tions: A critique and an annotated bibliography (with discussion). *Statist. Sci.* **1** 114–148.

Genest, C., Weerahandi, S. and Zidek, J. V. (1984). Aggregating opinions through logarithmic pooling. *Theory and Decision* **17** 61–70.

Halmos, Paul R. (1950). *Measure Theory*. Van Nostrand, New York.

Hewitt, Edwin and Stromberg, Karl (1965). *Real and Abstract Analysis*. Springer, New York.

Laddaga, Robert (1977). Lehrer and the consensus proposal. *Synthese* **36** 473–477.

Lindley, Dennis V. (1985). Reconciliation of discrete probability distributions. In *Bayesian Statistics* **2** (J. M. Bernardo, et al., eds.) 375–390. North-Holland, Amsterdam.

Madansky, Albert (1964). Externally Bayesian groups. Technical Report RM-4141-PR, RAND Corporation.

Madansky, Albert (1978). Externally Bayesian groups. Unpublished manuscript, University of Chicago.

McConway, Kevin J. (1978). The combination of experts' opinions in probability assessment: Some theoretical considerations. Ph.D. thesis, University College London.

McConway, Kevin J. (1981). Marginalization and linear opinion pools. *J. Amer. Statist. Assoc.* **76** 410–414.

Morris, Peter A. (1974). Decision analysis expert use. *Management Sci.* **20** 1233–1241.

Morris, Peter A. (1977). Combining expert judgments: a Bayesian approach. *Management Sci.* **23** 679–693.

Raiffa, Howard (1968). *Decision Analysis: Introductory Lectures on Choices under Uncertainty*. Addison-Wesley, Reading, Mass.

Wagner, Carl G. (1982). Allocation, Lehrer models, and the consensus of probabilities. *Theory and Decision* **14** 207–220.

Wagner, Carl G. (1984). Aggregating subjective probabilities: Some limitative theorems. *Notre Dame J. Formal Logic* **25** 233–240.

Wald, Abraham (1939). Contributions to the theory of statistical estimation and testing hypotheses. *Ann. Math. Statist.* **10** 299–326.

Weerahandi, Samaradasa and Zidek, James V. (1981). Multi-Bayesian statistical decision theory. *J. Roy. Statist. Soc. Ser. A* **144** 85–93.

Winkler, Robert L. (1968). The consensus of subjective probability distributions. *Management Sci.* **15** B61–B75.

Winkler, Robert L. (1981). Combining probability distributions from dependent information sources. *Management Sci.* **27** 479–488.

3.6

An Approach to Consensus and Certainty with Increasing Evidence

MARK J. SCHERVISH AND
TEDDY SEIDENFELD

ABSTRACT

We investigate conditions under which conditional probability distributions approach each other and approach certainty as available data increase. Our purpose is to enhance Savage's (1954) results, in defense of "personalism", about the degree to which consensus and certainty follow from shared evidence. For problems of consensus, we apply a theorem of Blackwell and Dubins (1962), regarding pairs of distributions, to compact sets of distributions and to cases of static coherence without dynamic coherence. We indicate how the topology under which the set of distributions is compact plays an important part in determining the extent to which consensus can be achieved. In our discussion of the approach to certainty, we give an elementary proof of the Lebesgue density theorem using a result of Halmos (1950).

I. INTRODUCTION

In his classic discussion of Bayesian inference, L. J. Savage (1954, Sections 3.6 and 4.6) illustrates how (finitely many) different investigators come to agree on the truth of one hypothesis, given an increasing sequence of shared observations. More precisely, Savage's result is this. Assume the following two conditions.

Reprinted from *Journal of Statistical Planning and Inference*, Volume 25, M. Schervish and T. Seidenfeld, "An Approach to Consensus and Certainty with Increasing Evidence," pp. 401–414, Copyright 1990, with kind permission from Elsevier Science – NL, Sara Burgerhartstraat 25, 1055 KV Amsterdam, The Netherlands.

1. The agents' initial opinions over a (finite) set of rival hypotheses are not too discrepant – there is agreement on which hypotheses have probability 0.
2. There is agreement also on the (distinct) likelihoods for these hypotheses over an infinite sequence of observations which are identically, independently distributed given an hypothesis.

Then, almost surely, the sequence of conditional probabilities for the hypotheses *converge* (given more of these data) *to a common*, 0–1 distribution focused on the true hypothesis.

Savage offers this finding as a partial rebuttal to the charge that his Bayesian theory of *personal* probability is overly "subjective" – that it cannot explain how scientific methods may be a source of "objective" knowledge. In short, he uses this result to explain how interpersonal agreements about what is practically certain can arise within the Bayesian paradigm. The object of our discussion here is to indicate how the two parts to Savage's conclusion – *consensus* and *certainty* of opinions – obtain (asymptotically) under more general conditions than are permitted by (1) and (2). That is, we indicate how Savage's reply may be enhanced.

The first part of Savage's conclusion, (almost certain) consensus for a pair of agents, does not depend upon the assumptions (2) so long as the hypotheses of interest are expressible in terms of (perhaps infinite) sets of observables, as was shown by Blackwell and Dubins (1962). (Also, like Blackwell and Dubins, we avoid Savage's restriction to conditional probability given non-null events but, instead, we require that probabilities are countably additive, unlike in Savage's argument.) We discuss how compactness of a set C of (mutually, absolutely continuous) probabilities affects the kind of consensus that may be achieved with increasing shared evidence. When the extreme points of C are compact in the discrete topology, consensus ("almost everywhere") follows from the theorem of Blackwell and Dubins: Corollary 1. When the extreme points of C are compact in the uniform-distance topology, convergence ("in-probability", but not "almost everywhere") of sequences of pairs of probabilities is proven: Corollaries 2, 3, and Example 3. And when the extreme points of C are weak-star compact, no consensus is assured at all: Example 2.

