Learning Convex Sets of Probability from Data

Fabio Cozman Lonnie Chrisman
e-mail: fgcozman@cs.cmu.edu, chrisman@lumina.com

CMU-RI-TR 97-25

The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213

June 20, 1997

Abstract

Several theories of inference and decision employ sets of probability distributions as the fundamental representation of (subjective) belief. This paper investigates a frequentist connection between empirical data and convex sets of probability distributions. Building on earlier work by Walley and Fine, a framework is advanced in which a sequence of random outcomes can be described as being drawn from a convex *set* of distributions, rather than just from a single distribution. The extra generality can be detected from observable characteristics of the outcome sequence. The paper presents new asymptotic convergence results paralleling the laws of large numbers in probability theory, and concludes with a comparison between this approach and approaches based on prior subjective constraints.

This research is supported in part by NASA under Grant NAGW-1175. Fabio Cozman was supported under a scholarship from CNPq, Brazil.

1 Introduction

This paper investigates the possibility of learning convex sets of probability distributions from data. Several theories of inference and decision employ sets of probability distributions as the fundamental representation of beliefs: in robust Statistics [1, 10], in relation to inner/outer measures for representation of subjective beliefs [7, 20, 24], as more flexible and general measures of uncertainty [2, 4, 5, 6, 9, 15, 21, 23, 25]. Usually such sets of distributions represent *subjective* opinions and preferences, and the indeterminacy of beliefs is *epistemic*.

Frequentist models depart from subjective interpretations and relate probability to observable phenomena, whereby an underlying probability reveals itself by way of asymptotic relative frequencies. This paper examines an analogous connection between convex sets of probability and observed outcome sequences. From an infinitely long sequence of outcomes, we attempt to recover the underlying convex set of distributions from which the data was generated. Our asymptotic results parallel and generalize the laws of large numbers used in probability theory. Existing literature does not provide an organized collection of asymptotic results for convex sets of distributions. The first results of this kind were proposed by Walley and Fine [27], and this paper can be understood as an adaptation of their results to more practical scenarios. The goals of our paper are:

1. To provide background on the theory of convex sets of distributions and motivation (Sections 2 and 3).
2. To describe a framework in which data can be viewed as being "generated" from an underlying convex set of distributions (Section 4).
3. To clearly define the notion of an estimator for convex sets of distributions (Section 5)
4. To describe Walley and Fine's estimator in an accessible fashion, and to improve upon it (Section 6).
5. To present new classes of estimators with asymptotic convergence results (Section 7).
6. To compare this approach to approaches that learn sets of distributions with prior subjective constraints (Section 8).

The paper presents novel asymptotic results (Section 7), which can be viewed as laws of large numbers for convex sets of distributions. The results show that by examining a finite number of subsequences of the observed trials, it is possible to learn a set of distributions that is guaranteed to dominate the set that generated the data. The theorems show how any estimator, including Walley and Fine's estimator, can be improved upon; our estimators lead to more realistic characteristics than Walley and Fine's estimator.

2 Convex sets of distributions

We consider theories that use convex sets of distributions to represent beliefs and to evaluate decisions. The set of distributions maintained by an agent is called the *credal* set [15]. To simplify terminology, we use the term credal set only when it refers to a set of distributions containing more than one element. Convex sets of conditional distributions are used to represent conditional beliefs. Inference is performed by applying Bayes rule to each distribution in a prior credal set; the posterior credal set is the union of all posterior distributions.

Given a credal set K, a probability interval can be created for every event A by defining lower and upper bounds, called the lower and upper envelopes:

$$\underline{p}(A) = \inf_{p \in K} p(A) \qquad \overline{p}(A) = \sup_{p \in K} p(A).$$

We say that a probability distribution $p(\cdot)$ dominates a lower envelope $\underline{p}(\cdot)$ if $p(A) \geq \underline{p}(A)$ for every event A.

We can also define an expected utility interval for every utility function $u(\cdot)$:

$$\underline{E}[u] = \inf_{p \in K} E_p[u] \qquad \overline{E}[u] = \sup_{p \in K} E_p[u]$$

Since utility functions induce expected utility intervals, it may be the case that decisions are incomparable (the ordering of possible decisions is a partial order) [15].

The upper envelopes and expectations can be obtained from the lower envelopes and expectations respectively. We have $\overline{p}(A) = 1 - \underline{p}(A^c)$ and $\overline{E}[u] = -\underline{E}[-u]$ for any event A and utility $u(\cdot)$.

Convex sets of distributions are interesting for several reasons, ranging from mathematical elegance to practical considerations of robustness (for an extensive discussion of this topic, consult Walley [25]). One of the common justifications is that assumptions of Bayesian theory are too strict: how can a real agent be required to specify a single number when explaining beliefs?

3 Interpretations of credal sets

Interpretations of probability often emphasize a frequentist approach, where probability is (only) a limiting frequency ratio. Another view is subjectivist, where probabilities are degrees of belief without necessarily having any physical manifestation.

