ and $L_{r_2}(-X_2)$ dominate each other. Any time the linear families generated by two risk assessments are mutually dominant, the set $K$ gener-

ated by the risk assessments is empty. However it is possible to have $K = \emptyset$ without having any linear family dominate any other, as the following example demonstrates.

#### Example of mutually incoherent assessments

Suppose we have the following three positions:

$$X_1 = \begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix} \qquad X_2 = \begin{bmatrix} -1 \\ 1 \\ -1 \end{bmatrix} \qquad X_3 = \begin{bmatrix} -1 \\ -1 \\ 1 \end{bmatrix}$$

with associated risk measures $r_i = 0.2$, for all $i$. The fact that $r_1 = 0.2$ implies $\forall P \in \text{cl}(\mathcal{P})$, $P[\omega_1] \geq 0.4$ and similarly for $r_2$ and $r_3$. Since every distribution $P \in \mathcal{P}$ must have $P(\omega_i) \geq 0.4$ for all $i$, $\mathcal{P} = \emptyset$. This is shown graphically in Figure 5-3. Notice that $K_1 \cap K_2 \cap K_3 = \emptyset$ but no linear family dominates any other. We now

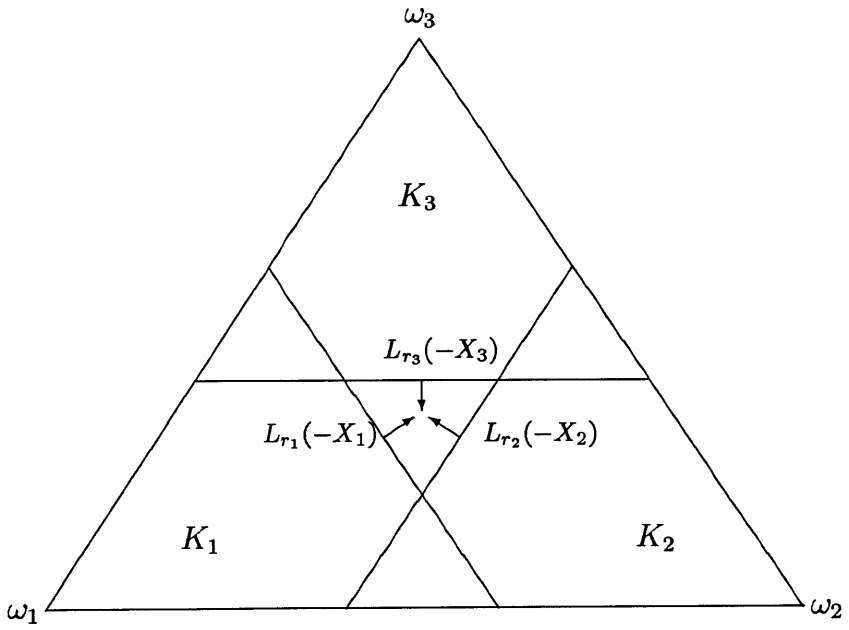



**Figure 5-3.** Non-dominated incoherent assessments

state a necessary and sufficient condition for a set of risk assessments to be coherent-equivalent (see Definition 5.4).

**Proposition 5.5.** *A set of risk assessments $\{r_i\}$ is coherent-equivalent if and only if $K \neq \emptyset$ and, for all $i, j$, $L_{r_i}(-X_i)$ does not dominate $L_{r_j}(-X_j)$.*

The only if portion is obvious. To show show the if portion we will construct a set $\mathcal{P}$ s.t. $\{r_i\}$ are the risk assessments generated by the associated $\rho$.

Let $\text{bd}(K)$ be the boundary of $K$ (note: K is closed by construction, so $\text{bd}(K) \subseteq K$ and $\text{bd}(K) \neq \emptyset$). For each $i$, define

$$\mathcal{P}_i = L_{r_i}(-X_i) \cap \text{bd}(K).$$

Since there does not exist $L_{r_j}(-X_j)$ such that $L_{r_i}(-X_i)$ dominates $L_{r_j}(-X_j)$ and since $\text{bd}(K) \neq \emptyset$, therefore $\mathcal{P}_i \neq \emptyset$. Let

$$\mathcal{P} = \bigcup_i \mathcal{P}_i$$

Then, letting $\rho$ be the coherent risk measure associated with $\mathcal{P}$, $r_i = \rho(X_i)$ for all $i$ and $\{r_i\}$ is coherent-equivalent. □

An equivalent statement of Proposition 5.5 would be that a set of assessments is incoherent if and only if at least one assessor can be shown to have *underestimated* the risk of his position, due to choosing a 'worst-case' scenario that is insufficiently 'bad,' relative to the set of all scenarios considered by the group of assessors.

### ■ 5.6.2 Fusion of Incoherent Risk Assessments

Proposition 5.5 gives us a tool to detect when there does not exist a coherent risk measure that could have led to a set of assessments. However a risk manager may still need to take the incoherent assessments and generate a fused risk assessment from them, preferably one that reflects the underlying assessments to the greatest degree possible while being coherent itself. To do this, we suggest applying the IGCAP developed in Chapter 3. The benefits of the IGCAP for coherent approximation of incoherent assessments were outlined extensively in Chapter 3 and 4.

More formally, let $\mathcal{P}^*$ denote the scenario set constructed to form $\rho^*$ which is the coherent risk measure we will use to assess the risk of the aggregate position. The IGCAP would suggest choosing $\mathcal{P}^*$ to minimize the average (or, under alternative assumptions, the maximum) divergence between the individual 'worst-case scenarios' under the coherent approximation and those under the original assessment.

We now state an adaptation of the IGCAP to the problem of coherently assessing aggregate risk based on a set of incoherent risk assessments.

$$Q^* = \underset{Q \in \Delta(\Omega)}{\arg\min} \sum_i D(L_{r_i}(-X_i)||Q) \tag{5.15}$$

where $D(L||Q) \triangleq \min_{P \in L} D(P||Q)$. Define $\mathcal{P}^* = \{Q^*\}$ and therefore

$$r = \rho^*(\sum_i -X_i) = \mathbb{E}_{Q^*}[\sum_i -X_i] = \sum_i \mathbb{E}_{Q^*}[-X_i] = \sum_i \rho^*(X_i)$$

We state the following fact about the solution mechanism suggested in Equation 5.15

**Fact 5.8.** *For any $i$ such that $L_{r_i}(-X_i)$ does not dominate any $L_{r_j}(-X_i)$, $r_i^* = \rho^*(X_i) \geq r_i$.*

Recall that a set of assessments is incoherent if and only if the 'worst-case' scenario chosen by at least one assessor is insufficiently 'bad' relative to the union of the scenarios of all individual assessors. Fact 5.8 states that the coherent fusion mechanism adopted can only *increase* the assessed risk for each position.

Equation 5.15 chooses as the ensemble of scenarios the single distribution that has minimum average divergence to the linear families defined by the individual risk assessments. To justify taking the set as a singleton, assume we let $\hat{\mathcal{P}}$ be some non-singleton set such that $Q^* \in \hat{\mathcal{P}}$. For any given $i$, $\hat{r}_i \geq r_i^*$, i.e. the 'worst-case' scenario under $\hat{\mathcal{P}}$ can only increase the assessment of the marginal risk of position $X_i$.

We thus have developed a complete mechanism for performing fusion of risk assessments in a coherent way. If the assessments are coherent-equivalent, Equation 5.14 provides an aggregate risk assessment based on a coherent risk measure that agrees perfectly with each of the individual risk assessments and makes no further assumptions on the underlying scenario set $\mathcal{P}$. If the assessments are incoherent, Equation 5.15 provides an aggregate risk assessment based on the IGCAP that minimizes the distance from the worst-case scenarios under the coherent risk measure to the families of the worst-case scenarios for each individual assessment.

### ■ 5.6.3 Potential Criticisms of Applying IGCAP to Coherent Risk Assessment

One may question why the coherent approximation should focus on minimizing the distance between scenarios rather than the assessments themselves. A more intuitive mathematical construction might choose $\hat{\mathcal{P}}$ such that the individual risk measures under the associated $\hat{\rho}$ were as close as possible to $r_i$. Such a formulation, however, suffers from non-linear scaling problems, as was pointed out in the discussion of generalizing the CAP in Chapter 3. As such, a coherent risk measure constructed to minimize the distance between $\hat{r}_i = \hat{\rho}(X_i)$ and $r_i$ would incentivize each individual risk assessor to increase the size of his position, leading to an unstable distributed assessment architecture.

Another criticism may be that the IGCAP was justified in Chapter 3 through an assessor model based on repeated observation of the random variable under assessment. Such an assessor model cannot justify the use of the IGCAP for coherent risk assessment, since the assessment is inherently not the result of a subjective expectation, but rather the outcome of a set of hypothetical scenarios. Critically, the IGCAP formulation assumed that the I-projection from $Q^*$ onto the linear families associated with each assessment was a good approximation of the subjective probability underlying each individual's assessment; this was formalized using a Conditional Limit Theorem. In this case, no such principle applies. It is not immediately clear why any point in the linear family $L_{r_i}(-X_i)$ should be favored as more likely to have been the source of the assessment than any other point.

We can partially address this concern by noting that a reasonable group of assessors would choose scenario ensembles "close" to each other, and also as limited a set of

points as possible to avoid cognitive or computational overload. As such, a reasonable estimate of the 'worst-case' scenarios are the single points from each $L_{r_i}(-X_i)$ that are 'closest' to each other. The IGCAP provides one idea of what closest might be, but unlike in the case of subjective expectation assessments, it shouldn't be treated as more reasonable than other, alternative distance definitions of close.

## 5.7 Conclusion

In this chapter we have analyzed the coherence properties of a specific example of distributed assessment, specifically the case of financial risk measurement. We presented several results related to the axiomatic classes of coherent and convex risk measures. First we showed that the minimal convex extension of the popular $\text{VaR}_\gamma$ risk measure is, itself, a VaR risk measure (usually with a different parameter, $\gamma^*$). Then we introduced a concept of risk measure robustness to divergence of opinion among assessors and analyzed the robustness properties of several popular risk measures. We next applied the principles of coherent approximation to the problem of distributed coherent risk measurement when risk assessments could have been generated coherently. Finally, we employed IGCAP to generate a combined risk assessment based on a set of distributed risk assessments which could not have been generated coherently.

There are several avenues of future work suggested by the foregoing analysis. The definition of robustness offered in the chapter as a uniform linear bound on the change in risk measurement under opinion divergence is not unique. Several other types of robustness could be considered. Furthermore, even under the current definition a rate analysis could be undertaken to determine which risk measures' differences decay fastest to zero with the opinion divergence. In the case of coherent approximation of risk measurements, we have offered a method for defining an approximation vector that is jointly coherent, but we have not suggested a method for mapping the approximation vector into a single-valued risk measure over the entire set of positions. It is our hypothesis that an appropriately designed fusion rule will exhibit a separation principle, such that the fused risk measure can be computed via a two-stage process of first coherently approximating the risk measurements and then applying a transform on the resulting vector to map it into the real numbers.

# Chapter 6

# Knightian risk and outcome indeterminacy

## 6.1 Introduction

In a seminal work, Frank Knight [144] distinguished between *risk* and *uncertainty*.

> Uncertainty must be taken in a sense radically distinct from the familiar notion of Risk, from which it has never been properly separated.... The essential fact is that 'risk' means in some cases a quantity susceptible of measurement, while at other times it is something distinctly not of this character; and there are far-reaching and crucial differences in the bearings of the phenomena depending on which of the two is really present and operating.... It will appear that a measurable uncertainty, or 'risk' proper, as we shall use the term, is so far different from an unmeasurable one that it is not in effect an uncertainty at all.

In essence, Knight suggests that there are two qualitatively different types of uncertainty: risk, which relates to foreseeable but unpredictable events, and uncertainty, which relates to unforeseeable events. Knight's view on the subject were so influential that in economics literature the phenomenon of 'unmeasurable' uncertainty has come to be known as 'Knightian uncertainty' (c.f. [145–147]).

Knight isn't alone in formulating this difference. Working independently but at the same time, Keynes [59] based his theory of probability on a similar differentiation between predictable and unpredictable uncertainty. The previously cited work of Anscombe and Aumann [64] made a similar distinction. More recently, the Dempster-Shafer theory of evidence [148–150] can be seen as an attempt to capture mathematically the difference between risk and uncertainty.

One provocative way of conceptualizing the concept of Knightian uncertainty is as a fundamental limitation on the knowability of a probability. This concept will be developed somewhat in Section 6.4.2, but the basic idea is that a useful way of viewing inconsistency among experts is not as some subset of the experts behaving irrationally, but rather as indicating in a very specific way the limits of the knowability of the outcomes in question.

## 6.2 Structural Uncertainty

Thus far in the thesis we have investigated the idea of coherent approximation from the perspective of 'correcting' experts' assessments. Alternatively, we could consider the assessments as correct, but the structure as being insufficiently expressive to capture the assessed behavior. This could come either from an inexperienced approximator, who assumes a structure insufficiently complex to cover the contingencies considered by the assessors, or it could come from a fundamental inability to model *a priori* observations as deterministic mappings from atomic events to an alphabet.