The second part of Savage's conclusion, that the conditional probabilities of a measurable event E converge (almost surely) to 1 or 0 as E occurs or fails to occur, follows from coherence alone. We offer two arguments for the approach to certainty. One proof (which we suspect is known to many) is as a consequence of Doob's martingale convergence theorem. The other argument uses just a basic result governing the extension of σ-finite measures from an algebra to its smallest σ-algebra. (A similar result is needed, also, in the application of Doob's martingale theorem.) We show how the approach to certainty for conditional probabilities provides an elementary proof of the Lebesgue density theorem: Corollary 6. Example 4 illustrates the importance of the measurability assumption for certainty, even with exchangeable probability distributions.

In addition, we show that consensus and certainty, viewed as claims about the asymptotic behavior of the agents' unconditional probabilities with increasing data, do not require the full force of temporal conditionalization – they do not require the use of Bayes' theorem to revise degrees of belief.[1] That is, though we do require (static) coherence, we do not assume that, over time, an agent updates his personal probability by conditionalization. We do not impose a constraint of (full) dynamic coherence. For consensus, it suffices that the agents use conditional probabilities arbitrarily chosen from a class C enveloped by finitely many (mutually absolutely continuous) distributions. Under the conditions of Corollary 1, asymptotic certainty follows from static coherence: Corollary 4.

II. THE STRUCTURAL ASSUMPTIONS FOR THE SPACE $(X, \mathcal{B}, P)$

II.1. *The Measurable Space* $(X, \mathcal{B})$

Consider a denumerable sequence of sets X_i $(i = 1, \ldots)$ with associated σ-fields $\mathcal{B}_i$. Form the infinite Cartesian product $X = X_1 \times \ldots$ of sequences $(x_1, x_2, \ldots) = x \in X$, where $x_i \in X_i$, that is, where each x_i is an atom of its algebra $\mathcal{B}_i$. (This is mild as the $\mathcal{B}_i$ may be unrelated.) In the usual fashion, let the measurable sets in X (the events) be the elements of the σ-algebra $\mathcal{B}$ generated by the set of measurable rectangles. (A measurable rectangle $A = A_1 \times \ldots$ is one where $A_i \in \mathcal{B}_i$ and $A_i = X_i$ for all but finitely many i.) Thus, $(X, \mathcal{B})$ is a measurable space.

Define the spaces of histories $(H_n, \mathcal{H}_n)$ and futures $(F_n, \mathcal{F}_n)$ where $H_n = X_1 \times \ldots \times X_n$, $\mathcal{H}_n = \mathcal{B}_1 \times \ldots \times \mathcal{B}_n$, and where $F_n \times X_{n+1} \times \ldots$ and $\mathcal{F}_n = \mathcal{B}_{n+1} \times \ldots$. Identify these as sub-$\sigma$-fields of $(X, \mathcal{B})$ by writing

$G_n \in \mathcal{H}_n$ as $G_n \times X_{n+1} \times \ldots$ and $E_n \in \mathcal{F}_n$ as $X_1 \times \ldots \times X_n \times E_n$. We shall be concerned, in particular, with the (increasing) sequence of histories $h_n \in H_n$ and the (decreasing) sequence of future events $E_n \in \mathcal{F}_n$, as these are judged given each history.

II.2. *The Probability* P

Let P be a (countably additive) probability over the measurable space $(X, \mathcal{B})$. Assume P is *predictive* (Blackwell and Dubins, 1962), so that there exist conditional probability distributions of events given past events, $P^n(\cdot|\mathcal{H}_n)$.[2] In particular, given a history h_n, there is a conditional probability distribution for the future, $P^n(\cdot|h_n)$ on $\mathcal{F}_n$.

Next, we show that conditional probability given some history, $P^n(\cdot|h_n)$ on $\mathcal{B}$, is characterized by the conditional probability for the future, $P^n(\cdot|h_n)$ on $\mathcal{F}_n$. We observe that when $P(B|\cdot)$ is defined as the Radon–Nikodym derivative of $P(\cdot \cap B)$ with respect to $P(\cdot)$, then if $D \cap E = \emptyset$: (i) $P(D \cup E|C) = P(D|C) + P(E|C)$ [a.e. P] and (ii) $P(D|E) = 0$ [a.e. P]. Thus, for all $A \in \mathcal{B}$ and for almost all x,

$$\begin{aligned} P^n(A|h_n) &= P^n(A \cap h_n|h_n) + P^n(A \cap h_n^c|h_n) \quad \text{by (i)} \\ \text{and} \quad &= P^n(A \cap h_n|h_n) \quad \text{by (ii).} \end{aligned}$$

However, $A \cap h_n$ can be written as $(x_1, \ldots, x_n) \times E_n$, where $E_n \in \mathcal{F}_n$, as desired. This elementary result will be helpful in our discussion (in Section IV) of convergence to certainty.

III. CONSENSUS THROUGH MARTINGALES

Consider any probability Q which is in agreement with P about events of measure 0 in $\mathcal{B}$, i.e., $\forall E \in \mathcal{B}, P(E) = 0$ iff $Q(E) = 0$, so that P and Q are mutually absolutely continuous. Then Q, too, is predictive with conditional probability distributions $Q^n(\mathcal{F}_n|h_n)$. In their important paper of 1962, Blackwell and Dubins establish (almost sure) asymptotic consensus between the conditional probabilities P^n and Q^n. In particular, they show:

Theorem 1. *For each P^n there is a Q^n so that, almost surely, the distance between them vanishes with increasing histories*:

$$\lim_{n\to\infty} \rho(P^n, Q^n) \to 0 \quad [\text{a.e. } P \text{ or } Q],$$

where ρ is the uniform distance metric between distributions. That is, with μ and ν defined on the same measure space $(M, \mathcal{M})$, $\rho(\mu, \nu)$ is the l.u.b., *over events $E \in \mathcal{M}$, of $|\mu(E) - \nu(E)|$.*

(Blackwell and Dubins prove this result about consensus by a "slightly generalized" martingale convergence theorem – their Theorem 2 (1962, p. 883).)