Most existing interpretations for credal sets fall squarely in the subjective regime (the same holds for related systems such as belief functions, etc.). The fact that probabilities can be directly related to observed frequencies gives probability a significant advantage over other subjective representations of belief. For example, as a result of this relationship, decision analysts are often able to measure the *calibration* of an expert's subjective assessments [16]. The lack of a similar connection to observable physical outcomes for credal sets is a troublesome deficiency for most existing theories. Only a few works have attempted to make such connections, most notably the work of Kyburg [13, 14], which proposes specific guidelines to transform finite data knowledge into intervals of probabilities; the work of Seidenfeld and Schervish [22] on the convergence properties of beliefs in a group of agents; and the work of Walley and Fine [27] on estimators for sets of distributions.

Is it possible to relate a convex set of distributions to observable repeated outcomes in a manner analogous to the relationship between probabilities and frequencies? Can credal sets similarly be induced from a limiting series of observations in a meaningful fashion? Results by Walley and Fine [27] prove that such a connection is indeed possible. In this paper, we explain, build upon and extend these results, and we present interpretations of the mathematical results that are both useful and understandable.

With these results, interpretations of credal sets can, like interpretations of probability, have an additional grounding in observable phenomena, making notions such as calibration meaningful even for credal sets.

4 Estimating a credal set

Our learning theorems (and Walley and Fine's theorems) are generalizations of various law of large numbers theorems. Just as a probability can be induced from the frequencies on an infinite sequence of independent and identically distributed (i.i.d.) outcomes, our results express the idea that a credal set can be induced from an *infinite* sequence of outcomes. We emphasize that the current theorems are only limiting results, with finite sample cases being deferred for future research.

We begin with some examples that highlight the subtleties of our task.

Example 1 *Consider a coin where the bias is regulated by an extraneous mechanism, which we call "nature". In a sequence of coin flips, the first, third, fifth, and successive odd flips land heads with probability 0.6, while on the second, fourth, sixth, and successive even flips land heads with probability 0.4.*

In this case, "nature" is choosing the bias of the coin from the probability interval $[0.4, 0, 6]$ in a deterministic fashion. An estimation task would be to recover this interval from the infinite series of flips.

Example 2 *Consider a slightly different sequence of coin flips. Suppose "nature" chooses each distribution for each trial independently from a uniform distribution ranging from* 0.4 *to* 0.6*.*

In this case, the trial outcomes are actually i.i.d., and a point probability 0.5 would accurately describe the sequence. Thus, although one could say "nature" is drawing from a credal set, in this example we have "nature" drawing samples from a single probability distribution. We have constructed a hierarchical model for an i.i.d. point probability.

These examples illustrate that a credal set may or may not reveal itself through a sequence of trials, even an infinite one. Therefore, the goal to recover *the* underlying credal set precisely would be ambiguous. We can still require no estimate to contain distributions that are not in the underlying credal set. This establishes our first requirement: any estimate for a lower envelope must dominate the lower envelope for the underlying credal set.

Examples 1 and 2 share an important characteristic. Suppose one measures the relative frequency of heads as the number of coins goes to infinity. In both cases the relative frequency of heads approaches 0.5. The next example displays a situation where this does not occur.

Example 3 *Suppose we observe a sequence of coins passing by on a conveyor belt. Each coin is being placed heads-up with a probability between* $1/3$ *and* $2/3$*. In fact, it turns out "nature" is picking these probabilities in a very regular fashion. For the* i^{th} *coin, if the second most significant bit of* i*, when* i *is written in binary, is* 0*, the coin is heads with probability* $1/3$*, while if it is* 1*, the coin is heads with probability* $2/3$*. For example, since* 634 *(*$= 1\underline{0}1111010_2$*) has a zero for the second most significant bit, the* 634^{th} *flip is heads with probability* $1/3$*.*

Many people's initial intuition is that the relative frequency also approaches 0.5 in this example (as "half" the coins have a bias of $1/3$, the other "half" a bias of $2/3$). However, this sequence of coins does *not* have a unique converging relative frequency. Call the relative frequency for heads at coin 2^n by r'_n. Then $\lim_{n\to\infty} r'_n = 1/2$ with probability 1. On the other hand, call the relative frequency for heads at coin $2^n + 2^{n-1}$ by r''_n. Then $\lim_{n\to\infty} r''_n = 4/9$. Depending on the way we generate *subsequences of relative frequencies*, we may get different converging relative frequencies. We conclude that a credal set may create infinite sequences of trials that *cannot* be represented by any probabilistic model (a single probability distribution cannot generate a sequence with more than one converging relative frequency).

We now formalize the concepts introduced by these examples. Section 4.1 discusses our model for how a sequence is drawn from a credal set. The use of credal sets enriches the basic notion(s) of statistical guarantees, and these generalized notions are discussed in Section 4.2. Section 5 then considers our estimation goal, i.e., what it would mean to estimate that credal set from a sequence of observations. Walley and Fine [27] constructed such an estimator; we present their estimator and results in Section 6.