#### Example of an inexperienced approximator

Suppose a political neophyte wished to use expert assessments to determine the probabilities of Democratic or Republican victory in the 2012 U.S. Presidential election. He considers as atomic events the possibility that each of several contenders 1) becomes their party's nominee and 2) wins the general election. For example, event $\omega_1$ corresponds to Barack Obama being the nominated (D), Mitt Romney being the nominated (R) and Barack Obama winning; event $\omega_2$ corresponds to Obama and Romney being nominees, but Romney winning; and so forth. For the events '(D) wins the election' and '(R) wins the election', the outcome matrix looks like:

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 1 & \dots & 0 \\ 0 & 1 & 0 & \dots & 1 \end{bmatrix}$$

The only unique columns in the matrix are $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$.

Our approximator now seeks expert input in the form of aggregate survey information. The result, event $A_1$ is deemed to have probability 0.55 and event $A_2$ is deemed to have probability 0.42. Given his assumed structure, these assessments are incoherent, so the approximator uses his favorite coherent approximation algorithm to 'correct' the assessments. However, unknown to the approximator, in the aggregate surveys the event 'Neither (R) nor (D)' received weight of 0.03. This atomic event, while considered by the assessors, was not considered by the approximator. In this case the incoherence was not fundamentally a property of the assessors, but rather of the insufficient structure assumed by the approximator.

#### Example of non-deterministic mappings

Suppose the phenomenon under observation and assessment is an electron of a hydrogen atom. In this case, the atomic events are the various quantum states of the electron, and the observables are its position and its momentum. There does not exist a deterministic mapping from quantum states to observables. Instead, each quantum state corresponds to a density of possible observations, all of which may occur no matter how precisely controlled the experiment is.

In this case, the relationship between observables (against which assessments are made) and pure states of the system is necessarily stochastic, meaning that the out-

come matrix $\mathbf{X}$ needs to be generalized to include ranges of outcomes for single atomic states.

### 6.2.1 Uncertainty and Indeterminacy

The two examples above represent two distinct ways in which an assumed structure may be deficient: it may fail to include possibilities considered by the assessors, and it may consider relationships between atomic states and assessed variables to be deterministic when they are not. I will refer to these circumstances as *uncertainty* and *indeterminacy* respectively.

In both cases, if assessments are incoherent w.r.t. an assumed structure, the problem remains of how to make the assessment coherent. In Chapters 3 and 4 we developed a framework in which the assessment was revised. In this Chapter we discuss possible methods to revise the structure.

The cases of uncertainty and indeterminacy represent two related, but distinctly different, ways in which the assumed structure requires revision. In the case of uncertainty, the challenge is to introduce new columns into the outcome matrix representing the additional outcomes considered by the expert assessors. In the absence of actual knowledge of which outcomes were considered, we adopt a maximum sparsity approach, in which we wish to minimize the number of additional columns introduced, subject to the constraint that the given assessment be coherent w.r.t. the final matrix.

In the case of indeterminacy we also adopt a sparsity approach, but in this case it is on the ranges of outcomes mapped from each input state to each observable. This range corresponds to the support of the density of the observable in each given state.

### 6.2.2 Background

Much has been written about Knightian uncertainty, although the association of such uncertainty with structural insufficiency is, to the best of my knowledge, unique to the current development. The concept of Knightian uncertainty has been applied to principles of asset pricing [145,146], labor markets [147], stock volatility [151] and many more. A closely related concept of 'ambiguity' has a similarly extensive development history in economics, largely stemming from the influential work of Ellsberg [152].

Indeterminacy has been investigated at length in the philosophical literature, including epistemology [153], linguistics [154,155], legalism [156,157], and philosophy of science [158,159]. The impact of indeterminacy as relates to the foundations of decision theory was explored in a fascinating paper by Churchman [160] where he argues that indeterminacy in value measurements, or more generally preference rankings, shakes the foundations of statistical decision theory laid out by Savage and others.

The problem of indeterminacy did not escape de Finetti, who included a lengthy discussion of it in an appendix to his Theory of Probability [26]. In Section 4 of the appendix he discusses situations in which the differentiation between possible and impossible are unclear, similar to the understanding of Knightian uncertainty we proposed above. Then, in Section 9, he discourses on the interaction of indeterminacy and verifiability, and then (in subsequent sections) suggests ways of reconciling his

development of probability theory with the indeterminacy inherent in the field of quantum mechanics, particularly the Heisenberg model.

## 6.3 Minimal Structural Revision to Induce Coherence

We first analyze the case when assessment incoherence is induced by an incomplete output matrix, identified in the introduction with 'uncertainty.' We first analyze the specific example of characteristic matrices, and then turn our attention to minimal revision of non-characteristic output matrices. We measure the amount of structural revision by the number of additional columns introduced in the output matrix, and analyze the viability of various algorithms for choosing the minimal structural revision such that the assessment becomes coherent.

### 6.3.1 Minimal Revision of Characteristic Matrices

In the example of the inexperienced approximator, the structure was seen to be insufficient to describe the set of experiences represented by the expert assessors due to the exclusion of certain outcomes from the considered set. The solution would be to alter the structure by inserting additional outcomes until the assessment can be represented as a convex combination of the set of considered outcomes. In the example, introducing the output 'neither Democrat nor Republican wins' would immediately induce coherence.

In the context of characteristic random variables, for any assessment vector of length $M$ there are at most $2^M$ possible output vectors that could be considered. The simplest solution would be to introduce all $2^M$ possible outcomes as columns of the characteristic matrix. In this way *any* assessment would be coherent. However, by introducing all possible outcomes, all situational structure is eliminated, resulting in a description of the assessment situation that is unnecessarily vague.

One might question the use of the term 'unnecessarily' in the above statement; could it not be that it is necessary to introduce all possible output vectors in order to introduce coherence? The substance of the Carathéodory Theorem states that if a point in $\mathbb{R}^M$ lies in the convex hull of some set of points, there exists a subset of size at most $M+1$ such that the point is representable as a convex combination of the subset. In this sense, simply eradicating all structure by introducing all possible output vectors is an unnecessary step.

Instead we suggest the following formulation. Consider $\underline{\mathrm{X}}$ to be the matrix of all $2^M$ possible output vectors (i.e. all binary sequences of length $M$). Let $\lambda$ be some vector of convex weights of length $2^M$ and let $S$ be some subset of the indices of $\lambda$ (i.e. $S \subset \{1,2,\ldots,2^M\}$) with $\bar{S} = \{1,2,\ldots,2^M\} \backslash S$. The structural approximation problem under uncertainty can be equivalently stated as:

$$\begin{aligned} \lambda^* &= \arg\min_\lambda ||\lambda_{\bar{S}}||_0 \\ &\text{s.t. } P = \underline{\mathrm{X}}\lambda \end{aligned} \tag{6.1}$$

where $||\lambda_{\bar{S}}||_0$ denotes the number of non-zero elements in $\lambda_{\bar{S}}$.

This problem formulation bears a striking resemblance to the problems of matching pursuit, basis pursuit and compressive sensing. In general, estimating sparse vectors from underdetermined systems of equations has proven hard without additional structure, such as the restricted isometry property (RIP). A general approach has been to employ the $L_1$ norm as a surrogate for the $L_0$ norm, with special problem structure determining when the $L_1$ relaxation provides an optimal solution to the sparse estimation problem. Another highly relevant problem formulation is that of Jagabathula and Shah [161,162], with the primary difference being the nature of the 'observation' matrix.

### ■ 6.3.2 Minimal Bases for Probabilistic Assessments

We turn our attention to a relaxed version of Equation 6.1. After a brief development, we will return to the original problem with a suggested suboptimal solution mechanism.

Consider the simpler problem of selecting the smallest possible set of vectors from $\underline{X}$ to represent $P$. Mathematically,

$$\begin{aligned} \lambda^* &= \arg\min_\lambda ||\lambda||_0 \\ \text{s.t. } P &= \underline{X}\lambda \end{aligned} \tag{6.2}$$

This is equivalent to the formulation in Equation 6.1 for the case where $S = \emptyset$. Let $N^* = ||\lambda^*||_0$.

While we do not undertake a complexity analysis of the problem here, it is not unreasonable to assume that, given the general challenges of sparse basis estimation, an optimal solution to Equation 6.2 is not easily computable. We therefore suggest an iterative algorithm to suboptimally approximate the solution to Equation 6.2. It should be noted that while this algorithm was independently derived, it bears a strong resemblance to that of [161], and the development was likely influenced by exposure to the techniques used therein.

In the algorithm definition, let the function min : $[0,1]^M \rightarrow [0,1]$ be understood to be the minimum of all *non-zero* entries of the vector, and the functional supp : $[0,1]^M \rightarrow \{0,1\}^M$ be the 'support' of the vector (i.e. if $s = \text{supp}(P)$, $s_i = 1$ iff $P_i > 0$).

*pre* $P_0 = P$

1. $X_t = \text{supp}(P_t)$
2. $\hat{\lambda}_t = \min(P_t)$
3. $P_{t+1} = P_t - X_t\hat{\lambda}_t$

*post* If $\sum_i \hat{\lambda}_i \neq 1$, $X_{T+1} = 0$ and $\hat{\lambda}_{T+1} = 1 - \sum_{i=1}^{T} \hat{\lambda}_i$

Let $\hat{N} = ||\hat{\lambda}||_0$. Note, to avoid notational complexity, we have written the above algorithm in such a way that $\hat{\lambda}$ is a dense vector of length $T$ or $T+1$ associated with a specific set of vectors $\{X_i\}_{i=1}^{\hat{N}}$, rather than a sparse vector of length $2^M$. An

equivalent, though more cumbersome, statement could be made in terms of generating a sparse vector to right multiply $\underline{X}$.

This algorithm has several attractive properties. Let $\bar{N}$ be the number of unique elements of $P$.

**Fact 6.1.** $\hat{N} \leq \bar{N} + 1 \leq M + 1$.

Thus, the algorithm never exceeds the minimum upper bound given by the Carathéodory Theorem. Furthermore, the sparsity determined by the algorithm can be ascertained *a priori*.

One might question how close to optimal the given algorithm is. For most $P \in [0,1]^M$ the algorithm is optimal.

**Fact 6.2.** *Taking $\mu$ as the Lebesgue measure,* $\mu(\{P : N^* \neq \hat{N}\}) = 0$.

The only points that are representable in fewer than $M+1$ points are, by definition, those which lie on lower dimensional hyperplanes. Let

$$\mathcal{Q} \triangleq \{Q \in [0,1]^M : \exists \lambda \text{ s.t. } Q = \underline{X}\lambda, ||\lambda||_0 \leq M\}.$$

Then, letting $\mathcal{P} \subseteq \mathcal{Q}$ be the set of points for which $\hat{N} = N^*$, we have

$$\mu(\mathcal{Q} \backslash \mathcal{P}) \leq \mu(\mathcal{Q}) \leq \sum_{i=1}^{K} \mu(\mathcal{Q}_i) = \sum_{i=1}^{K} 0 = 0$$

where $\mathcal{Q}_i$ is the set of points representable by a particular subset of basis vectors with cardinality less than $M + 1$, and $K$ is the total number of all such subsets (i.e. $K = \sum_{i=1}^{M} \binom{2^M}{i}$). The measure of each set $\mathcal{Q}_i$ is zero because it is contained on a lower-dimensional manifold.

Taken together, these facts seem to indicate that the algorithm performs reasonably well. However, it is also possible to show that in certain highly structured cases, the algorithm can generate a $\lambda$ with exponentially less sparsity than the optimal solution, as the following example demonstrates.

#### Example of algorithm failure

Consider the assessment vector $P$ with optimal (in terms of Equation 6.2) decomposition

$$P = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \lambda^*$$

where $\lambda^*$ is a vector of convex weights of length 3.

Assuming $\lambda_1^* > \lambda_2^* > \lambda_3^*$ but $\lambda_2^* + \lambda_3^* > \lambda_1^*$, it is not hard to see that the given algorithm will decompose $P$ into the matrix

$$P = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \hat{\lambda}.$$

Generalizing this example, for certain highly-structured assessments with sparse representations the decomposition determined by the algorithm can have $\hat{N} = 2^{N^*}$.

Two further possible questions to be answered about the suggested algorithm include analyzing the performance when $P$ is constrained to be an element of a discretized lattice, rather than any point within the hypercube, and determining whether handling highly structured cases such as in the example above could improve the theoretical algorithm performance (i.e. could the worst-case performance be moved from exponential to polynomial by handling a few degenerate cases).

In terms of the original problem, it's important to remember that the current algorithm is a suboptimal solution to a special case of the problem of interest.

## 6.4 Structural Revision in Cases of Indeterminacy

In this section we analyze the case when incoherence is due to non-deterministic mapping between atomic states and realizations of random variables. A minimality principle is formulated and related to the dual formulation of the CAP and the IGCAP defined in Chapter 3.

### 6.4.1 Defining Non-deterministic Output Matrices

We begin by formulating a concept of non-deterministic output 'matrices.' By a non-deterministic output matrix, we mean that each element of the matrix $\mathbf{X}$ is an interval of values, each being a possible output of the system under the given atomic state. For example, whereas before $\mathbf{X}_{ij}$ was equal to $X_i(\omega_j)$, now $\mathbf{X}_{ij}$ is a set-valued mapping from $\Omega$ to the Borel sigma-algebra over the interval $[\min \mathcal{X}_i, \max \mathcal{X}_i]$. Alternatively, one can view these intervals as sets of values which could possibly be observed as $X_i(\omega_j)$. Conceptually, non-deterministic output matrices can be viewed as a relaxation of a deterministic output matrix in which the mappings are indeterminant, for instance due to problems of representational precision or fundamental limits of observation.