What can be said about consensus when considering a set C of mutually absolutely continuous probabilities? Quite obviously, unless C is closed, there may be no consensus among conditional probabilities. In this vein, we note a simple corollary to Theorem 1.

Corollary 1. *Let C be a closed, convex set of probabilities all mutually absolutely continuous, and generated by finitely many of its extreme points. Denote this finite set by $C = \{P_1, \ldots, P_k\}$. Then, asymptotically, the conditional probabilities in C achieve consensus uniformly. That is, for almost all $x \in X, \forall \varepsilon > 0, \exists m, \forall n > m, \forall P, Q \in C, \rho(P^n, Q^n) < \varepsilon$.*

Proof. Because C is finite, by Theorem 1, for almost all $x \in X$,

$$\forall \varepsilon > 0, \exists m, \forall n > m, \max_{P_i, P_j \in C} \rho(P_i^n, P_j^n) < \varepsilon.$$

Recall that if $\rho(T, U) < \varepsilon$, then $\rho(T, S) < \varepsilon$ and $\rho(S, U) < \varepsilon$ for each convex combination $S = \alpha T + (1 - \alpha)U$, $0 \leq \alpha \leq 1$. Recall also that

$$\forall P \in C, \forall h_n, \exists \alpha_1, \ldots, \alpha_k \; (\alpha_i \geq 0, \textstyle\sum_i \alpha_i = 1)$$

$$P^n(\cdot|h_n) = \textstyle\sum_i \alpha_i P_i^n(\cdot|h_n), \quad i = 1, \ldots, k.$$

The corollary is immediate from these two observations. □

Note also that the corollary ensures consensus when agents are merely statistically coherent but take their updated probabilities from those in C.

Example 1. Let x_i be the i.i.d. Normal (μ, σ^2), where the conjugate priors for these parameters have hyperparameters which lie in a compact set. Given the observed history, h_n, let two agents choose probabilities P^n, Q^n (given h_n) in any manner (even as a function of h_n) from this set. Then, consensus obtains, almost surely, with observation of the x_i's.

The next result, which has a weaker conclusion, is true and helps to identify the role played by the topology in fixing closure of C.

Corollary 2. *Let C be a compact set (under the topology induced by ρ) of mutually absolutely continuous, predictive probabilities on the space $(X, \mathcal{B})$. Let $\{P_n, Q_n\}$ be any sequence of pairs from C. Then, $\forall R \in C$,*

$$\rho(P^n_n, Q^n_n) \xrightarrow{R} 0 \quad \text{as } n \to \infty.$$

That is,

$$\forall \varepsilon > 0, \lim_{n\to\infty} R(\{h_n : \rho(P^n_n, Q^n_n) > \varepsilon\}) = 0.$$

Proof. The result follows from the claim: $\rho(P^n_n, R^n) \xrightarrow{R} 0$ as $n \to \infty$. That is,

$$\forall \varepsilon > 0, \lim_{n\to\infty} R(\{h_n : \rho(P^n_n, R^n) > \varepsilon\}) = 0.$$

[The claim suffices for the corollary, since $\forall h_n$, $\rho(P^n_n, Q^n_n) \le \rho(P^n_n, R^n) + \rho(R^n, Q^n_n)$.] We demonstrate the claim using Blackwell–Dubins' theorem, Theorem 1, and a simple lemma.

Lemma. *For any predictive probabilities S and T,*

$$\text{if } \rho(S,T) < \alpha\beta, \text{ then } \forall n, S(\{h_n : \rho(S^n, T^n) > \alpha\}) < \beta.$$

The proof of the lemma is straightforward and is omitted.

Proof (of the claim). We argue indirectly. Suppose there is a subsequence, denoted by $\{n_i\}$, where $\exists \delta > 0$, $\forall n_i$, $R(\{h_{n_i}: \rho(P^{n_i}_{n_i}, R^{n_i}) > \varepsilon\}) > \delta$. C is compact. So it is sequentially compact and every sequence contains a convergent subsequence. Hence, $\{P_{n_i}\}$ contains a ρ-convergent subsequence, which we denote by $\{P_{m_j}\}$ where $\forall j, \exists i\ P_{m_j} = P_{n_i}$, and $\boldsymbol{P}$ is its uniform limit, which we denote by $P_{m_j} - \rho \to \boldsymbol{P} \in C$. Thus, $\forall k > 0$, $\exists i$, $\forall j > i$, $\rho(P_{m_j}, \boldsymbol{P}) < \varepsilon\delta/4k$. Then, by the lemma,

$$\exists i, \forall n, \forall j > i, \boldsymbol{P}\left(\left\{h_n : \rho\left(P^n_{m_j}, P^n\right) > \frac{1}{2}\right\}\right) < \delta/2k.$$

Since both $\boldsymbol{P}$ and R belong to the set C, by the Blackwell–Dubins theorem (going from "a. e." to "in probability" convergence),

$$\exists m, \forall n > m, \boldsymbol{P}\left(\left\{h_n : \rho(P^n, R^n) > \frac{1}{2}\varepsilon\right\}\right) < \delta/2k.$$