4.1 Data generation assumptions

Our data generation assumptions (taken from Walley and Fine) are as follows. For the i^{th} trial of the observed sequence, "nature" selects an underlying probability distribution, π_i. "Nature" may select a different distribution for different trials, i.e., it is possible that $\pi_i \neq \pi_j$. The manner in which these trial distributions are selected is not known to us; it may follow an (unknown) deterministic pattern (examples 1 and 3), there may be elements of randomness involved (example 2), and/or they may depend on actual previous outcomes. While no assumptions are made regarding how "nature" selects each trial distribution, we do assume that *every trial distribution is contained within a fixed credal set.* Once "nature" has selected a sequence of distributions, the individual trials are drawn independently and randomly from their corresponding distributions ($x_i \sim \pi_i$).

One may interpret the credal set as the most basic model of uncertainty and the selected distributions just as an explanatory device. A different interpretation is that there is a single distribution regulating the data, and this distribution is contained in the credal set [17]. Then our assumptions can be framed as a relaxation of the usual i.i.d. assumption for point probability. In this interpretation, while the trials are independent given the trial distributions, the underlying trial distribution would not have identically distributed marginals, and these marginals would need not be mutually independent.

One can see that our data generation assumptions are in fact appropriate for various physical phenomena. For example, the bias on the rolls of a die may slowly vary or oscillate by small amounts over time as the sides and corners become worn. It has been argued that the actual physical behavior of atomic clocks exhibits a similar type of non-stationarity that is most faithfully modeled by these assumptions [12, 8, 5].

Rather than view "nature" as actually drawing samples according to credal sets, the subjectivist may view the data generation somewhat differently. There are variables whose outcomes are to be assessed prior to observing the actual outcomes. However, due to lack of time or other factors, the assessments are to be completed without elaborating a full detailed model of the interactions or correlations between the variables. This interpretation of convex sets of probability is referred to the *ontological interpretation* in previous research [3, 27]. As actual values for the variables become observed, it is as if the values have been drawn from the perfectly calibrated subjectivist's belief set. In this way, inducing the underlying convex set of probabilities from an infinite observed sequence of data is equivalent to determining whether an agent's subjective interval-valued belief is properly calibrated.

4.2 Asymptotic certainty and favorability

In classical probability theory, asymptotic certainty is at the core of central limit theorems. For example, if a fair coin is flipped infinitely often, the frequency of heads will approach 0.5 *with asymptotic certainty*. This leaves open the possibility that a very unusual sample is generated by random chance, although as the length of the sequence grows, the chance of usual events become less and less significant.

Alternate versions of this type of limiting guarantee can be defined in the framework of convex probability. The two concepts of primary interest are asymptotic certainty and asymptotic favorability.

Let $\{A_1, A_2, \ldots\}$ be a sequence of events (an event here is a combination of outcomes that either occurs or does not occur when the sequence is generated). When $\lim_{n\to\infty} \underline{p}(A_n) \longrightarrow 1$, it is said that A is *asymptotically certain*, or "a.c." In this case, no matter what strategy "nature" uses to choose trial distributions, A will occur in the limit.

A weaker notion of convergence is also useful. When $\lim_{n\to\infty} \underline{p}(A_n^c)/\underline{p}(A_n) \longrightarrow 0$, where A_n^c denotes the complement of A_n, it is said that A is *asymptotically favored* (a.f.) [27]. For a point probability, asymptotic favorability and asymptotic certainty correspond. In general, asymptotic certainty implies asymptotic favorability; a.f. is much weaker than a.c. In terms of a credal set, asymptotic certainty of an event A means that, for all distributions $p(\cdot)$ in the credal set, $p(A)$ tends to 1; asymptotic favorability of an event A means that some distributions in the credal set have $p(A)$ tend to 1 and other distributions may have $p(A)$ tend to some non-negative number smaller than 1. Informally, asymptotic favorability only ensures that it is plausible that A occurs with probability 1, but this occurs only if "nature" happens to select trial distributions with the appropriate strategy (a "cooperative nature").

The concepts of a.c. and a.f. are most commonly applied to describe guarantees on sample statistics or estimators, by saying that statistic F will have property A with asymptotic certainty or favorability.

5 The estimation task

A naive description for the estimation task would be to recover (learn) a.c. the underlying credal set from an infinite sequence of outcomes. As examples 1 and 2 show, an underlying credal set does not necessarily reveal itself in any single infinite sequence of trials. In simple terms, this is just because our very loose assumptions about data generation have left totally open the manner in which "nature" selects individual trial distributions. The deeper ramifications of this are reflected in a series of estimation results [27, Theorems 5.1-5.4], which state that it is not possible to detect the full extent of the underlying credal set with asymptotic certainty, although it can be done with asymptotic favorability (i.e., if you happen to be fortunate).

We keep the requirement that a good estimator must produce estimates which dominate the lower envelope for the underlying credal set. This means that the estimated credal set is smaller than the underlying credal set; our requirement is that the estimate does not contain any distribution that is outside the credal set that generated the data.