One might question the utility of introducing non-deterministic output matrices. Defining random variables whose outcomes aren't deterministic mappings from $\Omega$

would indicate to some that $\Omega$ is improperly defined. Taking the quantum mechanical example from above, the critique would be that the quantum states of the hydrogen electron should not be taken as equivalent to the 'atomic' states of the system, exactly *because* they don't lead to a deterministic mapping from atomic events to observable outcomes. This critique is certainly valid, however we suggest that in some cases, such as the quantum mechanical example, no meaningful deterministic definition of atomic states *can* occur prior to observation. Because of the inherent uncertainty created by the impact of observation on the system itself, defining deterministic mappings from atomic events to random variables requires a posterior definition of atomic states *in terms* of observations. Such circularity seems problematic.

### ■ 6.4.2 Dual Program for the CAP

In the next two subsections we will develop the dual programs for both the CAP and IGCAP introduced in Chapter 3. We will first analyze the CAP under both Lagrangian and Fenchel duality in order to get a sense of the geometry, and then we will analyze a specific instance of the IGCAP under Fenchel duality. The purpose of these sections is to suggest a method for relaxing a given characteristic matrix into a non-deterministic matrix such that an assessment $P$ is coherent with respect to the relaxed structure.

Recall the original formulation of the CAP from Chapter 3. For a given assessment vector $P \in [0,1]^M$ we have:

$$\begin{aligned} Q^* = \arg\min_{Q \geq 0} \textstyle\sum_i (Q_i - P_i)^2 \\ \text{s.t. } Q \in \texttt{convhull}(\chi) \end{aligned} \tag{6.3}$$

Since $\texttt{convhull}(\chi)$ is a polyhedral set contained in the unit hypercube, we can find an $(A, b)$ pair such that the optimization problem can be equivalently written as

$$\begin{aligned} Q^* = \arg\min_{Q \geq 0} \textstyle\sum_i (Q_i - P_i)^2 \\ \text{s.t. } AQ = b \end{aligned} \tag{6.4}$$

This is simply a quadratic optimization problem, albeit with a potentially exponential number of linear constraints (although in certain special cases, as in the example below, the number of constraints can be quite small). Let $P^*$ representing the solution to Equation 6.3 (identically, Equation 6.4).

#### Example of defining $A$ to cast CAP as a standard quadratic program

Consider the simple system $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ with $A_i = \omega_i$ for $i = 1, 2, 3$. For this system we have

$$\chi = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{6.5}$$

yielding the coherent set upper bounded by the simplex and lower bounded by the origin. Thus $\texttt{convhull}(\chi)$ is the intersection of the halfplane defined by the pair

$A = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$, $b = 1$ and the positive orthant. The optimization problem can thus be recast into the standard quadratic optimization framework, in this case with a single linear constraint.

### ■ 6.4.3 Lagrangian Dual

It is informative to consider the dual formulation of this problem. A simple Lagrangian analysis gives

$$\begin{aligned}
||Q^* - P||_2^2 &= \min_{Q \geq 0} \max_{\lambda \geq 0} ||Q - P||_2^2 + \lambda^T (AQ - b) \\
&= \max_{\lambda \geq 0} \min_{Q \geq 0} ||Q - P||_2^2 + \lambda^T (AQ - b) \\
&= \max_{\lambda \geq 0} || - \frac{1}{2} A^T \lambda + P - P||_2^2 + \lambda^T \left( A(-\frac{1}{2} A^T \lambda + P) - b \right) \\
&= \max_{\lambda \geq 0} \frac{1}{4} \lambda^T A A^T \lambda - \frac{1}{2} \lambda^T A A^T \lambda + \lambda^T (AP - b) \\
&= \max_{\lambda \geq 0} \lambda^T (AP - b) - \frac{1}{4} \lambda^T A A^T \lambda
\end{aligned}$$

where the first equality is due to Lagrange multipliers, the second is due to the saddlepoint theorem, the third is due to the convexity in $p$ and the fourth and fifth are just algebra.

A few observations about the dual problem

- The dual objective is ellipsoidal in the dual space
- The first term (which we want to make large) corresponds to an inner product between $\lambda$ and the error vector $AP - b$
- The second term (which we want to make small) is a measure of the size of $A^T \lambda$

### ■ 6.4.4 Fenchel Dual

The basic equivalence of Lagrange and Fenchel duality means we will get similar results by considering the Fenchel dual. However, the geometric meaning of the dual which will factor significantly in our development, is potentially more obvious in using Fenchel duality. As such, we replicate the dual program, this time using Fenchel duality theory. This development is derived from [163], although other excellent treatments can be found in [164] or [165].

Using the vector notation, we define the objective function of the CAP as $f(P) = (Q - P)^T (Q - P)$. We can write the conjugate dual function as

$$f^*(\phi) = \sup_Q \left\{ \phi^T Q - (P - Q)^T (P - Q) \right\}$$

Optimizing, we have

$$Q^* = \frac{\phi}{2} + P$$

which leads to

$$f^*(\phi) = \frac{\phi^T\phi}{2} + \phi^T P - \frac{\phi^T\phi}{4} = \phi^T P + \frac{\phi^T\phi}{4}$$

Applying this to Corollary 3.3.11 from [163] we get a functional form of the dual problem of

$$\begin{aligned}\sup_{\lambda\geq 0}\left\{\lambda^T b - f^*(A^T\lambda)\right\} &= \sup_{\lambda\geq 0}\left\{\lambda^T b - \lambda^T AP - \frac{\lambda^T AA^T\lambda}{4}\right\} \\ &= \sup_{\lambda\geq 0}\left\{\lambda^T(b - AP) - \frac{\lambda^T AA^T\lambda}{4}\right\}\end{aligned}$$

which is equivalent to the Lagrangian dual formulation.

### ■ 6.4.5 A Geometric Interpretation of the Dual

Under the dual formulation, we solved for the optimal $P$ for a given $\lambda$. Specifically we found that

$$P^*(\lambda) = P - \frac{1}{2}A^T\lambda \tag{6.6}$$

This generates a geodesic in the primal space, originating at $P$ and parameterized by $\lambda$.

#### Example Cont.

Returning to the system in the example above, for any $P$ the geodesic defined by Equation (6.6) will have form

$$P^*(\lambda) = P - \frac{1}{2}\begin{bmatrix}1\\1\\1\end{bmatrix}\lambda = P - \begin{bmatrix}\frac{\lambda}{2}\\ \frac{\lambda}{2}\\ \frac{\lambda}{2}\end{bmatrix}$$

which represents a line segment originating at $P$ with slope equal to the negative normal of the constraint surface.

#### Geometry, Inside Out

As seen in Equation (6.6), one way of viewing the dual problem is as a movement from $P$ along a subspace defined by $A^T$ and parameterized by $\lambda$. Equivalently, we could consider the problem of expanding the constraint set until the point at which it includes $P$. Given that $AP^* \leq b$ by definition, we have

$$A\left(P - \frac{1}{2}A^T\lambda^*)\right) = b_i$$

or

$$AP = b_i + \frac{1}{2}AA^T\lambda^* = \hat{b_i}(\lambda^*)$$

Thus we can define a new set of constraints $\hat{b}$ such that the original point $P$ is coherent w.r.t. the updated constraints.

In terms of the original formulation of the problem as a projection into the coherent hull of the characteristic matrix, this inside-out interpretation would do the following: rather than events $\mathcal{E}$ being absolutely deterministic w.r.t. the atomic events $\Omega$, the events become indeterminant, meaning that there is some range of possible 'membership' values of each atomic event in the events under assessment. This range represents the set of possible interpretations of the assessment events w.r.t. the atomic events.

### ■ 6.4.6 Dual Program for the IGCAP

In the previous section we suggested a method for generating a non-deterministic output matrix based on the duality formulation of the original CAP. In this section we provide a similar analysis based on the IGCAP from Chapter 3.

Recall the mathematical definition of the IGCAP:

$$Q^* = \arg\min_{Q \in \Delta} \sum_{i=1}^{N} \min_{\pi \in L_{P_i}(X_i)} D(\pi||Q) \tag{6.7}$$

with $P^* = \mathbf{X}Q^*$ and $L_{P_i}(X_i)$ denoting the linear family generated by assessment $P_i$ of random variable $X_i$. Equivalently, we can write the IGCAP as a linearly constrained convex optimization problem:

$$\begin{array}{l} Q^* = \arg\min_{Q \geq 0} \sum_{i=1}^{N} \min_{\pi \in L_{P_i}(X_i)} D(\pi||Q) \\ \text{s.t. } AQ = 1 \end{array} \tag{6.8}$$

where $A = \mathbf{1}^M$ (a row vector of length $M$, uniformly equal to 1).

Let $f(Q) = \sum_{i=1}^{N} \min_{\pi \in L_{P_i}(X_i)} D(\pi||Q)$. Recall that if $X_i$ is a characteristic random variable for all $i$, then

$$f(Q) = \sum_{i=1}^{N} D_b(P_i||\chi_i Q)$$

where $\chi_i$ is the $i^{th}$ row of the characteristic matrix and $D_b(\cdot||\cdot)$ is the binary I-divergence. Assuming our random variables are characteristic, we compute the conjugate function as

$$\begin{aligned} f^\star(\phi) &= \sup_{Q \geq 0} \{\phi^T Q - f(Q)\} \\ &= \sup_{Q \geq 0} \left\{ \phi^T Q - \sum_i \left( P_i \log \frac{P_i}{\chi_i Q} + (1 - P_i) \log \frac{1 - P_i}{1 - \chi_i Q} \right) \right\} \end{aligned}$$

Using dual variable $\lambda$ we can write the dual program as

$$\sup_{\lambda \geq 0} \left\{ \lambda - f^{\star}(A^T \lambda) \right\} \tag{6.9}$$

For $\phi = A^T \lambda$ we have

$$f^{\star}(A^T \lambda) = \sup_{Q \geq 0} \left\{ \lambda \sum_i Q_i - \sum_i \left( P_i \log \frac{P_i}{\chi_i Q} + (1 - P_i) \log \frac{1 - P_i}{1 - \chi_i Q} \right) \right\}$$

Taking the derivative w.r.t. $Q_i$ we have

$$\begin{aligned} \frac{\partial f^{\star}(A^T \lambda)}{\partial Q_j} &= \lambda - \sum_{\{i : \chi_{ij} = 1\}} \frac{\partial D_b(P_i || \chi_i Q)}{\partial Q_j} \\ &= \lambda - \sum_{\{i : \chi_{ij} = 1\}} \frac{-P_i}{\chi_i Q} + \frac{1 - P_i}{1 - \chi_i Q} \\ &= \lambda - \sum_{\{i : \chi_{ij} = 1\}} \frac{-P_i + \chi_i Q}{\chi_i Q (1 - \chi_i Q)} \end{aligned}$$

where $\chi_i$ is the $i^{th}$ row of the characteristic matrix $\chi$. We wish to set this system of equations equal to zero in order to solve for the optimal $Q^*$ in terms of $\lambda$, but the coupling due to $\chi$ makes the problem algebraically intractable. Instead, assume $\chi = I$; then, simplifying the above equations and setting equal to zero yields:

$$\lambda + \frac{P_i - Q_i^*}{Q_i^* (1 - Q_i^*)} = 0$$

and solving using the quadratic equation gives

$$Q_i^* = \frac{P_i - \lambda}{2\lambda} \pm \frac{\sqrt{(P_i - \lambda)^2 + 4\lambda P_i}}{2\lambda}$$

We can't have $Q_i^* < 0$, so the lower solution is spurious, resulting in (after some simplification)

$$Q_i^* = \frac{P_i - \lambda}{\lambda} + \frac{\sqrt{(P_i - \lambda)^2 + 4P_i}}{\lambda} \tag{6.10}$$

We could use Equation 6.10 to complete the calculation of the conjugate, and use that to finish formulating the dual, but for our current purposes Equation 6.10 is sufficient. We interpret $Q_i^*$ as a parameterized function of the dual variable $\lambda$, just as we did $P_i^*$ in Equation 6.6. In this case, for $\lambda = 1$, $Q_i^* = P_i$. Thus the optimizing $Q^*$ starts at $P$ and moves a uniform amount along the geodesics defined by equation 6.10.

In terms of the question of structural revision, the structure of the output matrix should be relaxed along the geodesics defined in Equation 6.10. Given that to get an algebraic expression we had to assume that $\mathbf{X} = I$, the suggested non-deterministic

output 'matrix' is thus

$$\mathbf{X} = [x_{ij}] = \begin{cases} [\beta_i(\lambda^*)\ 0] & i = j \\ 0 & \text{o.w.} \end{cases}$$

where the functions $\beta_i$ are determined by Equation 6.10.

