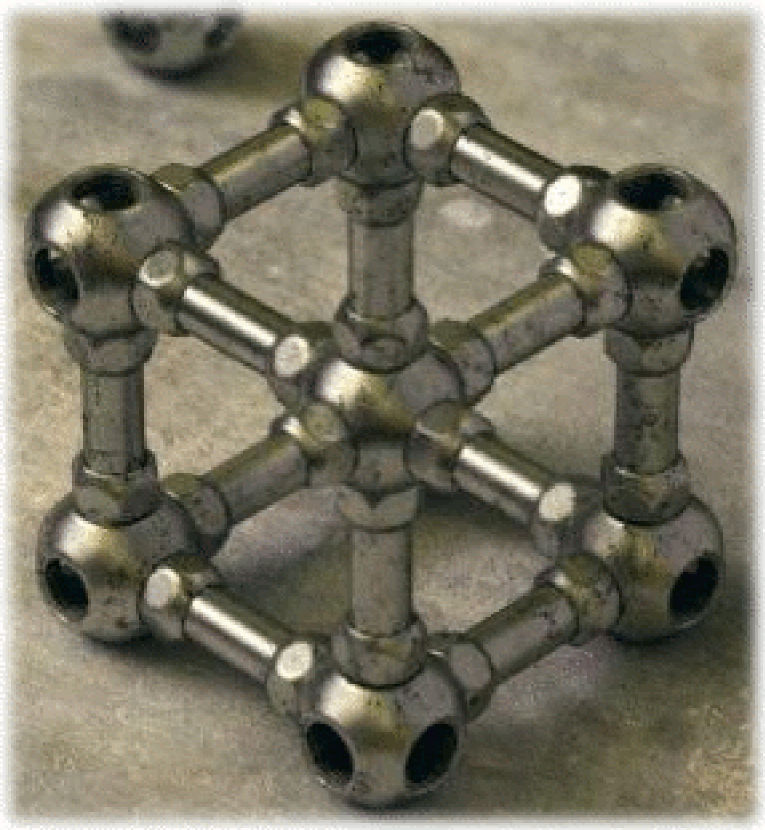

Concentration inequalities and laws of large numbers under epistemic and regular irrelevance [1]

Author(s):

Fabio Gagliardi Cozman

[1]This work was supported by Fapesp Project LogProb, grant 2008/03995-5, São Paulo, Brazil.


Volume 51, Issue 9, November 2010
ISSN 0888-613X
Special Issue: Imprecise probability in statistical inference and decision making
Edited by: Thomas Augustin, Frank Coolen, Serafin Moral and Matthias Troffaes
ELSEVIER
INTERNATIONAL JOURNAL OF
APPROXIMATE
REASONING
Imprecision and Uncertainty in Intelligent Systems
Editor-in-Chief
Thierry Denoeux
Available online at www.sciencedirect.com
ScienceDirect

International Journal of Approximate Reasoning 51 (2010) 1069–1084

ELSEVIER

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

Concentration inequalities and laws of large numbers under epistemic and regular irrelevance ☆

Fabio Gagliardi Cozman

Escola Politécnica, Universidade de São Paulo, São Paulo, SP, Brazil

A R T I C L E I N F O

Article history:
Available online 22 August 2010

Keywords:
Laws of large numbers
Lower and upper expectations and probabilities
Irrelevance and independence

A B S T R A C T

This paper presents concentration inequalities and laws of large numbers under weak assumptions of irrelevance that are expressed using lower and upper expectations. The results build upon De Cooman and Miranda's recent inequalities and laws of large numbers. The proofs indicate connections between the theory of martingales and concepts of epistemic and regular irrelevance.

1. Introduction

In this paper, we assume that a decision maker represents her uncertainty about a situation of interest through a *set* of expectation functionals. As each expectation functional induces a probability measure, our decision maker operates with a *set* of probability measures K instead of a *single* probability measure. There may be more than a single measure in K either because there are disagreements about the situation of interest, or because the decision maker is verifying the robustness of her assessments against perturbations, or because the decision maker has neither time nor resources to eliminate distributions from K. Perhaps the decision maker even wishes to abstract tedious details of the situation by not specifying point probabilities for some events. In any case, for each variable X we have its *lower* and *upper* expectations, respectively

$$\underline{E}[X] \doteq \inf E[X], \quad \overline{E}[X] \doteq \sup E[X],$$

where inf and sup are taken with respect to the set of expectation functionals. Similarly, for any event A, we have its *lower* and *upper* probabilities, respectively

$$\underline{P}(A) \doteq \inf P(A), \quad \overline{P}(A) \doteq \sup P(A),$$

where $P(A)$ is equal to the expectation of I_A, the indicator function of A.