Given two estimators that always dominate a credal set, which is best?

Even if an estimator asymptotically favors the underlying credal set and guarantees a dominating credal set with asymptotic certainty, this does not mean it is the best possible estimator. It is possible to have two different estimators, both with these properties, producing distinct credal sets from the same sequence. Often these estimators will be incomparable (in which case an even better estimator can be obtained using our Theorem 1). However, it is also possible that the first estimator will always dominate the second (a.c.). If this is the case, the second estimator, which is consistently dominated, is a better estimator. This is because both are guaranteed the dominate the underlying distribution, but the second estimator's resulting credal set will be larger, and therefore closer to the true underlying credal set.

In fact, if data is generated by any non-vacuous credal set, K, it is possible to construct a mathematically equivalent generator using any credal set dominated by K (i.e., larger than K) with a simple alteration to the method for selecting the distributions. In so far as a credal set (partially) summarizes the data generation process, one would always have the option of reducing the information content of the summarization by loosening the bounds. From all these equivalent generators, we are interested in the credal set conveying the most informative description of the data generation process — i.e., the credal set that dominates the others.

In short, our requirements are as follows. From an infinite sequence of outcomes, we desire an estimator that is guaranteed to dominate the underlying credal set with asymptotic certainty and contains as many distributions as possible.

6 Walley and Fine's estimation task

The estimation task that we have identified in the previous section differs somewhat from the estimation problem solved by Walley and Fine [27], although the assumptions behind data generation are identical. The difference is subtle, but important for avoiding confusion and fully understanding the results in this area. Although Walley and Fine's formulation has a mathematical elegance in that it allows them to identify the optimal estimator (in their sense), we believe the objective we have outlined is more representative of what one is pragmatically interested in learning in the framework of convex sets of distributions.

From a sequence of outcomes, $x_1, x_2, \ldots$, one can construct a sequence of relative frequencies, $r_1(A), r_2(A), \ldots$. Here $r_n(A)$ is the frequency of occurrences of event A during the initial n trials of the sequence. Rather than estimating the credal set that generates the data, Walley and Fine's characterize all possible subsequences of relative frequencies. For example, suppose one considers only the subsequence of odd frequencies, $r_2, r_3, r_5, r_7, r_{11}, r_{13}, r_{17}, \ldots$, and that this subsequence converges to a limiting frequency. Walley and Fine give estimators that capture this limiting frequency with asymptotic certainty. Popper [18, Section 63-66] calls these limiting frequencies "middle frequencies", and points out that sequences may have multiple middle frequencies.

Note that example 3 involves exactly this type of construction. Walley and Fine emphasize estimation with this type of sequence. On the other had, example 1 generates a sequence with a single middle frequency and does not produce a credal set estimate with Walley and Fine's approach.

One way to state the difference between our task and Walley and Fine's task is that we are interested in limiting frequencies for subsequences of outcomes, while Walley and Fine (and Popper) gave estimators for limiting points of subsequences of frequencies. The former approach characterizes the sequence of outcomes, and relates directly to the underlying credal set that generates the data, while the latter approach is a characterization of the sequence of relative frequencies, and how the sequence of frequencies may not converge in the classical sense. Walley and Fine's objectives can be pursued by throwing away all information contained in the sequence of actual outcomes, keeping only the sequence of frequencies.

It is not hard to see that an estimator for Walley and Fine's task is an estimator for our task. However, for our task, a Walley and Fine estimator can often be substantially improved upon.

Walley and Fine propose the following estimator for lower envelopes. Consider a sequence of trials $\{x_1, x_2, \ldots\}$. For any event A, the relative frequency $r_i(A)$ is the number of positive trials for A up to trial i. From the original sequence $\{x_1, x_2, \ldots\}$, we can compute a sequence of relative frequencies $\{r_1(A), r_2(A), \ldots\}$.

Walley and Fine define a class of estimators for the lower envelope having the following form:

$$\underline{r}(A) = \lim_{n \to \infty} \inf \{r_i(A) : k(n) \leq i \leq n\}, \tag{1}$$

where $k(n)$ is any function with the properties that $\lim_{n \to \infty} k(n) \longrightarrow \infty$ and $\lim_{n \to \infty}(k(n)/n) = 0$. For example, $k(n) = \sqrt{n}$ yields one such estimator.

The lower envelope formed through Walley and Fine's estimator can be extended to a convex set (the set of all distributions that dominate these estimates). Walley and Fine prove that this set dominates the credal set that generated the data [27, Theorem 4.1]. The dual upper envelope estimator is obtained by replacing the infimum with a supremum in (1).

For many of us who are used to thinking about relative frequencies in terms of single distributions and i.i.d. trials, the intuition behind Walley and Fine's estimator can be quite difficult to grasp. For example, $r_{k(n)}$ and r_n both converge to the relative frequency of the infinite sequence of trials, making it non-intuitive why looking at them together should uncover more information about the mechanism generating the data than simply looking at the common limiting relative frequency.