Let $m' = \max(m, m_i)$. Then, $\forall n > m'$ and $\forall m_j > m'$, $\boldsymbol{P}(\{h_n: \rho(P^n_{m_j}, R^n) > \varepsilon\}) < \delta/k$. Hence,

$$\forall \varepsilon > 0, \lim_{m'\to\infty} (\forall n > m' \, \forall m_j > m') \boldsymbol{P}\left(\left\{h_n : \rho\left(P^n_{m_j}, R^n\right) > \varepsilon\right\}\right) = 0.$$

That is, $\rho(P^n_{m_j}, R^n) \xrightarrow{P} 0$ as the pair $(m_j, n) \to \infty$. Because R is (finite and) absolutely continuous with respect to $\boldsymbol{P}$, *then* $\rho(P^n_{m_j}, R^n) \xrightarrow{R} 0$ as the pair $(m_j, n) \to \infty$. This contradicts the supposition, $\forall n_i$ $R(\{h_{n_i}: \rho(P^{n_i}_{n_j}, R^{n_i}) > \varepsilon\}) > \delta$, and proves the claim. □

However, Corollary 2 is not true when compactness of C is under the weak-star topology.[3] This is shown by a counterexample.

Example 2. Let $X_i = \{0, 1\}$ so that $(X, \mathcal{B})$ is the measurable space of (Borel sets) of infinite flips of a coin. Let R be the exchangeable probability on X given by the beta mixing prior $\alpha = \beta = 1$, the uniform distribution over the binomial parameter θ, for the deFinetti representation of R as a mixture of i.i.d. Binomial distributions.

Consider the set G_n of histories of length n for which the observed relative frequency of 1's is less than or equal to $\frac{1}{2}$. Let S_n be the exchangeable probability on X with beta mixing prior $\alpha = 6n$, $\beta = n$. Define probability P_n as follows:

$$P_n(\cdot) = \int \phi(\cdot | h_n) \mathrm{d}R(h_n),$$

where

$$\phi(\cdot | h_n) = \begin{cases} S^n_n(\cdot | h_n) & \text{if } h_n \in G_n, \\ R^n(\cdot | h_n) & \text{if } h_n \in G^c_n. \end{cases}$$

Then the sequence P_n converges weak-star to R, since each P_n agrees with R on all rectangles in $\mathcal{H}_n$ $(\supset \mathcal{H}_m, m \le n)$. (This follows by Halmos's Theorem 13.A, which we discuss in connection with Theorem 2, below.) Therefore, the set $W = \{R, P_n\ (n = 1, \ldots)\}$ is (weak-star) closed and every sequence in W contains a (weak-star) convergent subsequence.

So W is compact in the weak-star topology.[4] The elements of W are mutually absolutely continuous. (Use the Radon–Nikodym theory with the definition of P_n and the mutual absolute continuity of the S_n with R.) However, $\forall h_n \in G_n$, $\rho(P^n_n, R^n) > \frac{1}{4}$. This is because $\forall h_n \in G_n$, $P^n_n(x_{n+1} = 1 \mid h_n)$ $(= S^n_n(x_{n+1} = 1 \mid h_n)) \geq \frac{3}{4}$ while $R^n(x_{n+1} \mid h_n) \leq \frac{1}{2}$. Moreover, for each n, $R(G_n) \geq \frac{1}{2}$.

Note also that, given G_n, the difference $|P^n_n(\cdot) - R^n(\cdot)|$ fails to converge to 0, even pointwise, over events in $(X, \mathcal{B})$. Let E be the event that the limiting relative frequency of "1" exceeds $\frac{3}{4}$. Then,

$$\lim_{n\to\infty} \inf_{h_n \in G_n} |P^n_n(E) - R^n(E)| = \frac{1}{2}.$$

Last, observe that no superset of the P_n can be compact in the topology induced by ρ as $\rho(P_n, R) \geq \frac{1}{8}$.

The corollary can be strengthed to say:

Corollary 3. *Let the set C be as in Corollary 2. Then for all R in C,*

$$\forall \varepsilon > 0, \quad \lim_{n\to\infty} \sup_{P,Q \in C} R(\{h_n : \rho(P^n, Q^n) > \varepsilon\}) = 0.$$

Proof. We argue indirectly. Suppose $\exists \varepsilon > 0$ with

$$\liminf_{n\to\infty} \sup_{P,Q \in C} R(\{h_n : \rho(P^n, Q^n) > \varepsilon\}) = \delta > 0.$$

Equivalently, $\forall m$, $\exists n_m > m$, $\exists (P_m, Q_m) \in C$ such that

$$R(\{h_n : \rho(P^{n_m}_m, Q^{n_m}_m) > \varepsilon\}) = \delta.$$

Without loss of generality, assume $n_m > n_{m-1}$. Form two new sequences of pairs of elements of C, as follows. Let $A_n = P_m$ when $n = n_m$ and $A_n = R$ otherwise. Likewise, let $B_n = Q_m$ when $n = n_m$ and $B_n = R$ otherwise. Apply Corollary 2 to the sequence of pairs (A_n, B_n). We have, $\forall \varepsilon > 0$,

$$\lim_{n\to\infty} R(\{h_n : \rho(A^n_n, B^n_n) > \varepsilon\}) = 0,$$

which contradicts the supposition,

$$\forall m R(\{h_n : \rho(A^{n_m}_{n_m}, Q^{n_m}_{n_m}) > \varepsilon\}) = \delta. \quad \Box$$

Corollaries 2 and 3 guarantee convergence (in probability) for each sequence of pairs of probabilities chosen from the set C. These two results do *not* ensure the "in probability" convergence of conditional probabilities taken from C, analogous to the "almost everywhere" convergence of Corollary 1. The next example shows that compactness of C (under the uniform topology) is insufficient for "in probability" convergence of the conditional probabilities in C. Hence, trivially, ρ-compactness of C does not suffice for the "almost everywhere" convergence, which obtains according to the first corollary when C is generated by finitely many extreme points.