The goal of this paper is to present concentration inequalities and laws of large numbers under weak assumptions of "irrelevance" that are appropriate for such a decision maker. To illustrate the kind of result we seek, consider that De Cooman and Miranda [4, Def. 1] have recently identified an assumption of irrelevance based on lower and upper expectations, called *forward factorization*, that leads to laws of large numbers such as:

$$\text{for any } \epsilon > 0, \quad \lim_{n\to\infty} \underline{P}\left(\frac{\sum_{i=1}^n \underline{E}[X_i]}{n} - \epsilon \leqslant \frac{\sum_{i=1}^n X_i}{n} \leqslant \frac{\sum_{i=1}^n \overline{E}[X_i]}{n} + \epsilon \right) = 1.$$

Note that weaker assumptions (basic model is a *set* of expectation functionals) lead to weaker conclusions (average stays within *interval*). Inequalities and laws presented later are similar to these previous seminal results.

☆ This work has been financially supported by FAPESP, grant 2008/03995-5. The author is partially supported by CNPq.
E-mail address: fgcozman@usp.br

0888-613X/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2010.08.009

Section 2 presents some basic concepts. Section 3 considers several assumptions of irrelevance for sets of variables. Section 4 presents results for *bounded* variables. Regarding bounded variables the contribution here, when compared to De Cooman and Miranda's work, lies in offering tighter inequalities and alternative proof techniques that are closely related to established methods in standard probability theory (in particular, close to the Hoeffding and the Azuma inequalities). Section 5 offers more significant contributions as we lift the assumption of boundedness for variables, and use martingale theory to prove laws of large numbers under countable additivity. Section 6 explains the validity of results in Section 4 for full conditional measures and for Walley's theory of lower previsions. Section 7 comments on the significance of results.

2. Sets of expectations and probabilities, conditioning and irrelevance

Throughout the paper, we assume that an expectation functional E maps variables into real numbers, and satisfies:

(1) if $\alpha \leqslant X \leqslant \beta$ then $\alpha \leqslant E[X] \leqslant \beta$;
(2) $E[X+Y] = E[X] + E[Y]$;

where X, Y are variables and α, β are real numbers (inequalities are understood pointwise). From such an expectation functional E, a *finitely additive* probability measure P is induced by $P(A) \doteq E[I_A]$ for any event A with indicator function I_A (an event is a subset of the possibility space Ω). We sometimes denote the indicator function of event A simply by A. A finitely additive probability measure defined on the field of all subsets of Ω completely characterizes an expectation functional on bounded functions and vice versa [31, Thm. 3.2.2]. An important property of expectation functionals is that if a sequence of bounded variables $X_1, X_2, \ldots$ is such that $\lim_{j\to\infty} \sup|X_j - X| = 0$ for some variable X, then [31, Sec. 2.6.1(l)]

$$\lim_{j\to\infty} E[X_j] = E[X]. \tag{1}$$

A set of probability measures induced by a set of expectation functionals is called a *credal set* [22]. We do not assume that a credal set must be convex, nor closed, nor connected; an axiomatization of such general credal sets from preferences has been proposed by Seidenfeld et al. [28]. Given a credal set K, lower and upper expectations can be written, respectively, as $\underline{E}[X] \doteq \inf_{P\in K} E_P[X]$ and $\overline{E}[X] \doteq \sup_{P\in K} E_P[X]$. Lower and upper probabilities are similarly written as $\underline{P}(A) \doteq \inf_{P\in K} P(A)$ and $\overline{P}(A) \doteq \sup_{P\in K} P(A)$.

2.1. Countable additivity

Countable additivity is an assumption of continuity; for expectation functionals it reads [34, Sec. 2.2]: if $X_1, X_2, \ldots$ increase monotonically to a limit X, then $E[X] = \lim_i E[X_i]$. For a probability measure, countable additivity means: if $A_1 \supset A_2 \supset \cdots$ is a countable sequence of events such that $\cap_i A_i = \emptyset$, then $\lim_{n\to\infty} P(A_n) = 0$. For a credal set, countable additivity means that given a countable sequence of events

$$A_1 \supset A_2 \supset \cdots \quad \text{such that} \quad \cap_i A_i = \emptyset, \quad \text{then} \quad \lim_{n\to\infty} \overline{P}(A_n) = 0 \tag{2}$$

(hence, $\lim_{n\to\infty} P(A_n) = 0$ for every probability measure in the credal set; that is, every probability measure in the credal set satisfies countable additivity).

Countable additivity is assumed in the remainder of this section and in Sections 3–5. Whenever countable additivity is assumed, we assume that variables are measurable and all measures in the credal set of interest are specified using the same σ-field (so that $\sup_P E_P[X|Y]$ is measurable). Countable additivity is *not* assumed in Section 6.