In fact, instead of looking at just the limiting relative frequency of an infinite sequence, Walley and Fine's estimator simultaneously considers the whole set of possible limiting frequencies. If observations were being generated by an a single distribution through an infinite sequence of i.i.d. trials, each relative frequency in this set would converge (by the law of large numbers) to the same limiting relative frequency, making $r_{k(n)} = r_{k(n)+1} = \ldots = r_{i-1} = r_i$, and making the infimum in (1) uninteresting. When we drop the single distribution and i.i.d. assumptions, the estimates become far richer.

Walley and Fine [27, Theorem 4.1(a)] prove that their estimator produces a credal set that dominates the underlying credal set with asymptotic certainty. Their estimator will detect divergence from i.i.d. point probability with asymptotic favorability (their Theorem 4.1(d)). Also, any convergence subsequence of $r_1, r_2, .$ converges to a frequency contained in their estimate, and their estimate is the smallest credal set for which this is true (their Theorem 4.1). Thus, in a certain sense, their estimator optimally characterizes the asymptotic divergence of relative frequencies for a given sequence of outcomes.

The astute reader, however, will notice that Walley and Fine's estimator des not recover "nature's" exact credal set in this example in our previous examples. Since our goal is to recover, as best we can, "nature's" underlying credal set, there is clearly room for improvement. We now study Walley and Fine's estimator from a slightly different perspective, which helps clarify our own approach to the problem.

The above interpretation of Walley and Fine's estimator built an analogy between the estimator and the minimum of a sequence of estimates, $\{r_{k(n)}, \ldots, r_n\}$. This is a translation of expression (1) and serves the purpose of clarifying the logic behind the estimator.

Walley and Fine's estimator can also be described as the minimum estimate produced by a generator of sub-sequences; this is the description that interests us in this paper.

Consider an infinite sequence of trials $X = \{x_1, x_2, \ldots\}$. Consider a generator of selection rules, i.e., an algorithm that generates infinite sub-sequences X^s out of the sequence X, by specifying members of X that must also be members of X^s. Take the following algorithm, which generates subsequences for a given n:

- For all k from $k(n)$ to n, produce the sub-sequence $X^k = \{x_1 \ldots x_k\}$.

Each sub-sequence X^k has its relative frequency. Suppose now that $n \to \infty$. If the original sequence X has multiple converging frequencies, these frequencies must be attained in some of the sub-sequences and their minimum will be captured by Walley and Fine's estimator. The way to produce multiple converging frequencies was illustrated in the coin example: progressively "longer infinite" numbers of trials must be used to generate each one of the frequencies.

This procedure reveals the drawback of Walley and Fine's approach. The problem is that their estimator is geared toward capturing all possible limiting frequencies, regardless of the types of sequences it may find. Rarely a sequence of data will be maliciously produced as in the coin example, with progressively "longer infinite" segments generated by different probabilities. In general we expect trials to be generated by selecting distributions from the credal set in some defined, deterministic way, and then generating the data from the selected distributions. This is the most relevant situation in practice, where we are interested to assess how much our assumptions of randomness, and our abstractions in the modeling process, are justified.

The main goal of this paper is to develop estimators that are suited to deal with the situation described above. Suppose that we have some deterministic procedure selecting distributions. First suppose, for the sake of argument, that we know the deterministic procedure. For example, odd trials obey one distribution, even trials obey another. Then the logical way to proceed is to partition the data into even and odd sub-sequences and estimate relative frequencies in each one of them. This agrees with intuition and with statistical practice: if we suspect differences between blocks of data, we must run some form of cross-validation among the inferences obtained from different blocks.

Now suppose the distribution selection mechanism is unknown. We still proceed the same way, by partitioning the data into several sub-sequences, in the hope of matching the data patterns. If we take the lower bounds of the collections of relative frequencies that emerge, we obtain estimators that can capture aspects of the data that are not captured by Walley and Fine's estimator. The learning theorem proved in the next section demonstrates that this procedure in fact creates credal sets that dominate the "true" credal set with asymptotic certainty. We also indicate how to combine our procedure with Walley and Fine's estimator so as to improve both estimators.

7 The finite learning theorem for convex sets of distributions

We create a family of estimators whose main purpose is to capture aspects of a sequence of trials that cannot be captured by Walley and Fine's estimator. Consider an infinite sequence of trials $X = \{x_1, x_2, \ldots\}$. Each trial for event A is generated with a distribution $p(A)$ such that $p(A) > \underline{p}(A)$. Consider a sub-sequence X^s out of the sequence X. We assume that the probability of any trial is unaffected by the selection mechanism: $p(x_i \in A | x_i \in X^s) = p(x_i \in A)$. We must place restrictions on the possible sub-sequence selection rules, because we cannot select trials "after we see" the results. Otherwise it would be possible to construct a sub-sequence with only heads or only tails in the coin example. We must be able to specify sub-sequences in some definite way which cannot affect nor be affected by the trials. The definition of a sub-sequence generator that complies with such requirements can be taken from the theory of random numbers, where selection rules are studied to great length. We adopt the definitions of *computable selection rules* given by Knuth [11] to indicate which entities we consider. We assume that sub-sequences are defined such that there are infinitely many elements in each sub-sequence for an infinitely long original sequence.