### 6.4.7 Dual Program for Reversed-Divergence Cost Function

In the previous subsection we derived a parameterized expression for $Q^*$ when $\mathbf{X} = I$ under the IGCAP, which amounts to the optimization problem

$$Q^* = \arg\min_{Q \in \Delta} \sum_{i=1}^{M} D_b(P_i||Q_i) \tag{6.11}$$

The structure of the IGCAP was driven by a particular operator model, and discussed at some length in Chapter 3. However, for illustrative purposes, we here consider the sister program to the above equation, reversing the order of $P$ and $Q$ under the binary divergence.

$$Q^* = \arg\min_{Q \in \Delta} \sum_{i=1}^{M} D_b(Q_i||P_i) \tag{6.12}$$

Equation 6.12 is identical to Equation 6.11 except the variable order under the divergence has been reversed. As with the IGCAP derived cost function, we can write the dual program as

$$\sup_{\lambda \geq 0} \{\lambda - f^\star(A^T\lambda)\}$$

where, as before $A = \mathbf{1}^M$ and $f^\star(\cdot)$ is now the conjugate of the cost function under the reversed ordering. Writing this out we have

$$f^\star(A^T\lambda) = \sup_{Q \geq 0} \left\{ \lambda \sum_i Q_i - \sum_i Q_i \log \frac{Q_i}{P_i} + (1 - Q_i) \log \frac{1 - Q_i}{1 - P_i} \right\}$$

Taking the derivative w.r.t. $Q_j$ we get

$$\begin{aligned} \frac{\partial f^\star(A^T\lambda)}{\partial Q_j} &= \lambda - \left( \log \frac{Q_j}{P_j} + Q_j \frac{P_j}{Q_j} \frac{1}{P_j} - \log \frac{1 - Q_j}{1 - P_j} + (1 - Q_j) \frac{1 - P_j}{1 - Q_j} \frac{-1}{1 - P_j} \right) \\ &= \lambda - \log \frac{Q_j}{P_j} + \log \frac{1 - Q_j}{1 - P_j} \\ &= \lambda - \log \frac{Q_j(1 - P_j)}{(1 - Q_j)P_j} \end{aligned}$$

Setting this equal to zero and solving for $Q_j^*$ we get

$$\begin{aligned} \lambda - \log \frac{Q_j^*(1-P_j)}{(1-Q_j^*)P_j} &= 0 \\ \frac{Q_j^*(1-P_j)}{(1-Q_j^*)P_j} &= e^{\lambda} \\ Q_j^* = \frac{\frac{P_j}{1-P_j}e^{\lambda}}{1+\frac{P_j}{1-P_j}e^{\lambda}} &= \frac{P_j e^{\lambda}}{(1-P_j)+P_j e^{\lambda}} \end{aligned} \tag{6.13}$$

Note that the geodesics defined by the cost function with reversed order under divergence are the binary tilted exponential families with bases $P_i$. We thus have, just as in the case of the CAP and the IGCAP, the interpretation of $Q^*$ as moving a uniform distance $\lambda$ along the geodesics (in this case defined by the exponential families) from $P$.

Before we finish the computation of the conjugate function, we consider the similarities between Equation 6.13, Equation 6.10 and Equation 6.6. In all three, a family of binary distribution parameters are defined, with one element of the family being the assessed value $P_i$. The families defined in Equations 6.6 and 6.13 are both well known, but the form of Equation 6.10 is not familiar from the literature. The development suggests it is an analog, in some limited sense, to the exponential family. Where the exponential family naturally occurs under optimization of the first variable in the divergence, the family defined in Equation 6.10 occurs under optimization of the second variable. The relationship here is merely suggestive, but might have interesting implications.

Returning to the development of the dual program for the reversed-divergence

cost function, we can use Equation 6.13 to finish writing the conjugate function.

$$
\begin{aligned}
f^{\star}(A^T\lambda) &= \lambda\sum_i Q_i^* - \sum_i Q_i^* \log\frac{Q_i^*}{P_i} + (1-Q_i^*)\log\frac{1-Q_i^*}{1-P_i} \\
&= \lambda\sum_j \frac{P_j e^{\lambda}}{(1-P_j)+P_j e^{\lambda}} - \sum_j \frac{P_j e^{\lambda}}{(1-P_j)+P_j e^{\lambda}} \log\frac{\frac{P_j e^{\lambda}}{(1-P_j)+P_j e^{\lambda}}}{P_j} \\
&\quad -\sum_j \left(1 - \frac{P_j e^{\lambda}}{(1-P_j)+P_j e^{\lambda}}\right) \log\frac{1-\frac{P_j e^{\lambda}}{(1-P_j)+P_j e^{\lambda}}}{1-P_j} \\
&= \sum_j \frac{\lambda P_j e^{\lambda}}{(1-P_j)+P_j e^{\lambda}} - \sum_j \frac{P_j e^{\lambda}}{(1-P_j)+P_j e^{\lambda}} \log\frac{e^{\lambda}}{(1-P_j)+P_j e^{\lambda}} \\
&\quad -\sum_j \left(\frac{1-P_j}{(1-P_j)+P_j e^{\lambda}}\right) \log\frac{1}{(1-P_j)+P_j e^{\lambda}} \\
&= -\sum_j \frac{P_j e^{\lambda}}{(1-P_j)+P_j e^{\lambda}} \log\frac{1}{(1-P_j)+P_j e^{\lambda}} \\
&\quad -\sum_j \left(\frac{1-P_j}{(1-P_j)+P_j e^{\lambda}}\right) \log\frac{1}{(1-P_j)+P_j e^{\lambda}} \\
&= \sum_j \log\left[(1-P_j)+P_j e^{\lambda}\right]
\end{aligned}
$$

This gives a dual program of

$$
\sup_{\lambda \geq 0} \left\{ \lambda - \log \prod_j \left((1-P_j)+P_j e^{\lambda}\right) \right\}
$$

### ■ 6.4.8 Import of Relaxed Structure in Cases of Indeterminacy

Viewing the problem of coherent approximation as a minimal expansion (relaxation) of the constraint set rather than a projection of an incoherent point provides a very different situational interpretation. The projection perspective suggests that assessors are in some sense illogical; they're unable to come up with the 'right' assessments (i.e. ones that are probabilities).

The alternative geometry of constraint relaxation has a different implication. Rather than the experts being incompetent, it is the space itself which is indeterminant. In this view, the characteristic matrix representation is insufficient to describe the relation between assessed events and atomic events. If we view atomic events as 'observables' and assessment events as 'assertions,' the relaxation would correspond to allowing that while we all may see the same thing, our interpretation of our observations my differ.

Philosophically, the situation is akin to that described by John Wisdom in his metaphor of the garden:

> Two people return to their long neglected garden and find among the weeds a few of the old plants surprisingly vigorous. One says to the other, 'It must be that a gardener has been coming and doing something about these plants'. Upon inquiry they find that no neighbour has ever seen anyone at work in their garden. The first man says to the other, 'He must have worked while people slept'. The other says, 'No, someone would have heard him and besides, anybody who cared about the plants would have kept down these weeds'. The first man says, 'Look at the way these are arranged. There is purpose and a feeling for beauty here. I believe that someone comes, someone invisible to mortal eyes. I believe that the more carefully we look the more we shall find confirmation of this'. They examine the garden ever so carefully and sometimes they come on new things suggesting that a gardener comes and sometimes they come on new things suggesting the contrary and even that a malicious person has been at work.

In the metaphor, both observers see exactly the same thing, and there is no logical justification for preferring one explanation over the other. It is, paraphrasing Savage, a sentiment in logic similar to *de gustibus non est disputandem.*

## ■ 6.5 Conclusion

In this brief chapter we have analyzed the issue of coherent approximation from an alternative perspective. Rather than assuming the assessments are incoherent because they are in error, the problem is analyzed from the perspective that incoherence is an issue of the assumed coupling structure between atomic events and assessed random variables. Two ways in which the assumed structure may be deficient were introduced and methods for correcting the structural deficiencies were proposed and analyzed. Finally, a connection was made to the broader question of subjectivity of experience and whether ambiguity of experience is an essential element of life.

In the future, further analysis should be done on the connection to basis pursuit and matching pursuit with regards to the problem of minimal structural revision under conditions of uncertainty. Also, the comparison between the optimal structural revision under uncertainty and compressed sensing should be analyzed. In terms of structural revision under indeterminacy, the connection between various dual formulations and optimal structural deformations should be further analyzed.

# Chapter 7

# Conclusion

## 7.1 Coherent Approximation

Taking action in our complicated, interconnected world requires the best possible situational awareness. Gathering information in the form of expert opinion and assessments, whether those experts are human or machine, can improve a decision maker's understanding of the consequences of his actions. However, when expert assessments fail to be coherent, the decision maker's awareness is impeded. What is needed is a method for approximating the inconsistent views with consistent ones that still reflect, as much as possible, the expert opinion that underlies the assessments.

In this thesis we have proposed and developed a framework for performing coherent approximation, based on principles of information geometry. We have shown that it is robustly applicable to various types of quantitative assessment, and verified the effectiveness of the framework in simulations of dynamically varying environments. We have compared it to previous attempts at coherent approximation and shown specific limitations of previous approaches that are overcome under the proposed algorithm.

### 7.1.1 Contributions

This thesis has developed several methods and applications of coherent approximation or combination of expert assessments.

- In Chapter 3 a coherent approximation principle based on information geometry (the IGCAP) was motivated, formulated, justified, and analyzed.

- In Chapter 4 a mechanism for coherently approximating sequences of assessments generated by mismatched likelihood models was developed and demonstrated in simulation.

- Also in Chapter 4 an application of IGCAP to perform approximate Bayesian filtering based on an incoherent sequence of expert assessments was derived and verified in simulation.

- In Chapter 5 a method for measuring aggregate financial risk coherently was developed, based in part on the IGCAP.

- In Chapter 6 an alternative interpretation of incoherence as a structural limitation was put forth and two suggested methods for relaxing the structure were suggested.

We cover each of these contributions in greater depth below.

#### Information Geometric Coherent Approximation Principle

Previous work [76,77] had formulated the problem of coherently approximating a set of assessments as a Euclidean projection. While intuitive and relatively simple to compute, this approach had significant limitations. It treats probability as a fungible commodity rather than a representation of internalized uncertainty; it fails to extend invariantly to assessments of non-characteristic random variables; and the principle lacks justification in terms of a fundamental model of assessor error.

To overcome some of these limitations, in Chapter 3 we introduced a new coherent approximation principle based in information geometry, which we refer to as the Information Geometric Coherent Approximation Principle (IGCAP). The key principle underlying the IGCAP is that assessments define linear families on the probability simplex, and that the coherent solution $\lambda^*$ is the point that lies 'closest' (on average or, in an alternate formulation, in minimax) to those linear families. We showed that the IGCAP can be viewed as an ML estimate of an observation generating distribution, with incoherence among assessors explained through the independence of the distributed observation process. We used a large deviations principle to demonstrate a conditional limit theorem which allowed us to interpret the projection of $Q^*$ onto the assessment-generated linear families as a good approximation of each assessor's private type.

Several benefits of the IGCAP were demonstrated, including:

1. The solution $P^* = \mathbf{X}\lambda^*$ is (a) unique and (b) equal to $P$ if $P$ is coherent
2. The computation is particularly tractable for characteristic random variables, with a special case that agrees with the operator model suggested in Section 3.3.2
3. The mechanism can extend invariantly to assessments of non-characteristic random variables
4. The formulation can be expanded to allow for assessments given as ranges rather than points, or to account for potential risk aversion (or risk seeking) among assessors
5. Given a sequence of assessments $P_n$ such that $\lim_{n\to\infty} P_n = \bar{P}$ is coherent, the sequence of solutions $P_n^* \to \bar{P}$.

#### Coherent Approximation of Sequential Assessments

In Chapter 4 we developed three distinct sequential assessment regimes and demonstrated mechanisms, based in the IGCAP, by which incoherent sequences of assessments could be coherently approximated.

The Subjective Likelihood (SL) model consisted of a set of assessors, each responsible for generating assessments of a particular characteristic random variable based on a sequence of globally-observable random variables (assumed i.i.d.). Each assessor has a subjective likelihood function that encodes the relation between the observational sequence and the value of their particular random variable. We introduced the concepts of Step-Wise Coherence (SWC) and Weak, Asymptotic Coherence (WAC) as classifications of the set of subjective likelihood models, and showed that WAC is a strictly weaker sense of likelihood coherence. Using large deviations techniques and the asymptotic coherence structure of a set of likelihood models, we formulated a principle of conserving predictive uncertainty which provided a coherent approximation for a sequence of incoherent assessments generated by subjective likelihood functions.

In the Conditional Assessment (CA) model, each assessment may be conditioned on some subset of the atomic events. This conditioning may be based on side information known to the assessor, or it may be easier for the assessor to only consider a subset of the possibilities. It was shown that, just as with unconditional assessments, conditional assessments generate a linear family on the simplex over the atoms of $\Omega$. It was further shown that under IGCAP, given two assessments conditioned on the same event their coherent approximation could be derived in two equivalent ways: 1) as the optimal solution to the IGCAP in the subset of the simplex spanned by the conditioning event or 2) as the conditional revision of the optimal solution to the IGCAP on the simplex over the whole of $\Omega$.