2.2. Conditioning

The conditional expectation for variable X given a nonempty event A, denoted by $E[X|A]$, is constrained by $E[X|A]P(A) = E[XA]$. The "standard" approach to conditioning is to define $E[X|A]$ as $E[XA]/P(A)$ when $P(A) > 0$, and to leave $E[X|A]$ undefined when $P(A) = 0$. If we have two random variables X and Y, the standard (Kolmogorovian) approach to conditioning takes $E[X|Y]$ to be a random variable that solves the following equation for every B in the σ-algebra generated by Y [24, Sec. B.1.2]:

$$E[B(X - E[X|Y])] = 0. \tag{3}$$

The Radon–Nikodym theorem guarantees, given the assumption of countable additivity in the standard theory, existence of $E[X|Y]$, unique up to probability zero changes. Moreover, the following *disintegrability* result holds: $E[X] = E[E[X|Y]]$.

To motivate some of the definitions proposed in the next section, consider the definition of conditional upper expectations when we have a *set* K of expectation functionals. It might seem reasonable to define conditional upper expectations as follows:

$$\overline{E}^{\oslash}[X|A] \doteq \sup_{E\in K} E[X|A] \quad \text{if } \underline{P}(A) > 0 \qquad (4)$$
$$\overline{E}^{\oslash}[X|A] \text{ undefined} \quad \text{if } \underline{P}(A) = 0,$$

and likewise for conditional lower expectations; that is, $\underline{E}^{\oslash}[X|A]$ is equal to $\inf E[X|A]$ if $\underline{P}(A) > 0$ and undefined otherwise. This sort of conditioning appears in theories that ignore events of lower probability zero, such as Giron and Rios' theory [15]; later we indicate that this definition does not seem to lead to interesting laws of large numbers. A possibly more sensible idea, that we indicate through the superscript >, is to discard those distributions for which $P(A) = 0$ [32,33]:

$$\overline{E}^{>}[X|A] \doteq \sup_{E\in K:P(A)>0} E[X|A] \quad \text{if } \overline{P}(A) > 0, \qquad (5)$$
$$\overline{E}^{>}[X|A] \text{ undefined} \quad \text{if } \overline{P}(A) = 0,$$

and likewise for conditional lower expectations; that is, $\underline{E}^{>}[X|A]$ is equal to $\inf_{E\in K:P(A)>0} E[X|A]$ if $\overline{P}(A) > 0$ and undefined otherwise. We refer to this strategy as *regular conditioning*, inspired by Walley [31, Ap. J], who uses the term *regular extension* for a similar idea. Appendix A further comments on regular conditioning.

2.3. Irrelevance

Suppose we have a set of probability measures and two variables X and Y. Walley defines *epistemic irrelevance* of Y to X to mean that

$$\overline{E}[f(X)|Y] = \overline{E}[f(X)]$$

for all bounded functions f of X (Section 6 further comments on Walley's theory). One might take epistemic irrelevance as a relaxed version of stochastic independence, perhaps suitable for robustness analysis, or as the appropriate definition of irrelevance in the presence of disagreeing, incomplete or imprecise assessments of probability. Note that epistemic irrelevance is *much weaker* than requiring that each expectation functional satisfies standard stochastic independence of X and Y.

Because Walley's concept requires a theory of conditioning that departs from the standard one (Section 6), we present here a modified concept of irrelevance that employs the intuition behind regular conditioning. Assume countable additivity and suppose all measures of interest are specified using the same σ-field; further assume that for each expectation functional E, the conditional expectation $E[\cdot|Y]$ is a random variable obtained through the standard approach to conditioning.

In rough terms, our approach is to associate with each probability measure P in the credal set an event N_P such that $P(N_P) = 0$, and to require that for all functions f of X,

$$\underline{E}[f(X)] \leqslant E_P[f(X)|Y=y] \leqslant \overline{E}[f(X)] \quad \text{for all } y \notin N_P. \qquad (6)$$

We start with some preliminary definitions. Given a credal set K, an *exclusion set* $\mathbf{N}$ is a set containing an event N_P for each probability measure P in K, such that $P(N_P) = 0$. Define the random variable $\overline{E}^{>}_{\mathbf{N}}[X|Y]$ as follows:

$$\overline{E}^{>}_{\mathbf{N}}[X|Y=y] \doteq \begin{cases} \sup_{P:y\notin N_P} E_P[X|Y=y] & \text{when } \{P : y \notin N_P\} \neq \emptyset, \\ 0 & \text{otherwise}. \end{cases}$$