To prove the main theorem, we need the following result from Walley and Fine [27, Lemma3.2 summarized]:

Lemma 1 *Suppose $\underline{p}(\cdot)$ is a lower envelope in a space with a finite number of elements, and we choose a collection of events A_{jn} for $n \geq 1$ and $1 \leq j \leq J$, such that the events A_{jn} are a.c. as $n \to \infty$. Then the finite intersection $\cap_j A_{jn}$ is a.c. as $n \to \infty$.*

Since the Lemma fails for countably infinite collections of events, we cannot extend our main theorem to countably infinite collections of events with the same tools used in this proof. Whether this can be done with other techniques is an open problem.

Now we can prove:

Theorem 1 (Finite Convex Learning Theorem) *For a given algorithm s as specified above, define:*

$$r_n^s(A) = \frac{|\{i : 1 \leq i \leq n, x_i^s \in A\}|}{n^s}.$$

where n^s is the number of elements in s up to trial n. Let S be a finite set of algorithms s; define

$$\underline{r}_n^S(A) = \min_{s \in S} r_n^s(A),$$

then $\underline{r}_n^S(\cdot)$ dominates a.c. (as $n \to \infty$) the lower envelope that generated the original sequence.

Proof.

Call $\underline{p}(\cdot)$ the lower envelope that generated the data. The conjugate upper envelope is $\overline{p}(\cdot)$.

First, $\underline{r}_n^S(\cdot)$ is a lower envelope: the lower envelope of $r_n^s(\cdot)$ for all $s \in S$. Now, take each algorithm s_j from S. Each s_j defines an infinite sequence of trials with probability larger than $\underline{p}(A)$ for each event A. Now apply Theorem 4.1.a from Walley and Fine [27] on each sub-sequence:

$$\forall \eta > 0, \; \bigcap_{A \in \mathcal{A}} [\overline{p}(A) + \eta > r_n^{s_j}(A) > \underline{p}(A) - \eta] \text{ a.c. as } n \to \infty \text{ under } \underline{p}^{\infty}(\cdot).$$

In other words, for large enough n, the value of $r_n(A)$ will (almost always) be within $\underline{p}(A)$ and $\overline{p}(A)$. So the event $\{r_n^{s_j}(A)$ almost within $\underline{p}(A)$ and $\overline{p}(A)\}$ is a.c.

Now due to Lemma 1, we know that as $n \to \infty$, the event

$$\left\{ \bigcap_{s_j \in S} r_n^{s_j}(A) \text{ almost within } \underline{p}(A) \text{ and } \overline{p}(A) \right\}$$

is a.c.; so the event

$$\left\{ \min_{s_j \in S} r_n^{s_j}(A) \text{ almost within } \underline{p}(A) \text{ and } \overline{p}(A) \right\}$$

is a.c.. □

This result suggests that, given a finite space of events and a sequence of trials, it is possible to find estimates for the lower envelope of the distributions.

A drawback of the estimator $r_n^S(\cdot)$ is that it may not capture *all* limits of pointwise convergent sub-sequences of relative frequencies. The following result indicates how to solve this problem.

Theorem 2 *The estimator defined by $r_n^{S'}(A) = \min(r_n^S(A), r_n(A))$, where $r_n^S(A)$ is defined in the previous theorem, and $r_n(A)$ is Walley and Fine's estimator, dominates a.c. (as $n \to \infty$) the lower envelope that generated the original sequence and contains the lower envelope of all limits of pointwise convergent sub-sequences of relative frequencies.*

So far the discussion has concentrated on the estimation of lower envelopes. A lower envelope corresponds to an infinite number of convex sets of distributions, so statements about estimation of convex sets are stronger than statements about lower envelopes. To be able to attack this problem, we note that there is a one-to-one correspondence between credal sets and lower expectations [25]. If we can estimate lower expectations, we can recover the underlying (unique) convex set of distributions. Walley and Fine also approach this problem and prove that for a measurable utility function $u(\cdot)$:

$$\forall \eta > 0, \quad \left[\overline{E}[u] + \eta > \frac{\sum_{i=1}^n u(x_i)}{n} > \underline{E}[u] - \eta \right] \text{ a.c. as } n \to \infty \text{ under } \underline{p}^\infty(\cdot).$$

We can use this result and adapt our Theorem 1 to obtain:

Theorem 3 *For a given algorithm s as specified above, define:*

$$E_n^s[u] = \frac{\sum_{i=1}^n u(x_i)}{n^s},$$

the sample average of $u(x)$ in the sub-sequence s. Let S be a finite set of algorithms s; define

$$\underline{E}_n^S[u] = \min_{s \in S} E_n^s[u],$$

then for all $\eta > 0$ the event $\left\{ \overline{E}[u] + \eta > E_n^S[u] > \underline{E}_n^S[u] - \eta \right\}$ is a.c. (as $n \to \infty$).