In the Markov Chain (MC) model, the individual random variables of the assessors were assumed to be random sequences that varied over time according to some underlying Markov process. Furthermore, each assessor was assumed to have access to private observations which were integrated into his assessment using a likelihood model that was unknown to the other assessors or to the approximator. A method for approximate Bayesian filtering based on the sequence of distributed, incoherent assessments was derived and demonstrated in simulation to perform comparably to the optimal, centralized Bayesian filter with full information.

#### Coherent Approximation of Risk Assessments

In Chapter 5 we analyzed several aspects of financial risk assessment, particular in the context of the axiomatic class of *coherent* and *convex* risk measures. After defining the concept of minimal convex extension of a risk measure as the one with minimal acceptance set that 1) includes the acceptance set of the original risk measure and 2) adheres to the necessary axioms, we demonstrated that the minimal convex extension of the Value-at-Risk (VaR) risk measure for any parameter $\gamma$ is still a VaR risk measure with parameter $\gamma^*$.

We then defined a robustness principle for risk measures and analyzed the robustness of several suggested risk measures to small variations in the subjective probability distribution. It was shown that several risk measures, including 'nicely behaved' coherent risk measures, may be fragile under minor divergences of opinion.

Next, we developed a mechanism for generating a coherent risk measure of an

aggregate financial position given a distributed set of risk assessments of the individual financial positions. When the risk assessments were generated coherently (or were coherent-equivalent in the sense defined in the chapter) a method was derived to approximate the scenario set of the coherent risk measure that gave rise to the assessments. When the risk assessments were incoherent, the IGCAP was employed to generate a coherent approximation of the aggregate risk.

#### Knightian Uncertainty and Indeterminacy

Chapter 6 challenged the fundamental assumption of the thesis that incoherent assessments were 'wrong.' Instead, it was posited that the assessments themselves were accurate but that the structure of the problem was at fault. Two mechanisms in which the structure could be insufficient were suggested: 1) a failure to foresee possible outcomes leading to an incomplete output matrix and 2) a fundamental inability to deterministically map atomic states into assessors' observables. These two mechanisms were identified with 'uncertainty' and 'indeterminacy' respectively and two methods for revising the structure accordingly were suggested.

In the case of uncertainty, it was suggested that a minimal number of additional columns should be inserted in the output matrix. It was shown that, considering the columns of the initial output matrix as a subset of all potential columns (drawn from a finite field), the problem could equivalently be viewed as determining a sparse augmentation to the current set of bases. The special case, in which the initial set of columns is empty, was analyzed and a greedy algorithm was proposed. It was shown that for $\mu$-almost all assessments $P$ the greedy algorithm performs optimally, but that in certain highly-structured examples it can perform exponentially badly.

In the case of indeterminacy, the concept of a non-deterministic output matrix, in which the matrix elements are intervals of values rather than point values, was proposed. It was suggested that the geometry of the coherent approximation problem could be adapted to determine intervals that would optimally approximate the original structure while sufficiently relaxing the structure to incorporate the assessment. For three different coherent formulations, the CAP, the IGCAP and a variant of the IGCAP with the divergence order reversed, the conjugate function was utilized to determine geodesics along which to optimally relax the structure of the output matrix.

### ■ 7.1.2 Future Work

The IGCAP developed in Chapter 3 has many attractive properties, including guaranteed existence and uniqueness, affine invariance, and many others. The current development does not suppose to be an exhaustive development of the principles, benefits, and limitations of IGCAP. Further attention should be given to developing the principle, particularly as regards its application to actual problems of distributed assessment.

The connection between the subjective likelihood model in Chapter 4 and the IGCAP is mainly conceptual. Further investigation to determine if a more fundamental connection exists between the two could be undertaken. Also, the Markov Chain

(MC) model from Chapter 4 could be generalized to more general Markov structures such as trees and graphs.

In the theoretical development of coherent approximation of distributed risk assessment in Chapter 5, the choice of aggregate coherent risk measure was chosen based on the IGCAP. However, in the final analysis only a relatively weak justification was offered, relating to the increase of assessed marginal risk under alternative measures that decrease the sum of divergences between 'worst case' scenarios. Essentially, the IGCAP is a Pareto efficient balance between marginal risk and maximum divergence between worst-case scenarios. However, it is a single point on a Pareto frontier. A more significant result would justify choosing it over alternative Pareto efficient risk measures.

The development in Chapter 6 of methods for structural relaxation in cases of uncertainty could be significantly developed. The current analysis applies directly only to the case when $S = \emptyset$, with a suggestion that it could be adapted to the more general case. The algorithm itself could be further analyzed, for instance determining performance when assessments are constrained to live on a lattice rather than in $[0,1]^M$, and developed. Determining the relation to previously proposed greedy algorithms for matching pursuit, and comparing to the relaxed case of using $L_1$ norm as a surrogate for $L_0$ which has worked well in problems of basis pursuit and compressive sensing are additional areas of possible further development.

Also, in Chapter 6 the development of structural relaxation in cases of indeterminacy leaves open the question of the precise mapping between optimal distances along geodesics (parameterized by the optimal dual variable) and definition of a non-deterministic output matrix. Determining in what sense each of the suggested relaxations (based on duals of the CAP, the IGCAP, and the order-reversed IGCAP) results in an 'optimal' relaxation of the characteristic matrix would be fruitful. Particularly interesting would be an investigation of whether the parameterized family arising in the definition of the conjugate of the IGCAP objective has further application. It's analogs, the least-squares Euclidean geodesic and the exponential family, both have significant application, suggesting that perhaps it is of more fundamental interest.

Finally, the proof (as they say) is in the pudding. Applying the IGCAP or the related dynamic approximation methods developed in Chapters 3 and 4 to real world data sets is a very important avenue of future development. First steps in this direction may be suggested by the application of IGCAP to financial risk assessment in Chapter 5. More should be done to verify that the coherent approximation framework developed here can improve situational understanding for actual decision makers.

## ■ 7.2 Concluding Remarks

One of the most important endeavors we undertake as humans, both individually and as countries, communities and societies, is to reconcile individually divergent values and beliefs. I hope this thesis has helped in some small way to further that goal of reconciliation. Although the development has been highly mathematical and tied to

a very specific model of valuation, my desire throughout has been to say something about the larger problem of reconciling contradictions generally while maintaining as much as possible the rights, prerogatives, and values of the individual. I'm personally skeptical of man's ability to improve society through simply thinking hard. However, I feel that when animated by a worthy cause, driven by a 'moral outrage' (in the words of West Churchman [166]), we are not only capable of helping make society better, but *obligated* to do so.

> It would be a good thing if the systems planner's germination was moral outrage and not just a mild felt need. In other words, I do not think we should view the major problems of the world today with calm objectivity. We shouldn't first ask ourselves for a precise and operational definition of malnutrition. We should begin with 'kids are starving in great numbers, damn it all!'

While it may not have come through in the words of this thesis, I feel strongly the moral need to address the growing problem of egocentrism and lack of perspective within individuals and societies. The challenge of coherence, in the end, is the challenge of finding a way to live together, despite our sometimes contradictory values and beliefs. I hope we can increasingly meet that challenge head on.

# Appendix A

# Ramsey, de Finetti, and Savage

## A.1 Introduction

In this appendix I will attempt to briefly trace the concept of probabilistic coherence. I will devote a section each to the developments of Ramsey [27], de Finetti [26] and Savage [28]. I will then briefly analyze the similarities and differences between the various developments.

## A.2 Ramsey

Ramsey's life was cut tragically short due to illness in 1930. Before his death he made several contributions to the modern theories of economics and probability. In this section we will analyze the contributions of his paper "Truth and Probability" to the theory developed in this thesis, with a particular focus on his views about how probability is quantified and the importance of coherence, or (in Ramsey's terms) consistency of degree of belief.

Ramsey's development begins with a critique of Keynesian probability theory. Particularly troubling to Ramsey is Keynes' view that in probability, as in logic, no room can be given to personal belief. To Ramsey, the idea that specific probabilities flow necessarily from certain fixed beliefs is contradicted by the fact that there seems to be little consensus about particular probabilities. Ramsey compares it to geometry, saying "it is as if everyone knew the laws of geometry but no one could tell whether any given object were round or square."

Ramsey's own view of probability is highly personalistic and based in the logic of choice under uncertainty. In his words, "[t]he difference [between believing something more or less firmly] seems to me to lie in how far we should act on these beliefs." He terms this view beliefs *qua* bases of action, and relates the quantification of probability to the lowest possible odds an individual is willing to accept on a wager, or, equivalently, a point of indifference between betting for or against a proposition. Ramsey recognizes the reality that humans experience, to varying degrees, attraction or aversion to risk; however, as an idealization he accepts the idea of a risk-neutral agent.

Ramsey then proposes a set of "axioms of consistency" which are sufficient to imply the fundamental laws of probability. In addition to consistency (e.g. transitivity) of action and belief, Ramsey requires the existence of a maximally uncertain

proposition (i.e. one "believed to degree $\frac{1}{2}$"). There are also two axioms relating to continuity of beliefs. However, it is in the axioms of consistency that the concept of coherence is found.

> These are the laws of probability, which we have proved to be necessarily true of any consistent set of degrees of belief. Any definite set of degrees of belief which broke them would be inconsistent in the sense that it violated the laws of preference between options, such as that preferability is a transitive asymmetrical relation, and that if $\alpha$ is preferable to $\beta$, $\beta$ for certain cannot be preferable to $\alpha$ if $p$, $\beta$ if not-$p$. If anyone's mental condition violated these laws, his choice would depend on the precise form in which the options were offered him, which would be absurd. He could have a book made against him by a cunning better and would then stand to lose in any event.

## ■ A.3 de Finetti

The following summary of de Finetti's development is taken primarily from Chapter 3 of [26].

In his *Theory of Probability* [26], De Finetti independently derived a view of probability similar to Ramsey's based on very similar principles. After spending two chapters introducing the principles and developing a view of possible/impossible, in Chapter 3 de Finetti proceeds to the 'logic of uncertainty.' He here begins the technical development of a theory of subjective probability based on a rational bettor's willingness to accept certain odds as 'fair' (meaning the bettor is indifferent to either side of a wager at the given odds). In this sense he follows Ramsey in viewing probability theory as being based on the theory of decision-making, albeit assuming an idealized (in terms of risk aversion) decision maker.

For de Finetti, as with Ramsey, the critical decision-making requirement is one of consistency among decisions, which de Finetti calls *coherence.*

> It turns out, in fact, that there exist simple (and, in the last analysis, obvious) conditions, which we term conditions *of coherence*: any transgression of these results in decisions whose consequences are manifestly undesirable (leading to certain loss). The 'one must'...is not to be taken as an obligation that someone means to impose from the outside, nor as an assertion that our evaluations are always automatically coherent. On the contrary, it is precisely because this is an area where it is particularly easy to slip into incoherence that it is important to learn the *art of prevision.*

The development begins with the positing of uncertain quantities (denoted $X$ and $Y$) and a 'price' function $\mathbf{P}$ which maps uncertain quantities to the real numbers. The interpretation of the price function $\mathbf{P}$ is that $\mathbf{P}(X)$ is the certain amount deemed equivalently valuable to the decision-maker as the uncertain quantity $X$. De Finetti then states the following two consistency axioms:

(a) the price $\mathbf{P}$ is an additive function: $\mathbf{P}(X+Y) = \mathbf{P}(X) + \mathbf{P}(Y)$.

(b) the price $\mathbf{P}$ must satisfy the inequality $\inf X \leq \mathbf{P}(X) \leq \sup X$.

In de Finetti's words, "the two extremely simple conditions...are *not only necessary but also sufficient* for coherence - i.e. for avoiding undesirable decisions. This is all that is needed for the foundation of the whole theory of probability."

De Finetti concludes his basic development by noting two criteria, each consisting of a decision framework and a consistency principle, which are equivalent in terms of eliciting probabilities (or, more generally, previsions). The first decision framework is the betting proposition familiar from Ramsey's work, with the associated consistency requirement being the non-existence of a book with guaranteed positive payoff. The second decision framework posits a penalty function equal to the squared difference of a prevision and the random quantity's realized value, with an associated coherence principle that the chosen prevision is not dominated (in the sense of there existing one with a guaranteed lower value) by another prevision.

## ■ A.4 Savage

In *The Foundations of Statistics* [28], Savage develops a theory of probability very similar to those developed by de Finetti and Ramsey. After concisely introducing the concepts of actions, states, and consequences, Savage proposes a set of seven axioms that are sufficient to define probability.

One of the framing contributions Savage makes is in considering rationality not as an *open-loop* property, but explicitly calling out it's closed-loop nature. "When certain maxims are presented for your consideration, you must ask yourself whether you try to behave in accordance with them, or, to put it differently, how you would react if you noticed yourself violating them." This seems a valuable conceptual deviation from the deductive approach taken by both Ramsey and de Finetti.

The first axiom Savage proposes, which he views as fundamental, is that actions can be simply ordered with respect to preferences. Specifically, for all actions $x, y, z$, Savage assumes 1) either $x$ is not preferred to $y$ or $y$ is not preferred to $x$ (or both) and 2) if $x$ is preferred to $y$ and $y$ to $z$ then $x$ is preferred to $z$. Thus, de Finetti style coherence is again presented here as fundamental to the process of developing a probability theory based in decision theory.