Example 3. Let $(X, \mathcal{B})$ and R be as in Example 2. Define the (integer-valued) function $n^*(m)$ $(m \geq 2)$, recursively, as follows:

$$n^*(2) = 1, n^*(m) = n^*(m-1) + \left\llbracket \frac{1}{2}(m+1) \right\rrbracket \quad \text{for } m = 3, \ldots,$$

where $\llbracket r \rrbracket$ is the integer part of r. For $k = 0, \ldots, \llbracket \frac{1}{2}m \rrbracket$, define

$$A_{k,m,n} = \left\{ h_n = (x_1, \ldots, x_n) : \left| (1/n)\left(\sum_{i=1}^{n} x_i \right) - k/m \right| \right.$$
$$< 1/2m + 1.001[\log\log n^*(m)/2n^*(m)]^{1/2};$$
$$\left. n = n^*(m), \ldots, n^*(m+1) - 1 \right\}$$

and define $m^*(n)$ to be that integer m such that $n^*(m) \leq n < n^*(m+1)$. Next, define the probability

$$Q_{k,m,n} = \int \phi(\cdot \mid h_n) \mathrm{d}R(h_n),$$

where

$$\phi(\cdot \mid h_n) = \begin{cases} S^n_n(\cdot \mid h_n) & \text{if } h_n \in A_{k,m,n}, \\ R^n(\cdot \mid h_n) & \text{if } h_n \notin A_{k,m,n}. \end{cases}$$

(This is analogous to the definition of P_n in Example 2). Since $R(A_{k,m,n}) = Q_{k,m,n}(A_{k,m,n}) \to 0$ uniformly in k and n as $m \to \infty$, any sequence of probabilities of the form $P_n = Q_{k,m,n}$, for $m = m^*(n)$ and $k \in \{0, \ldots, \llbracket \frac{1}{2}m \rrbracket\}$, will converge to R uniformly as $n \to \infty$.

By the law of the iterated logarithm for the binomial probability P_θ $(0 < \theta < 1)$,

$$P_\theta\left(\left\{h_n : \left|(1/n)\left(\sum_{i=1}^{n} x_i\right) - \theta\right| > \theta(1-\theta)(1.001)[2\log\log n/n]^{1/2}\right.\right.$$
$$\left.\left.\text{infinitely often in } n\right\}\right) = 0.$$

For all n and all $\theta < \frac{1}{2}$,

$$\left\{h_n : \left|(1/n)\left(\sum_{i=1}^{n} x_i\right) - \theta\right| \le \theta(1-\theta)(1.001)[2\log\log n/n]^{1/2}\right\}$$
$$\subset \left\{h_n : \left|(1/n)\left(\sum_{i=1}^{n} x_i\right) - \theta\right| \le 1.001[\log\log n^*(m)/2n^*(m)]^{1/2},\right.$$
$$\left.\text{for } m = m^*(n)\right\}$$
$$\subset \left\{h_n : \left|(1/n)\left(\sum_{i=1}^{n} x_i\right) - k/m\right| \le 1/2m + 1.001[\log\log n^*(m)/2n^*(m)]^{1/2},\right.$$
$$\left.\text{for } |k/m - \theta| \le 1/2m\right\}.$$

Now choose P_n as follows. Let $P_n = Q_{k,m,n}$ for $m = m^*(n)$ and $k = n - n^*(m) = k^*(n)$. From this it follows (by the law of the iterated logarithm) that, for all $\theta < \frac{1}{2}$, $P_\theta(\{A_{k^*(n),m^*(n),n}$ infinitely often in $n\}) = 1$. Then, we have that $R(\{A_{k^*(n),m^*(n),n}$ infinitely often in $n\}) \ge \frac{1}{2}$. Let E be the event that the limiting frequency of "1" is at most 0.65. For $n \ge 600$, and for all histories $h_n \in A_{k^*(n),m^*(n),n}$, $|P_n^n(E) - R^n(E)| \ge 0.9$. Hence,

$$\{h_n : A_{k^*(n)m^*(n),n} \text{ infinitely often}\} \subset \{h_n : \rho(P_n^n, R^n) \nrightarrow 0\},$$

and it follows that $R(\{h_n : \rho(P_n^n, R^n) \to 0\}) \le \frac{1}{2}$.

Next, select a sequence of probabilities, $\{P_{n,h_n}\}$, where the n-th term in the sequence depends on the history h_n, as follows. If $(1/\mathrm{n})(\Sigma_{i=1}^{n} x_i) \le \frac{1}{2}$, let $P_{n,h_n} = Q_{k,,m^*(n),n}$, where k is chosen so that $|(1/\mathrm{n})\Sigma_{i=1}^{n} x_i) - k/m^*(n)| \le 1/2m^*(n)$. Otherwise, if $(1/\mathrm{n})(\Sigma_{i=1}^{n} x_i) > \frac{1}{2}$, let $P_{n,h_n} = R$. The set C of R together with all the P_{n,h_n} (the $Q_{k,m^*(n),n}$ for $k = 0, \ldots, [\![\frac{1}{2}m^*(n)]\!]$) is compact (in the uniform topology), since every subsequence converges uniformly to R, and the elements of C are mutually absolutely continuous. But, for $n \ge 600$, $R(\{h_n : \rho(P_{n,h_n}^n, R^n) > 0.9\}) \ge \frac{1}{2}$. Thus,

$$\liminf_{n\to\infty} R\left(\left\{h_n : \sup_{P \in C} \rho(P^n, R^n) > 0.9\right\}\right) \ne 0.$$

Therefore,

$$\sup_{P,Q \in C} \rho(P^n, Q^n) \nrightarrow 0 \quad \text{(in probability)}, \qquad (*)$$

in marked contrast with Corollary 3.