8 Comparison with subjective learning of convex sets of distributions

This paper concentrates on the connection between data and convex sets of distributions. Results do not use the possible existence of prior distribution for events. An alternative approach is to use prior distributions and Bayes rule to obtain posterior measures for events. The objective of this paper is not to replace Bayes rule, but rather to enhance one's intuition about probabilities constructed solely from data. There has been work on subjective approaches to the process of learning convex sets of distributions; we mention two approaches that are relevant to Bayesian networks.

8.1 Ramoni and Sebastiani approach to missing data

The estimation of parameters for a Bayesian network usually has to deal with missing data, i.e., observations for some variables are not collected. The standard Bayesian assumption is that missing data happens at random; if this assumption is violated, inferences may be biased. Ramoni and Sebastiani propose to lift the "missing at random assumption" [19] in a Bayesian network learning scenario. They consider all possible ways in which missing data could have happened, and create a convex set of joint distributions that represent the gamut of possibilities for the data actually collected. The idea is to avoid using unjustified assumptions and replacing those by sets of distributions, so that the effects of missing data can be evaluated.

8.2 Walley's imprecise Dirichlet prior

The imprecise Dirichlet prior has been proposed by Walley [26] as a model for inferences associated with multinomial sampling. Here we indicate how this model can be used to learn Bayesian networks associated with convex sets of distributions.

An imprecise Dirichlet distribution for a vector valued variable θ is:

$$p(\theta) = \mathrm{Dir}(\theta|s,t) \sim \prod_{i=2}^{|\theta|} \theta_i^{st_i - 1},$$

where s is a real number larger than zero and t is a vector where $\sum t_i = 1$ and $0 < t_i < 1$ for all t_i.

This class of distributions can be used as a prior credal set; the prior assumptions are much less restrictive than standard Bayesian assumptions. Note that for any event A, the prior imprecise Dirichlet model induces the bounds $\underline{p}(A) = 0$ and $\overline{p}(A) = 1$.

First consider standard Bayesian network learning when complete data is available. A Bayesian network codifies a joint distribution through the expression:

$$p(\tilde{x}) = \prod_i p(x_i | \mathrm{pa}(x_i)),$$

where $\mathrm{pa}(x_i)$ are the parents of variable x_i. For each variable, the vector of parameters θ_i contains elements $\theta_{ijk} = p(x_i = k | \mathrm{pa}(x_i) = j)$, where $\theta_{ij1} = 1 - \sum_{k=2}^{|x_i|} \theta_{ijk}$. The vector $\theta_{ij} = \{\theta_{ijk}\}_{k=1}^{|x_i|}$ contains the relevant parameters for the distribution $p(x_i | \mathrm{pa}(x_i) = j)$. The vector $\Theta = \{\theta_1, \ldots, \theta_n\}$ contains all parameters to be estimated. The usual assumption for the prior $p(\Theta)$ is parameter independence:

$$p(\Theta) = \prod_{i=1}^{n} \prod_{j=1}^{\mathrm{pa}(x_i)} p(\theta_{ij}).$$

Finally, the prior distributions for each vector θ_{ij} are assumed to come from an imprecise Dirichlet family. The posterior is then an imprecise Dirichlet distribution with parameters that depend on the prior parameters and the data.

Suppose that every vector θ_{ij} is associated with an imprecise Dirichlet prior:

$$p(\theta_{ij}) = \mathrm{Dir}(\theta_{ij} | s_{ij}, t_{ij}) \sim \prod_{k=2}^{|x_i|} \theta_{ijk}^{s_{ij} t_{ijk} - 1},$$

where s_{ij} is a real number larger than zero and t_{ij} is a vector such that $\sum t_{ijk} = 1$ and $0 < t_{ijk} < 1$ for all t_{ijk}. We assume that the convex set of prior joint distributions is obtained by taking the convex hull of all prior marginals defined by imprecise Dirichlet distributions.

Suppose data n_{ij} observations are made with $\mathrm{pa}(x_i) = j$ and n_{ijk} observations are made with $x_i = k$, $\mathrm{pa}(x_i) = j$.

The posterior distribution for θ_{ij} is given by imprecise Dirichlet distributions, due to the parameter independence assumption and the convexification convention. We have θ_{ij} with marginals:

$$p(\theta_{ij} = \mathrm{Dir}(\theta_{ij} | s'_{ij}, t'_{ij}),$$

where $s'_{ij} = n_{ij} + s_{ij}$ and $t'_{ijk} = \frac{n_{ijk} + s_{ij} t_{ijk}}{n_{ij} + s_{ij}}$.