The second Savage axiom is related to what he calls 'the sure-thing principle', by which he means that that if an action is preferred in every possible state, then it must be globally preferred. The third axiom deals with acts with constant (across all possible states) consequences, and states that such an action is not preferred to another such action if and only if the consequence of the first is less than or equal to that of the second. This is the first move to connecting preferred actions to preferred payoffs (consequences), which Savage continues with axioms 4-6, specifically: scale invariance in preferences over consequences, non-universal indifference over consequences, and fineness and tightness in the preferences over consequences (essentially meaning there exists some partition of the space such that any strict preference over consequences can't be rendered weak or overturned by varying the consequence on each element of the partition individually).

The final step toward quantification of probability is made by applying utility to consequences, and considering (a la von Neumann-Morgenstern) 'gambles' between consequences. The final axiom (number seven) is a variant of the sure-thing principle presented as axiom two, but extending to utility of consequences over states

## A.5 Discussion

There are several remarkable similarities and a few differences between the developments of Ramsey, de Finetti, and Savage.

All three treat as fundamental the concept of coherence. Ramsey and de Finetti found this in the argument of Dutch Books, although de Finetti adds the secondary criteria based on dominated scoring functions. In Savage's development, the axiom of simple ordering over acts doesn't attempt to justify itself philosophically, but is seen as a fundamental expression of human decision making. In this sense, the coherence in Savage's development is more primitive than the coherence from Ramsey and de Finetti.

Another commonality among the three developments is the view of decision theory as being foundational to the theory of probability. This personalistic approach is in stark contrast to the objectivist approach that was ascendant at the time. However, all three recognize the challenge in treating humans as 'rational' (in the sense of expectation maximizing) decision makers when there is evidence (even before Kahneman and Tversky's landmark studies) that such an idealization was poorly justified in practice. Ramsey refers to this principle is stating that pursuing his line of development further would be akin to seeking the seventh decimal of a number that can't be precisely determined beyond the second decimal. The inherent reality of non-rational human decision making obviously weighs heavily on all their minds.

For Savage, the role of gambling plays a fairly minor part of the overall development, relative to de Finetti. Indeed, much of Savage's theory of probability is developed prior to introducing gambling and utility as a method of mixing across consequences, whereas the idea of gambling is the origin of de Finetti's development. For Ramsey, the role of gambling (in terms of viewing probability as fair odds for a gamble) is similarly fundamental, although it is largely implied in his development. Savage chooses to focus first on general actions under uncertainty, and arguably views the final introduction of the specific action of setting odds as a relatively minor aspect of the overall process.

All three developments, to varying degrees, point to the idea of probability as being an outgrowth of logic. This is partially seen through the reliance on coherence and consistency axioms or principles, as already outlined. In the case of both Ramsey and de Finetti it goes significantly beyond this as they explicitly call out the relationship between logical consistency and probability. This is more explicit in Ramsey's treatment, which adopts much of its notation from the field of analytical logic and dwells at some length on the interrelation between the two fields.

In the end, each of these three seminal developments adds something to our picture of subjective probability. To grossly oversimplify the unique contributions of each

development, from Ramsey we get the strong connection to the field of logic, from de Finetti we get an extremely concise development in terms of wagering, and from Savage we get a broader connection to decision theory as well as a strikingly elegant mathematical construction of states, acts, consequences and utilities.

# Appendix B

# Simulation Code

## B.1 Subjective Likelihood Simulation

```
clear all;

T = 3*ones(3)+diag(ones(1,3));
T = T/sum(sum(T));

nB = sum(T(:,[2 3]),2)*3/2;
nB = nB([1 2 3;2 1 3;3 2 1]);
x = 1;
for idx=1:1000
    rand('seed',idx);
    B = T*3+(diag(1.5*ones(1,3))-.5).*(ones(3,1)*rand(1,3)/3);
    z = rand(1,2000);
    z = sum(cumsum(T(:,x)*3)*ones(1,length(z))<ones(3,1)*z)+1;

    P = ones(3,1)/3;
    for a=1:length(z)
        P(:,a+1) = P(:,a).*B(z(a),:)'./...
          (P(:,a).*B(z(a),:)'+(1-P(:,a)).*nB(z(a),:)');
    end

    Phat1 = P./(ones(3,1)*sum(P));

    Phat2 = P;
    [mns,idcs] = min(Phat2);
    Phat2 = Phat2-ones(3,1)*min([mns;(sum(Phat2)-1)/3]);
    tfs = find(min(Phat2)==0);
    Phat2(:,tfs) = max(Phat2(:,tfs)-...
      ones(3,1)*(sum(Phat2(:,tfs))-1)/2, zeros(size(Phat2(:,tfs))));

    l1 = log(diag(B)./diag(nB));
    l2 = log(diag([0 0 1;1 0 0;0 1 0]*B)/.35);
    C = l2./(l2-l1);
```

```
    rho_hat = cumsum([z==1;z==2;z==3],2)./(ones(3,1)*[1:length(z)]);

    wt = rho_hat(2,:)./(rho_hat(2,:)+rho_hat(3,:));
    rho_star = [C(1)*ones(1,length(z));(1-C(1))*wt;(1-C(1))*(1-wt)];
    wt = rho_hat(1,:)./(rho_hat(1,:)+rho_hat(3,:));
    rho_star(:,:,2) = [(1-C(2))*wt;...
      C(2)*ones(1,length(z));(1-C(2))*(1-wt)];
    wt = rho_hat(1,:)./(rho_hat(1,:)+rho_hat(2,:));
    rho_star(:,:,3) = [(1-C(3))*wt;...
      (1-C(3))*(1-wt);C(3)*ones(1,length(z))];
    threshold_breakers = rho_hat>(C*ones(1,length(z)));
    for a=1:length(z)
        brkrs = find(threshold_breakers(:,a));
        if length(brkrs)==1
            rho_star(:,a,brkrs) = rho_hat(:,a);
        elseif length(brkrs)==2
            tmp = rho_star(:,a,brkrs(1));
            rho_star(:,a,brkrs(1)) = rho_star(:,a,brkrs(2));
            rho_star(:,a,brkrs(2)) = tmp;
        end
    end

    %calculate weights based on e^-nD(rho_hat||rho_star)
    wts = [];
    for a=1:3
        wts(a,:) = exp(-[1:length(z)].*sum(rho_hat.*...
          log(rho_hat./rho_star(:,:,a))));
    end
    %keyboard;
    wts = wts./(ones(3,1)*sum(wts));
    Phat3(:,1) = ones(3,1)/3;
    Phat3(:,2:length(z)+1) = diag(ones(1,3))*wts;

    D = (1:3==x)'*ones(1,length(z)+1)-P;
    Dhat1 = (1:3==x)'*ones(1,length(z)+1)-Phat1;
    Dhat2 = (1:3==x)'*ones(1,length(z)+1)-Phat2;
    Dhat3 = (1:3==x)'*ones(1,length(z)+1)-Phat3;

    M(:,:,idx) = [sum(D.*D);sum(Dhat1.*Dhat1);...
      sum(Dhat2.*Dhat2);sum(Dhat3.*Dhat3)];
end
```

## ■ B.2 Bayesian Filtering Simulation

```
%NOTE: Uses the cvx convex optimization solver
for idx = 1:100
    %clear all; close all;
    clear s
    %rand('seed',1);
    cvx_quiet(1);
    N = 3;
    K = 200;

    Theta = 1:N;
    %pt = .5+.5*rand;
    pt = .5+.2*rand;
    pnt = (1-pt)/(N-1);
    T = (pt-pnt)*diag(ones(1,N))+pnt;

    cT = cumsum(T);
    %py = .4+.6*rand;
    py = .4+.4*rand;
    pny = (1-py)/(N-1);
    %Py = diag(ones(1,3));
    Py = (py-pny)*diag(ones(1,N))+pny;
    cPy = cumsum(Py);

    X = diag(ones(1,N));
    %I'm going to interleave the following Probability matrices,
    %so each odd column is a 'prediction'
    %and each even column is an 'update'
    Pl = (1/N)*ones(N,1);
    %Pl is the local estimate of the marginal probability;
    %unlike the other probability matrices, it's not a
    $probability mass function (i.e. each column is an
    $assessment, and the assessments are generally incoherent)

    Pg = (1/N)*ones(N,1);
    %Pg is the estimate using IGCAP to approximate likelihoods
    Po = Pg;
    %Po is the optimal centralized estimate, using all information

    s(1) = ceil(rand*N);
    pi_star_old = (T*1/N*ones(N,1))*ones(1,N);

    for a=1:K
        if ~mod(a,round(K/10))
```

```
    fprintf('.');
end
for b=1:N
    y(b) = find(rand<cPy(:,s(a)),1,'first');
    %local update step
    tPl = Pl(b,2*a-1);
    if y(b)==b
        Pl(b,2*a) = py*tPl/(py*tPl+pny*(1-tPl));
    else
        Pl(b,2*a) = pny*tPl/(pny*tPl+py*(1-tPl));
    end
end
%global update step
P = Pl(:,2*a);
cvx_begin
variable Q(N)
maximize sum(P.*log(Q)+(1-P).*log(1-Q))
subject to
ones(1,3)*Q==1
Q>=0
cvx_end
for b=1:N
    Z(b) = (1-Q(b))/(1-Pl(b,2*a));
    pi_star(:,b) = Q/Z(b);
    %pi_star(b,b) = Q(b)/Z(b)*exp(th(b));
    pi_star(b,b) = Pl(b,2*a);
end
Pg(:,2*a) = prod([Pg(:,2*a-1) pi_star./pi_star_old],2);
Pg(:,2*a) = Pg(:,2*a)/sum(Pg(:,2*a));
pi_star_old = T*pi_star;

%optimal update step
Po(:,2*a) = prod(Py(y,:))'.*Po(:,2*a-1)/...
  (prod(Py(y,:))*Po(:,2*a-1));

%update state
s(a+1) = find(rand<cumsum(T(:,s(a))),1,'first');

%local prediction step
for b=1:N
    Pl(b,2*a+1) = pt*Pl(b,2*a)+pnt*(1-Pl(b,2*a));
end

%global prediction step
Pg(:,2*a+1) = T*Pg(:,2*a);
```

```
        %centralized prediction step
        %    Pc(:,2*a+1) = T*Pc(:,2*a);

        %optimal prediction step
        Po(:,2*a+1) = T*Po(:,2*a);
    end
    fprintf('\n');
    Po_perf(idx,:) = 1-Po([3:6:3*(size(Po,2)-2)]+s(1:end-1));
    Pg_perf(idx,:) = 1-Pg([3:6:3*(size(Po,2)-2)]+s(1:end-1));
    Pl_perf(idx,:) = 1-Pl([3:6:3*(size(Po,2)-2)]+s(1:end-1));
    pt_hist(idx) = pt;
    py_hist(idx) = py;
end
```

# Bibliography


[1] P. K. Varshney, *Distributed Detection and Data Fusion.* Springer, 1996. 17

[2] R. R. Tenney and N. R. Sandell, "Detection with distributed sensors," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. AES-17, no. 4, pp. 501 –510, Jul. 1981. doi: 10.1109/TAES.1981.309178 17

[3] ——, "Structures for distributed decision making," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. 11, no. 8, pp. 517 –527, Aug. 1981. doi: 10.1109/TSMC.1981.4308739 17

[4] J. Tsitsiklis, "Problems in decentralized decision making and computation," PhD Thesis, Massachusetts Institute of Technology, 1984. 17

[5] ——, "Decentralized detection by a large number of sensors," in *Mathematics of Control, Signals and Systems*, vol. 1, 1988, pp. 167–182. 17

[6] ——, "Decentralized detection," in *Advances in Statistical Signal Processing*, 1993, vol. 2, pp. 297 –344. 17

[7] Z. Chair and P. Varshney, "Optimal data fusion in multiple sensor detection systems," *Aerospace and Electronic Systems, IEEE Transactions on*, vol. AES-22, no. 1, pp. 98 –101, Jan. 1986. doi: 10.1109/TAES.1986.310699 17

[8] V. Blondel, J. Hendrickx, A. Olshevsky, and J. Tsitsiklis, "Convergence in multiagent coordination, consensus, and flocking," in *Decision and Control, IEEE Conference on*, 2006. 17

[9] W. Ren, R. Beard, and E. Atkins, "A survey of consensus problems in multi-agent coordination," in *American Control Conference*, 2005. 17

[10] R. Olfati-Saber and R. Murray, "Consensus problems in networks of agents with switching topology and time-delays," *Automatic Control, IEEE Transactions on*, vol. 49, no. 9, Sep. 2004. 17

[11] H. G. Tanner, A. Jadbabaie, and G. J. Pappas, "Flocking in fixed and switching networks," *Automatic Control, IEEE Transactions on*, vol. 52, no. 5, pp. 863 –868, May 2007. 17

[12] R. Olfati-Saber, "Flocking for multi-agent dynamic systems: Algorithms and theory," *Automatic Control, IEEE Transactions on*, vol. 51, no. 3, pp. 401–420, Mar. 2006. 17

[13] D. Mosk-Aoyama and D. Shah, "Fast distributed algorithms for computing separable functions," *Information Theory, IEEE Transactions on*, vol. 54, no. 7, pp. 2997 –3007, Jul. 2008. 17