Last, note that the set C' consisting just of R and the P_n is ρ-compact and contains mutually absolutely continuous elements. It is easy to see (in effect by Corollary 3) that

$$\sup_{P,Q \in C'} \rho(P^n, Q^n) \xrightarrow{R} 0. \qquad (**)$$

Thus, "in probability" consensus of the form (**) does not entail "almost everywhere" consensus of the form $\rho(P_n^n, R^n) \to 0$ [a.e. R].

IV. THE ALMOST CERTAIN APPROACH TO CERTAINTY FOR EVENTS MEASURABLE WITH RESPECT TO $(X, \mathcal{B})$

Though Blackwell and Dubins do not make mention of it, another consequence of Doob's martingale convergence theorem (Doob, 1953, Theorem 7.4.1, p. 319) is the desired result about the approach to certainty in these conditional probabilities. Denote the characteristic function of a set E by $\chi_E(x) = 1$ if $x \in E$ and $\chi_E(x) = 0$ if $x \notin E$, for $x \in X$ a complete history.

Theorem 2. $\forall E \in \mathcal{B}$, $\lim_{n\to\infty} P^n(E|h_n) = \chi_E(x)$ [a.e. P].

Theorem 2 asserts that, for each event E and for all but a set of complete histories of P-measure 0 (depending upon E), the sequence of conditional probabilities, $P^n(E|h_n)$, converges to 1 or to 0 as E occurs or not.

Proof. Theorem 2 is a substitution instance of Doob's Theorem 7.4.3 (p. 331) which reads (in relevant parts) as follows: Let z be a random variable with finite absolute expectation, $E\{|z|\} < \infty$ and let $\mathcal{G}_1 \subset \mathcal{G}_2 \subset \ldots$ be Borel (σ-)fields of measurable sets. Let $\mathcal{G}_\infty$ be the smallest Borel (σ-)field of sets with $\cup_n \mathcal{G}_n \subset \mathcal{G}_\infty$. Then, $\lim_{n\to\infty} E\{z|\mathcal{G}_n\} = E|\{Z|\mathcal{G}_\infty\}$ [a.e.]. Theorem 2 obtains by taking $z = \chi_E$, $\mathcal{G}_n = \mathcal{H}_n$, and $\mathcal{G}_\infty = \mathcal{B}$. □

Doob's Theorem 7.4.3 is proven with Theorem 7.4.1 (his martingale convergence theorem) and a familiar measure-theoretic result (in effect, Theorem 13.A of Halmos, 1950, p. 54), which asserts that a σ-finite measure μ on an algebra $\mathcal{A}$ induces a unique σ-finite extension of μ on the smallest σ-algebra containing $\mathcal{A}$, $\sigma[\mathcal{A}]$.[5] It is our purpose, next, to show that Theorem 2 can be derived from a measure-theoretic result (Halmos, Theorem 13.D, p. 56), related to the extension Theorem 13.A, without appeal to martingale theory. But first we state without proof a simple consequence of Theorem 2 and Corollary 1.

Corollary 4. *Let the set C be as in Corollary 1. If $f(h_n) \in \{P^n(E|h_n): P \in C\}$, then the sequence of conditional probabilities, $f(h_n)$, converges to $\chi_E(x)$* [a.e. $P \in C$}.

Thus, when the set C is as in Corollary 1, static coherence suffices for asymptotic certainty.

Besides the Theorem 13.A, which ensures the unique extension of the measure μ from $\mathcal{A}$ to $\sigma[\mathcal{A}]$, also there is an important result about approximation (in measure) of elements of $\sigma[\mathcal{A}]$ by elements of $\mathcal{A}$. This is reported by Halmos (1950, p. 56).

Theorem 13.D. *If μ is a σ-finite measure on a ring R, then for every set E of finite measure in $\sigma[R]$ and for every $\varepsilon > 0$, there is a set $E_0 \in R$ with $\mu(E \triangle E_0) \leq \varepsilon$ ($\triangle$ is symmetric difference).*

Theorem 2 can be derived from this directly (as noted by Halmos, Theorem 49.B, p. 213), without supposing P is predictive.[6]

The technique used in the alternative proof establishes the approach to certainty for the extension of $(X, \mathcal{B}, P)$ to its measure completion.

Corollary 5. *Let $(X, \overline{\mathcal{B}}, \overline{P})$ be the measure completion of $(X, \mathcal{B}, P)$. Then*

$$\forall E \in \overline{\mathcal{B}}, \lim_{n\to\infty} \overline{P}^n(E|h_n) = \chi_E(x) \quad [\text{a.e.}\ \overline{P}].$$

Also, Theorem 2 provides an elementary proof of the Lebesgue density theorem (see Lebesgue, 1904, and Oxtoby, 1971, pp. 16–18), which we give as a corollary. Let μ be Lebesgue measure. For each measurable set E on the real line $\mathcal{R}$, define the density at the point x by:

$$\lim_{h\to\infty} \mu(E \cap [x-h, x+h])/2h,$$

and denote by $\phi(E)$ the set of points at which E has density 1.

Corollary 6. The one-dimensional Lebesgue density theorem: *For each measurable set $E \subset \mathcal{R}$; $\mu(E \triangle \phi(E)) = 0$.*

Proof. Use Corollary 5 and the fact that the unit interval [0, 1] is Borel equivalent to 2^ω. (See Royden, 1968, p. 268.)