9 Conclusion

This paper advances a frequentist framework based on convex sets of probability distributions. From a sequence of outcomes generated by repetitive experiments, we are able to learn meaningful convex sets of probability distributions from the data. This learning is accomplished using estimators that examine relative frequencies over a finite collection of subsequences of the data. The estimators are guaranteed in a strong sense (i.e., with asymptotic certainty) to dominate the convex set of distributions that generated the data. Our theorems also demonstrate that any estimator based on a finite collection of subsequences can always be improved.

The work started by Walley and Fine and extended in this paper opens several important doors for advocates of belief representations based on convex sets of distributions. First, it demonstrates that these representations can actually be learned from observed data. Second, and perhaps most importantly, is that the connection to observed outcomes addresses what has been a critical weakness of these convex set representations. Bayesians have had the philosophical upper hand primarily because of the connection between probability and observed frequency. Among other things, this connection implies that it is possible to detect when a Bayesian degree of belief is or is not properly calibrated. No such notion has previously been possible for convex set representations of belief. We now know that the connection of subjective probability to observed frequencies is not exclusive property of the Bayesian interpretation, but can indeed be enjoyed by belief frameworks based on credal sets as well.

References

[1] J. O. Berger. Robust bayesian analysis: Sensitivity to the prior. *Journal of Statistical Planning and Inference*, 25:303–328, 1990.

[2] J. Cano, M. Delgado, and S. Moral. An axiomatic framework for propagating uncertainty in directed acyclic networks. *International Journal of Approximate Reasoning*, 8:253–280, 1993.

[3] L. Chrisman. Independence with lower and upper probabilities. *Proc. Twelfth Conference Uncertainty in Artificial Intelligence*, pages 169–177, 1996.

[4] L. Chrisman. Propagation of 2-monotone lower probabilities on an undirected graph. *Proc. Twelfth Conference Uncertainty in Artificial Intelligence*, pages 178–186, 1996.

[5] T. L. Fine. Lower probability models for uncertainty and nondeterministic processes. *Journal of Statistical Planning and Inference*, 20:389–411, 1988.

[6] F. J. Giron and S. Rios. Quasi-bayesian behaviour: A more realistic approach to decision making? In J. M. Bernardo, J. H. DeGroot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics*, pages 17–38. University Press, Valencia, Spain, 1980.

[7] I. J. Good. *Good Thinking: The Foundations of Probability and its Applications.* University of Minnesota Press, Minneapolis, 1983.

[8] Grize. *Towards a Stationary Continuous Lower Probability Based Model for Flicker Noise.* PhD thesis, Cornell University, 1984.

[9] J. Y. Halpern and R. Fagin. Two views of belief: Belief as generalized probability and belief as evidence. *Artificial Intelligence*, 54:275–317, 1992.

[10] P. J. Huber. *Robust Statistics.* Wiley, New York, 1980.

[11] D. E. Knuth. *The Art of Computer Programming*, volume 2. Addison-Wesley Pub. Co., Reading, Mass., 1973.

[12] Kumar and Fine. Stationary lower probabilities and unstable averages. *Z. Wahrsh. verw Gebiete*, 69:1–17, 1984.

[13] H. E. Kyburg Jr. *The Logical Foundations of Statistical Inference.* D. Reidel Publishing Company, New York, 1974.

[14] H. E. Kyburg Jr. Higher order probabilities and intervals. *International Journal of Approximate Reasoning*, 2:195–209, 1988.

[15] I. Levi. *The Enterprise of Knowledge.* The MIT Press, Cambridge, Massachusetts, 1980.

[16] M. G. Morgan and M. Henrion. *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis.* Cambridge, New York, 1990.

[17] A. Papamarcou and T. Fine. Unstable collectives and envelopes of probability measures. *Annals of Probability*, 19(2):893–906, 1991.

[18] K. R. Popper. *The logic of scientific discovery.* Harper, New York, 1965.

[19] M. Ramoni and P. Sebastiani. Robust learning with missing data. Technical Report KMI-TR-28, Knowledge Media Institute, The Open University, July 1996.

[20] E. H. Ruspini. The logical foundations of evidential reasoning. Technical Report SRIN408, SRI International, 1987.

[21] T. Seidenfeld. Outline of a theory of partially ordered preferences. *Philosophical Topics*, 21(1):173–188, Spring 1993.

[22] T. Seidenfeld and M. Schervish. Two perspectives on consensus for (bayesian) inference and decisions. *IEEE Transactions on Systems, Man and Cybernetics*, 20(1), 1990.

[23] C. A. B. Smith. Consistency in statistical inference and decision. *Journal Royal Statistical Society B*, 23:1–25, 1961.

[24] P. Suppes. The measurement of belief. *Journal Royal Statistical Society B*, 2:160–191, 1974.

[25] P. Walley. *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, New York, 1991.

[26] P. Walley. Inferences from multinomial data: Learning about a bag of marbles. *Journal Royal Statistical Society B*, 58(1):3–57, 1996.

[27] P. Walley and T. L. Fine. Towards a frequentist theory of upper and lower probability. *The Annals of Statistics*, 10(3):741–761, 1982.