[14] G. N. Frederickson and N. A. Lynch, "Electing a leader in a synchronous ring," *Journal of the ACM*, vol. 34, pp. 98–115, January 1987. 17

[15] N. Malpani, J. L. Welch, and N. Vaidya, "Leader election algorithms for mobile ad hoc networks," in *Proceedings of the 4th international workshop on Discrete algorithms and methods for mobile computing and communications*, 2000, pp. 96–103. 17

[16] H. Kopetz and W. Ochsenreiter, "Clock synchronization in distributed real-time systems," *Computers, IEEE Transactions on*, vol. C-36, no. 8, pp. 933 –940, Aug. 1987. 17

[17] Q. Li and D. Rus, "Global clock synchronization in sensor networks," *IEEE Transactions on Computers*, vol. 55, pp. 214–226, 2006. 17

[18] B. D. Anderson and J. B. Moore, *Optimal Filtering.* Englewood Cliffs, NJ: Prentice-Hall, 1979. 17

[19] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series.* New York: Wiley, 1949. 17

[20] R. E. Kalman, "A new approach to linear filtering and prediction problems," *Transactions of the ASME–Journal of Basic Engineering (Series D)*, vol. 82, pp. 35–45, 1960. 17

[21] R. Kalman and R. S. Bucy, "New results in linear filtering and prediction theory," *Transactions of the ASME–Journal of Basic Engineering (Series D)*, vol. 83, pp. 95–108, Mar. 1961. 17

[22] J. Speyer, "Computation and transmission requirements for a decentralized linear-quadratic-gaussian control problem," *Automaic Control, IEEE Transactions on*, vol. 24, no. 2, pp. 266–269, Feb. 1979. 17

[23] B. Rao, H. Durrant-Whyte, and J. Sheen, "A fully decentralized multi-sensor system for tracking and surveillance," *Int'l Journal of Robotics Research*, vol. 12, no. 1, pp. 20–44, Feb. 1993. 17

[24] R. Olfati-Saber, "Distributed Kalman filtering with embedded consensus filters," in *IEEE Conference on Decision and Control*, 2005. 17

[25] ——, "Distributed Kalman filtering for sensor networks," in *IEEE Conference on Decision and Control*, 2007. 17

[26] B. de Finetti, *Theory of Probability.* Wiley New York, 1974, vol. 1-2. 19, 27, 31, 32, 34, 36, 63, 125, 145, 146

[27] F. P. Ramsey, "Truth and probability," in *The Foundations of Mathematics and other Logical Essays*, R. Braithwaite, Ed. New York: Harcourt, Brace and Co., 1931, ch. VII, pp. 156–198. 19, 27, 33, 145

[28] L. J. Savage, *The Foundations of Statistics.* New York: Wiley, 1954. 19, 25, 27, 31, 145, 147

[29] B. Dowden and N. Swartz. (2004, Sep.) The internet encyclopedia of philosophy: Truth. [Online]. Available: http://www.iep.utm.edu/truth/ 20

[30] B. Russell, "On the nature of truth," *Proceedings of the Aristotelian Society*, vol. 7, pp. 228–249, 1907. 20

[31] L. Wittgenstein, *Tractatus Logico-Philosophicus.* London, UK: Kegan Paul, Trench, Trubner & Co., LTD., 1922, transl. by C.K. Ogden, with introduction by B. Russell. 20

[32] F. Waismann, "Verifiability," in *Logic and Language*, A. G. N. Flew, Ed. London: Blackwell, 1951, pp. 117–144. 20

[33] H. Putnam, *Representation and Reality.* Cambridge, MA: MIT Press, 1988. 20

[34] K. Popper, *Logik der Forschung.* Springer-Verlag, 1935. 20

[35] ——, *The Logic of Scientific Discovery.* Hutchinson & Co., 1959. 20

[36] T. Kuhn, *The Structure of Scientific Revolutions.* University of Chicago Press, 1962. 20

[37] H. Hart, *The Concept of Law.* Oxford,UK: Oxford University Press, 1961. 21

[38] B. Barry, *Sociologists, Economists and Democracy.* The University of Chicago Press, 1978. 21

[39] N. Chomsky, *Aspects of the Theory of Syntax.* Cambridge, MA: The MIT Press, 1965. 21

[40] C. J. Fillmore, "Frame semantic and the nature of language," *Annals of the New York Academy of Sciences*, vol. 280, p. 2032, 1976. 21

[41] G. Lakoff and M. Johnson, "The metaphorical structure of the human conceptual system," *Cognitive Science: A Multidisciplinary Journal*, vol. 4, no. 2, p. 195208, Apr. 1980. 21

[42] G. Shafer and V. Vovk, *Probability and Finance: It's Only a Game!*, ser. Wiley series in probability and statistics. John Wiley & Sons, Inc., 2001. 22, 33, 101

[43] E. Fama, "The behavior of stock market prices," *Journal of Business*, vol. 38, p. 34105, 1965. 22

[44] S. A. Ross, "The arbitrage theory of capital asset pricing," *Journal of Economic Theory*, vol. 13, pp. 341–360, 1976. 22, 34

[45] J. Nash, "Non-cooperative games," *The Annals of Mathematics*, vol. 54, no. 2, pp. 286–295, 1951. 22

[46] R. Aumann, "Subjectivity and correlation in randomized strategies," *Journal of Mathematical Economics*, vol. 1, pp. 67–95, 1974. 22

[47] J. Harsanyi, *Rational Behavior and Bargaining Equilibrium in Games and Social Situations.* Cambridge, UK: Cambridge University Press, 1977. 22

[48] D. Kahneman and A. Tversky, "Prospect theory: an analysis of decision under risk," *Econometrica*, vol. 47, no. 2, pp. 263–292, Mar. 1979. 22

[49] D. Kahneman and A. Tversky, Eds., *Judgment under uncertainty: Heuristics and biases.* Cambridge, UK: Cambridge University Press, 1982. 22

[50] S. Mullainathan and R. H. Thaler, "Behavioral economics," in *International Encyclopedia of the Social and Behavioral Sciences*, N. J. Smelser and P. B. Baltes, Eds. Oxford, UK: Oxford University Press, 2001, vol. 20, p. 10941100. 22

[51] D. M. Kreps, *Notes on the Theory of Choice (Underground Classics in Economics).* Boulder, CO: Westview Press, 1988. 25

[52] K. Binmore, *Rational Decisions.* Princeton, NJ: Princeton University Press, 2008. 25

[53] F. N. David, *Probability Theory for Statistical Methods.* Cambridge: University Press, 1951. 25

[54] P. L. Bernstein, *Against the Gods: The remarkable story of risk.* John Wiley & Sons, Inc., 1996. 26, 101

[55] R. von Mises, "On the foundations of probability and statistics," *The Annals of Mathematical Statistics*, vol. 12, no. 2, pp. 191–205, Jun. 1941. 26

[56] ——, *Probability, Statistics, and Truth*, 2nd ed. George Allen & Unwin Ltd., 1957. 26

[57] R. Fisher, "On the mathematical foundations of theoretical statistics," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 222, pp. 309–368, 1922. 26

[58] J. Neyman and E. S. Pearson, "On the use and interpretation of certain test criteria for purposes of statistical inference: Part i," *Biometrika*, vol. 20A, no. 1/2, pp. 175–240, Jul. 1928. 26

[59] A. M. Keynes, *A Treatise on Probability.* London, UK: Macmillan and Co., Ltd, 1921. 26, 123

[60] R. Carnap, *Logical Foundations of Probability.* University of Chicago Press, 1950. 26

[61] B. Ventelou, *Millenial Keynes: An Introduction to the Origin, Development, and Later Currents of Keynesian Thought.* New York, NY: M.E. Shaw, 2005, translated and edited, with an introduction by Gregory P. Nowell. 27

[62] L. Savage, "The theory of statistical decision," *Journal of the American Statistical Association*, vol. 46, pp. 55–67, 1951. 27

[63] H. Jeffreys, *Theory of Probability*, 3rd ed. Oxford University Press, 1961. 27, 78

[64] F. Anscombe and R. J. Aumann, "A definition of subjective probability," *The Annals of Mathematical Statistics*, vol. 34, no. 1, pp. 199–205, Mar. 1963. 27, 123

[65] C. Smith, "Consistency in statistical inference and decision," *Statisical Society. Series B(Methodological)*, vol. 23, no. 1, pp. 1–37, 1961. 27

[66] D. Schmeidler, "Subjective probability and expected utility without additivity," *Econometrica*, vol. 57, no. 3, pp. 571–587, May 1989. 27

[67] J. Y. Halpern, *Reasoning about Uncertainty.* Cambridge, MA: The MIT Press, 2003. 28

[68] L. Zadeh, "Toward a generalized theory of uncertainty (gtu)an outline," *Information Sciences*, vol. 172, p. 140, 2005. 28

[69] V. Borkar, V. Konda, and S. Mitter, "On de Finetti coherence and Kolmogorov probability," *Statistics and Probability Letters*, vol. 66, no. 4, pp. 417–421, Mar. 2004. 32, 33

[70] I. Karatzas, *Methods of Mathematical Finance.* New York, NY, USA: Springer-Verlag, 1998. 33, 34

[71] B. Skyrms, "Diachronic coherence and radical probabilism," in *Degrees of Belief*, ser. Synthese Library, F. Huber and C. Schmidt-Petri, Eds. Springer Netherlands, 2009, vol. 342, pp. 253–261. 33, 78

[72] H. Shin, "Review of *The Dynamics of Rational Deliberation*," *Economics and Philosophy*, vol. 8, pp. 109–138, 1992. 33

[73] S. A. Clark, "The valuation problem in arbitrage price theory," *Journal of Mathematical Economics*, vol. 22, no. 5, pp. 463 – 478, 1993. 34

[74] D. B. Hausch and W. T. Ziemba, "Arbitrage strategies for cross-track betting on major horse races," *The Journal of Business*, vol. 63, no. 1, pp. 61–78, 1990. 41

[75] ——, "Locks at the racetrack," *Interfaces*, vol. 20, no. 3, pp. 41–48, 1990. 41

[76] D. Osherson and M. Vardi, "Aggregating disparate estimates of chance," *Games and Economic Behavior*, vol. 56, no. 1, pp. 148–173, Jul. 2006. 41, 42, 43, 52, 60, 140

[77] J. Predd, D. Osherson, S. Kulkarni, and H. Poor, "Aggregating forecasts of chance from incoherent and abstaining experts," *Decision Analysis*, vol. 5, pp. 177–189, 2008. 41, 43, 52, 60, 140

[78] J. Predd, R. Seiringer, E. Lieb, D. Osherson, S. Kulkarni, and H. Poor, "Probabilistic coherence and proper scoring rules," *IEEE Transactions on Information Theory*, vol. 55, no. 10, pp. 4786–4792, Oct. 2009. 41, 52, 59, 73

[79] E. Eisenberg and D. Gale, "Consensus of subjective probabilities: The pari-mutuel system," *The Annals of Mathematical Statistics*, vol. 30, no. 1, pp. 165–168, Mar. 1952. 42

[80] L. Brown and Y. Lin, "Racetrack betting and consensus of subjective probabilities," *Statistics & Probability Letters*, vol. 62, pp. 175–187, 2003. 42

[81] C. R. Plott, J. Wit, and W. Yang, "Parimutuel betting markets as information aggregation devices: experimental results," *Economic Theory*, vol. 22, no. 2, pp. 311–351, Sep. 2003. 42

[82] B. Fischhoff, "Debiasing," in *Judgment under uncertainty: Heuristics and biases*, 1st ed., D. Kahneman, P. Slovic, and A. Tversky, Eds. Cambridge, UK: Cambridge University Press, 1982, pp. 422–444. 42

[83] M. Berhold, "Procedures to increase the validity of subjective probability estimates," *Decision Science*, vol. 6, no. 4, pp. 721–730, 1975. 42

[84] R. Nisbett, D. Krantz, C. Jepson, and G. Fong, "Improving inductive inference," in *Judgment under uncertainty: Heuristics and biases*, 1st ed., D. Kahneman, P. Slovic, and A. Tversky, Eds. Cambridge, UK: Cambridge University Press, 1982, pp. 445–459. 42

[85] M. G. Morgan and M. Henrion, *Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis.* Cambridge University Press, 1990, ch. 7.7. 42

[86] R. Clemen and R. Winkler, "Combining probability distributions from experts in risk analysis," *Risk Analysis*, vol. 19, pp. 187–203, 1999. 42

[87] R. Jacobs, "Methods for combining experts' probability assessments," *Neural Computation*, vol. 7, pp. 867–888, 1995. 42, 43

[88] C. Genest and J. Zidek, "Combining probability distributions: A critique and annotated bibliography," *Statistical Science*, vol. 1, pp. 114–148, 1986. 42

[89] J. Kittler, "Combining classifiers: A theoretical framework," *Pattern Analysis and Applications*, vol. 1, no. 1, pp. 18–27, Mar. 1998. 42

[90] E. Bauer and R. Kohavi, "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants," *Machine Learning*, vol. 36, no. 1-2, pp. 105–139, 1999. 42

[91] R. Clemen, "Combining forecasts: A review and annotated bibliography," *International Journal of Forecasting*, vol. 5, pp. 559–583, 1989. 42