V. CONCLUSIONS

Theorem 2 asserts that, with increasing evidence, conditional probabilities for an event E approach certainty, almost surely. We alert the reader to the requirement that $E \in \mathcal{B}$, so that asymptotic certainty about a parameter μ depends upon the measurability of μ.

Example 4a. Let x_1 be binary, $X_1 = \{0, 1\}$, with probability $P(x_1 = 1) = \mu$. Consider the infinite sequence $X = (x_1, x_1, x_1, \ldots)$ generated by repeating the outcome x_1. This is (trivially) an exchangeable sequence. By deFinetti's representation theorem, the probability P on X is given as a mixture of i.i.d. Binomial distributions with some "prior" (mixing) probability distribution $\pi(\theta)$ over the binomial parameter θ. However, $\mu \neq \theta$. On the contrary, P is represented by the "prior" mixture $\pi(\theta = 0) = (1 - \mu)$, $\pi(\theta = 1) = \mu$ and, obviously, observations tells us about θ only, not about the parameter $\mu (= \pi)$ which is not measurable in the σ-field $(X, \mathcal{B})$.

Example 4b. A less obvious version of this problem is as follows. Let x_i form an exchangeable sequence where, for each integer $k, x_{i_1}, \ldots, x_{i_k}$ have the Multivariate Normal distribution $N_k = (\mu, \Sigma)$, with $\mu' = [\mu, \ldots, \mu]$ and with Σ equal to the $k \times k$ matrix having main diagonal elements all 2's and off-diagonal entries all 1's. Then μ fails to be measurable with respect to the σ-field $(X, \mathcal{B})$. Repeated observations of the x_i do not make the posterior probability concentrate about μ. Rather, the asymptotic posterior distribution of μ from these data, with limiting sample average = $\hat{x}$, is identical to the posterior one would obtain (using the same prior over μ) from a sample of two, independent Normal(μ, 2) observations with sample average $\hat{x}$.

Against the background of Theorem 2, we report conditions which guarantee asymptotic consensus (under the uniform distance metric) for conditional probabilities taken from a set C of unconditional distributions. Not surprisingly, depending upon how C is closed, different conclusions obtain.

When C is (contained within or) generated by finitely many (mutually absolutely continuous) elements, i.e., when C is convex and its extreme points form a compact set in the discrete topology, consensus of conditional probabilities occurs, almost surely. When the extreme points of C form a compact set in the uniform topology, there is "in probability" (but not "almost sure") consensus for sequences of pairs of conditional probabilities. When C is compact in the weak-star topology, there may be no limiting consensus. In this case, there can even be an event E and a sequence of pairs (P_n, Q_n) from C where the (paired) conditional probabilities of E differ by a fixed amount, $|P^n_n(E) - Q^n_n(E)| > \delta > 0$.

Of course, these "large sample" results fail to provide bounds on the rates with which consensus and certainty occur. What they do show, however, is the surprising fact that these asymptotic properties of conditional probabilities do *not* depend upon exchangeability or other kinds of symmetries of the (unconditional) probabilities in C. Rather, agreement on events of zero probability and a suitable closure suffice for consensus, while "the approach to certainty" is automatic.

NOTES

1 A careful discussion of static and dynamic coherence is given by Levi (1980, Chapter 4). Simply stated, static coherence requires that an agent have (conditional) degrees of beliefs captured by a (conditional) probability which respects the total evidence principle. Let $P_k(\cdot|\cdot)$ be one such (static) representation, with background knowledge K. If the agent learns some new evidence E, so that K' (the closure under implication of K with E) is the new knowledge, then temporal conditionalization requires identifying the updated (static) representation $P_{K'}(\cdot|\cdot)$ with the conditional probability $P_k(\cdot|\cdot, E)$. Hence, temporal conditionalization provides a dynamic constraint.

Savage's analysis (§3.6 of 1954) is ambiguous between the static and dynamic reading of the convergence in conditional probabilities. The ambiguity persists even in his subsequent essay, "Implications of Personal Probability for Induction" (1967), though we think he appears inclined there to endorse temporal conditionalization. M. Goldstein concentrates on this distinction in his "Exchangeable Belief Structures" (1986) and other essays of his cited therein. Besides Levi's clear presentation, different philosophic perspectives on the question of static vs. dynamic coherence include: Kyburg's "Conditionalization" (1980), van Fraassen's "Belief and the Will" (1984), and additional references given therein.

2 Predictive probabilities are those which admit *regular conditional distributions* in the sense of Breiman (1968, p. 77). Without regularity of conditional distributions, all that can be shown about the existence of conditional probabilities follows from the Radon–Nikodym theorem. In Section IV we point out that, for event E, the convergence to certainty of the conditional probabilities, $P^n(E|h_n)$, does not require that P be predictive. We alert the reader to minor variations in the definitions of "predictive", as found in Breiman (1968, p. 77), Doob (1953, p. 26), and Halmos (1950, pp. 209–210). See, in particular, Doob's discussion (1953, p. 624).

Last, we note that the sufficient condition for P being predictive, as given by Breiman (1968, Theorem 4.34, p. 79), is also sufficient for extending a set function μ given "marginally" as a probability on n-dimensional sets of a ring R (that is, for each n, μ is given as a probability on n-dimensional rectangles), to a coherent probability on $\sigma[R]$, its infinite dimensional σ-ring, as shown by Halmos (1950, T.A, p. 212).