[92] T. Dietterich, "Ensemble methods in machine learning," in *1st Int'l Conference on Multiple Classifier Systems*. Spinger Berlin/Heidelberg, Jun. 2000, pp. 1–15. 42

[93] R. M. Dawes, "The robust beauty of improper linear models in decision making," in *Judgment under uncertainty: Heuristics and biases*, 1st ed., D. Kahneman, P. Slovic, and A. Tversky, Eds. Cambridge, UK: Cambridge University Press, 1982, pp. 391–407. 42

[94] D. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, pp. 241–259, 1992. 42

[95] R. Winkler, "The consensus of subjective probability distributions," *Management Science*, vol. 15, no. 2, pp. B61–B75, Oct. 1968. 43

[96] D. Lindley, "Reconciliation of discrete probability distributions," *Operations Research*, vol. 31, no. 5, pp. 866–880, Sep. 1983. 43

[97] R. Batsell, L. Brenner, D. Osherson, S. Tsavachidis, and M. Vardi, "Eliminating incoherence from subjective estimates of chance," in *Proceedings of the 8th International Conference on the Principles of Knowledge Representation and Reasoning (KR 2002)*, Toulouse, France, 2002, pp. 353–364. 43

[98] P. Jones, S. Mitter, and V. Saligrama, "Revision of marginal probability assessments," in *13th International Conference on Information Fusion*, Edinburgh, UK, 2010. 43, 59

[99] A. Abbas, "A Kullback-Leibler view of linear and log-linear pools," *Decision Analysis*, vol. 6, no. 1, pp. 25–37, Mar. 2009. 51

[100] I. Csiszár, "$I$-divergence geometry of probability distributions and minimization problems," *The Annals of Probability*, vol. 3, no. 1, pp. 146–158, 1975. 53, 59, 109

[101] I. Csiszár and F. Matus, "Information projections revisited," *Information Theory, IEEE Transactions on*, vol. 49, no. 6, pp. 1474 – 1490, Jun. 2003. 53, 109

[102] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, 2nd ed. Springer-Verlag, 1998. 55

[103] S. L. Zabell, "Rates of convergence for conditional expectations," *The Annals of Probability*, vol. 8, no. 5, pp. 928–941, Oct. 1980. 57

[104] I. Csiszár, "Sanov property, generalized i-projection and a conditional limit theorem," *The Annals of Probability*, vol. 12, no. 3, pp. 768–793, Aug. 1984. 57

[105] I. Csiszár, T. Cover, and B.-S. Choi, "Conditional limit theorems under markov conditioning," *IEEE Transactions on Information Theory*, vol. IT-33, no. 6, pp. 788–801, Nov. 1987. 57

[106] E. Jaynes, *Probability Theory: The Logic of Science.* Cambridge University Press, 2003. 61

[107] D. Becherer and M. H. Davis, *ArrowDebreu Prices.* John Wiley & Sons, Ltd, 2010. 66

[108] K. Back, T. R. Bielecki, C. Hipp, S. Peng, W. Schachermayer, and W. Schachermayer, "Utility maximisation in incomplete markets," in *Stochastic Methods in Finance*, ser. Lecture Notes in Mathematics. Springer Berlin / Heidelberg, 2004, vol. 1856, pp. 3–14. 66

[109] D. Kelsey and F. Milne, "The arbitrage pricing theorem with non-expected utility preferences," *Journal of Economic Theory*, vol. 65, no. 2, pp. 557 – 574, 1995. 73

[110] J. Wolfers and E. Zitzewitz, "Prediction markets," *The Journal of Economic Perspectives*, vol. 18, no. 2, pp. pp. 107–126, 2004. 75

[111] J. E. Berg and T. A. Rietz, "Prediction markets as decision support systems," *Information Systems Frontiers*, vol. 5, pp. 79–93, 2003. 75

[112] J. Wolfers and E. Zitzewitz, "Interpreting prediction market prices as probabilities," NBER Working Paper No. 12200, Tech. Rep., 2007. 75

[113] C. F. Manski, "Interpreting the predictions of prediction markets," *Economics Letters*, vol. 91, no. 3, pp. 425 – 429, 2006. 75

[114] R. Forsythe, F. Nelson, G. R. Neumann, and J. Wright, "Anatomy of an experimental political stock market," *The American Economic Review*, vol. 82, no. 5, pp. 1142–1161, Dec. 1992. 75

[115] P. Diaconis and S. Zabell, "Updating subjective probability," *Journal of the American Statistical Association*, vol. 77, no. 380, pp. 822–830, Dec. 1982. 76, 78

[116] B. Skyrms, "Dynamic coherence and probability kinematics," *Philosophy of Science*, vol. 54, no. 1, pp. 1–20, 1987. 76, 78

[117] D. Freedman and R. Purves, "Bayes' method for bookies," *The Annals of Mathematical Statistics*, vol. 40, no. 4, pp. 1177–1186, Aug. 1969. 78

[118] D. Heath and W. Sudderth, "On finitely additive priors, coherence, and extended admissibility," *The Annals of Statistics*, vol. 6, no. 2, pp. 333–345, Mar. 1978. 78

[119] D. Lane and W. Sudderth, "Coherent and continuous inference," *The Annals of Statistics*, vol. 11, no. 1, pp. 114–120, Mar. 1983. 78

[120] E. Regazzini, "De Finetti's coherence and statistical inference," *The Annals of Statistics*, vol. 15, no. 2, pp. 845–864, Jun. 1987. 78

[121] ——, "Coherent statistical inference and Bayes theorem," *The Annals of Statistics*, vol. 19, no. 1, pp. 366–381, Mar. 1991. 78

[122] B. Skyrms, "Coherence, probability and induction," *Philosophical Issues*, vol. 2, pp. 215–226, 1992. 78

[123] C. E. Alchourrn, P. Gärdenfors, and D. Makinson, "On the logic of theory change: Partial meet contraction and revision functions," *The Journal of Symbolic Logic*, vol. 50, no. 2, pp. 510–530, 1985. 78

[124] K. Scarfone and P. Mell, "Guide to intrusion detection and prevention systems (IDPS)," National Institute of Standards and Technology, Technology Administration, US Dept. of Commerce, Tech. Rep. 800-94, Feb. 2007. 79

[125] T. M. Cover and J. A. Thomas, *Elements of information theory.* New York, NY, USA: Wiley-Interscience, 1991. 81, 84, 97, 109

[126] M. Feder and N. Merhav, "Universal composite hypothesis testing: A competitive minimax approach," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1504–1517, Jun. 2002. 81

[127] H. M. Markowitz, *Portfolio Selection.* New York: Wiley, 1959. 101

[128] M. Rubinstein, "Bruno de Finetti and mean-variance portfolio selection," *Journal of Investment Management*, vol. 4, no. 3, 2006. 101

[129] B. de Finetti, "Il problema dei "pieni"," *Giorn. Ist. Ital. Attuari*, vol. 11, pp. 1–88, 1940. 101

[130] H. Markowitz, "de Finetti scoops Markowitz," *Journal of Investment Management*, vol. 4, no. 3, 2006. 101

[131] C. French, "The Treynor capital asset pricing model," *Journal of Investment Management*, vol. 1, no. 2, pp. 60–72, 2003. 102

[132] W. Sharpe, "Capital asset prices: A theory of market equilibrium under conditions of risk," *The Journal of Finance*, vol. 19, no. 3, pp. 425–442, Sep. 1964. 102

[133] J. Lintner, "The valuation of risk assets and the selection of risky investments in stock portfolios and capital budgets," *Review of Economics and Statistics*, vol. 47, no. 1, pp. 13–37, 1965. 102

[134] P. Jorion, *Value at Risk.* New York: McGraw-Hill, 1997. 102

[135] D. Duffie and J. Pan, "An overview of value at risk," *The Journal of Derivatives*, vol. 4, no. 3, pp. 7–49, 1997. 102

[136] P. Artzner, F. Delbean, J. Eber, and D. Heath, "Coherent measures of risk," *Mathematical Finance*, vol. 9, no. 3, pp. 203–228, Jul. 1999. 102, 103, 104, 105, 107

[137] R. Rockafellar and S. Uryasev, "Optimization of conditional value at risk," *Journal of Risk*, vol. 2, pp. 21–41, 2000. 102

[138] ——, "Conditional value-at-risk for general loss distributions," *Journal of Banking & Finance*, vol. 26, pp. 1443–1471, 2002. 102, 110

[139] D. Bertsimas, G. L. Lauprete, and A. Samaraov, "Shortfall as a risk measure: properties, optimization and applications," *Journal of Economic Dynamics & Control*, vol. 28, pp. 1353–1381, 2004. 102

[140] F. Delbean, "Coherent measures of risk on general probability spaces," in *Advances in Finance and Stochastics: Essays in Honour of Dieter Sondermann*, K. Sandmann and P. Schönbucher, Eds. Springer–Verlag, 2002, ch. 1, pp. 1–38. 102, 103

[141] H. Föllmer and A. Schied, "Robust preferences and convex measures of risk," in *Advances in Finance and Stochastics: Essays in Honour of Dieter Sondermann*, K. Sandmann and P. Schönbucher, Eds. Springer–Verlag, 2002, ch. 1, pp. 1–38. 102, 105, 106, 110

[142] ——, "Convex measures of risk and trading constraints," *Finance and Stochastics*, vol. 6, pp. 429–447, 2002. 102, 104

[143] ——, *Stochastic finance: An introduction in discrete time*, 2nd ed., ser. de Gruyter Studies in Mathematics 27. Berlin NewYork: de Gruyter, 2004. 102, 104, 105

[144] F. H. Knight, *Risk, Uncertainty, and Profit.* Boston, MA, USA: Hart, Schaffner & Marx; Houghton Mifflin Company, 1921. 123

[145] L. G. Epstein and T. Wang, "Intertemporal asset pricing under Knightian uncertainty," *Econometrica*, vol. 62, no. 2, pp. 283–322, Mar. 1994. 123, 125

[146] M. Basili, "Knightian uncertainty in financial markets: An assessment," *Economic Notes*, vol. 30, no. 1, p. 126, Feb. 2001. 123, 125

[147] K. G. Nishimura and H. Ozaki, "Search and Knightian uncertainty," *Journal of Economic Theory*, vol. 119, no. 2, pp. 299 – 333, 2004. 123, 125

[148] G. Shafer, *A Mathematical Theory of Evidence.* Princeton, NJ, USA: Princeton University Press, 1976. 123

[149] A. Dempster, "Upper and lower probabilities induced by a multivalued mapping," *The annals of Mathematical Statistics*, vol. 38, no. 2, pp. 325–339, 1967. 123

[150] ——, "A generalization of Bayesian inference," *Journal of the Royal Statistical Society, Series B*, vol. 30, pp. 205–247, 1968. 123

[151] J. Dow and S. R. da Costa Werlang, "Excess volatility of stock prices and Knightian uncertainty," *European Economic Review*, vol. 36, no. 2-3, pp. 631–638, April 1992. 125

[152] D. Ellsberg, "Risk, ambiguity, and the Savage axioms," *The Quarterly Journal of Economics*, vol. 75, no. 4, pp. 643–669, Nov. 1961. 125

[153] W. Quine, *Ontological Relativity and Other Essays.* New York: Columbia University Press, 1969. 125

[154] T. A. O. Endicott, "Linguistic indeterminacy," *Oxford Journal of Legal Studies*, vol. 16, no. 4, pp. 667–697, 1996. 125

[155] J. Dore and R. P. McDermott, "Linguistic indeterminacy and social context in utterance interpretation," *Language*, vol. 58, no. 2, pp. 374–398, 1982. 125

[156] K. J. Kress, "Legal indeterminacy," *77 California Law Review*, vol. 283, pp. 320–331, 1989. 125

[157] C. L. Kutz, "Just disagreement: Indeterminacy and rationality in the rule of law," *The Yale Law Journal*, vol. 103, no. 4, pp. 997–1030, 1994. 125

[158] D. Freundlieb, "Epistemological realism and the indeterminacy of meaning. is systematic interpretation possible?" *Journal for General Philosophy of Science / Zeitschrift für allgemeine Wissenschaftstheorie*, vol. 22, no. 2, pp. 245–261, 1991. 125

[159] P. L. Peterson, "Semantic indeterminacy and scientific underdetermination," *Philosophy of Science*, vol. 51, no. 3, pp. 464–487, 1984. 125

[160] C. Churchman, "Problems of value measurement for a theory of induction and decisions," in *Proceedings of the 3rd Berkeley Symp. on Mathematical Statistics and Probability*, 1956, pp. 53–59. 125

[161] S. Jagabathula and D. Shah, "Inferring rankings under constrained sensing," in *Proceedings of Neural Information Processing Systems*, Vancouver, B.C., Dec. 2008. 127

[162] S. Jagabathula, V. Farias, and D. Shah, "A nonparametric approach to modeling choice with limited data," *Management Science*, submitted. 127

[163] J. Borwein and A. Lewis, *Convex Analysis and Nonliear Optimization: Theory and Examples*, 2nd ed. Springer, 2006. 131, 132

[164] D. Bertsekas, *Nonlinear Programming*, 2nd ed. Athena, 1999. 131

[165] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004. 131

[166] C. Churchman, *Thought and Wisdom*. Seaside, CA: Intersystems Publications, 1982. 144