3 Diaconis and Freedman (1986, appendix) discuss "weak-star merging" of posterior probabilities in a setting with i.i.d. data. Their interesting Theorem A.1 (1986, p. 18) equates such "merging" of opinions with their rather strict notion of "consistency" of posterior probabilities.

4 Theorem I.6.13, p. 21 of Dunford and Schwartz (1958) asserts that, in a metric space, a closed and sequentially compact set is compact. For discussion of the metrizability of the weak-star topology see Dudley (1968, §1). Those results on the metrizability of the weak-star topology apply to our probabilities on $(X, \mathcal{B})$ since the unit interval $[0, 1]$, a separable metric space, is Borel equivalent to 2^ω (Royden, 1968, p. 268).

5 This note offers some details about how Doob's Theorem 7.4.3 is demonstrated in order to emphasize the role of the extension Theorem 13.A. Define y_n by $y_n = E\{\chi_E(x)|h_n\}$. Then the random variables $y_1, y_2, \dots$ constitute a martingale. That is, as required for a martingale: (a) $E\{|y_n|\}$ $(= P(E)) < \infty$, and (b) $E\{y_{n+1}|y_1, \dots, y_n\} = E[y_{n+1}|y_n\} = y_n$.

The latter is a special case of Doob's Example 1 (1953, p. 92), with $\eta = \chi_E$, and $\xi_n = x_n$, since $h_n = (x_1, \dots, x_n)$.

By Doob's Theorem 7.4.1(i), $\lim_{n\to\infty} y_n = w$ exists (almost surely). But by 7.4.1(ii), w behaves exactly like the conditional probability $P(E|x)$ (which is the Radon–Nikodym integrand representation of $P(E \cap \cdot)$ over $\mathcal{B}$) for all elements of the field $\cup_n \mathcal{H}_n$. That is, $\forall n, \forall A \in \mathcal{H}_n$,

$$\int_A w\mathrm{d}P = \int_A y_n \mathrm{d}P(h_n) = \int_A P(E|x)\mathrm{d}P = P(E \cap A).$$

By the extension Theorem 13.A, a measure on $\mathcal{B}$ is uniquely determined by its values on the field $\cup_n \mathcal{H}_n$. Therefore, $w = P(E|x) = \chi_E(x)$, almost surely, as was to be shown.

6 Halmos proves Theorem 49.B by showing "almost uniform convergence". Actually, his argument stops short of that. The proof is completed by a familiar construction relating to Egoroff's theorem, e.g., Exercise 30 of Royden (1968, p. 72).

REFERENCES

Blackwell, D. and L. Dubins (1962). Merging of opinions with increasing information. *Ann. Math. Statist.* **33,** 882–887.

Breiman, L. (1968). *Probability*. Addison-Wesley, Reading, MA.

Diaconis, P. and D. Freedman (1986). On the consistency of Bayes estimates. *Ann. Statist.* **14**, 1–26.

Doob, J.L. (1953). *Stochastic Processes.* Wiley, New York.

Dudley, R.M. (1968). Distances of probability measures and random variables. *Ann. Math. Statist.* **39**, 1563–1572.

Dunford, N. and J.T. Schwartz (1958). *Linear Operators*, Part I. Interscience, New York.

van Fraassen, B.C. (1984). Belief and the will, *J. of Philos*. **81**, 235–256.

Goldstein, M. (1986). Exchangeable belief structures. *J. Amer. Statist. Assoc.* **81** (396), 971–976.

Halmos, P.R. (1950). *Measure Theory*. Van Nostrand, New York.

Kyburg, H.E.K. (1980). Conditionalization. *J. of Philos*. **77**, 98–114.

Lebesgue, H. (1904). *Leçons sur l'Intégration et la Recherche des Fonctions Primitives.* Paris.

Levi, I. (1980). *The Enterprise of Knowledge.* MIT Press, Cambridge, MA.

Oxtoby, J.C. (1971). *Measure and Category.* Springer-Verlag, New York.

Royden, H.L. (1968). *Real Analysis.* Macmillan, New York.

Savage, L.J. (1954). *The Foundations of Statistics.* Wiley, New York.

Savage, L.J. (1967). Implications of personal probability for induction. *J. of Philos.* **64**, 593–607.

3.7

Reasoning to a Foregone Conclusion

JOSEPH B. KADANE, MARK J. SCHERVISH, AND TEDDY SEIDENFELD

ABSTRACT

When can a Bayesian select an hypothesis *H* and design an experiment (or a sequence of experiments) to make certain that, given the experimental outcome(s), the posterior probability of *H* will be greater than its prior probability? In this chapter we discuss an elementary result that establishes sufficient conditions under which this reasoning to a foregone conclusion cannot occur. We illustrate how when the sufficient conditions fail, because probability is finitely but not countably additive, it may be that a Bayesian can design an experiment to lead his/her posterior probability into a foregone conclusion. The problem has a decision theoretic version in which a Bayesian might rationally pay not to see the outcome of certain cost-free experiments, which we discuss from several perspectives. Also, we relate this issue in Bayesian hypothesis testing to various concerns about "optional stopping."

I. INTRODUCTION

In a lively (1962) discussion of some foundational issues, several noted statisticians, especially L. J. Savage, focused on the controversy of whether an experimenter's stopping rule is relevant to the analysis of his or her experimental data. Savage wrote (1962, p. 18):

The [likelihood] principle has important implications in connection with optional stopping. Suppose the experimenter admitted that he had seen 6 red-

This research was supported by National Science Foundation Grants SES-9123370, DMS-9005858, DMS-9302557, and SES-9208942 and by Office of Naval Research Contracts N00014-89-J-1851 and N00014-91-J-1024.

Reprinted with permission from the *Journal of the American Statistical Association*, 91, no. 435 (September 1996): 1228–1236.