Define also $\underline{E}^{>}_{\mathbf{N}}[X|Y=y]; \doteq -\overline{E}^{>}_{\mathbf{N}}[-X|Y=y]$; that is:

$$\underline{E}^{>}_{\mathbf{N}}[X|Y=y] \doteq \begin{cases} \inf_{P:y\notin N_P} E_P[X|Y=y] & \text{when } \{P : y \notin N_P\} \neq \emptyset, \\ 0 & \text{otherwise}. \end{cases}$$

Write

$$\overline{E}^{>}_{\mathbf{N}}[X|Y] \cong \alpha$$

to indicate that $\overline{E}^{>}_{\mathbf{N}}[X|Y=y] = \alpha$ for those y such that $\{P : y \notin N_P\} \neq \emptyset$. Likewise, write

$$\overline{E}^{>}_{\mathbf{N}}[X|Y] \lessapprox \alpha \quad \text{and} \quad \underline{E}^{>}_{\mathbf{N}}[X|Y] \gtrapprox \alpha$$

to indicate inequalities that hold for y such that $\{P : y \notin N_P\} \neq \emptyset$. Finally:

Definition 1. *Regular irrelevance* of Y to X obtains when

$$\overline{E}^{>}_{\mathbf{N}}[f(X)|Y] \cong \overline{E}[f(X)] \qquad (7)$$

for every function f of X and for some exclusion set $\mathbf{N}$.

Under countable additivity and standard conditioning (expression (3)), regular irrelevance of Y to X implies

$$\overline{E}^{>}[f(X)|A(Y)] = \overline{E}[f(X)] \qquad (8)$$

for any function f of X and for any event $A(Y)$ defined by variable Y such that $\overline{P}(A(Y)) > 0$. Throughout the paper the expression "event $A(Y)$ defined by variable Y" means that A has an indicator function that is a zero/one function of Y.

3. Irrelevance assumptions and factorizations

We now introduce the main irrelevance assumptions for sets of random variables. To simplify the notation, a set of variables $\{X_1, \ldots, X_n\}$ is denoted by $X_{1:n}$. Later we refer to infinitely long sequences of variables $X_1, X_2, \ldots$; all concepts of irrelevance apply to an infinite sequence if they apply to every subsequence $X_{1:n}$.

3.1. Forward regular irrelevance and weak forward regular irrelevance

Our starting point is De Cooman and Miranda's assumption of *forward irrelevance* [4,5] for random variables $X_{1:n}$:

- for each $i \in [2, n]$, for any function f of X_i,

$$\overline{E}[f(X_i)|X_{1:i-1}] = \overline{E}[f(X_i)].$$

We adapt their assumption to the definition of regular irrelevance, and define *forward regular irrelevance* as follows:

- for each $i \in [2, n]$, there is an exclusion set $\mathbf{N}$ such that for any function f of X_i,

$$\overline{E}^{>}_{\mathbf{N}}[f(X_i)|X_{1:i-1}] \cong \overline{E}[f(X_i)]. \tag{9}$$

A weaker condition, that we refer to as *weak forward regular irrelevance*, follows the intuition behind expression (6):

- for each $i \in [2, n]$, there is an exclusion set $\mathbf{N}$ such that

$$\overline{E}^{>}_{\mathbf{N}}[X_i|X_{1:i-1}] \lessapprox \overline{E}[X_i] \quad \text{and} \quad \underline{E}^{>}_{\mathbf{N}}[X_i|X_{1:i-1}] \gtrapprox \underline{E}[X_i]. \tag{10}$$

Another variant of forward irrelevance, now based on the intuition behind expression (4), is:

- for each $i \in [2, n]$, for any function f of X_i,

$$\overline{E}[f(X_i)|X_{1:i-1} = x_{1:i-1}] = \overline{E}[f(X_i)] \quad \text{whenever } \underline{P}(X_{1:i-1} = x_{1:i-1}) > 0. \tag{11}$$

This latter condition is really too weak to produce any sensible law of large numbers, as the following example demonstrates.[1] For this reason, we do not deal with condition (11) further in this paper.

Example 1. Consider binary variables $X_1, X_2, \ldots$ (values 0 and 1). Define events $A_0 \doteq \cap_{i \geqslant 1}\{X_i = 0\}$ and $A_1 \doteq \cap_{i \geqslant 1}\{X_i = 1\}$. Consider a convex and closed set K of joint distributions built as the convex hull of three distributions P_1, P_2 and P_3, as follows.

Distribution P_1 simply assigns probability one to A_1. Distribution P_2 assigns probability δ to A_0 and probability $1 - \delta$ to A_1, for some $\delta \in (0, 1)$. Distribution P_3 is the product of identical marginals: for any integer $n > 0$, $P_3(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^n P_3(X_i = x_i)$, where $P_3(X_i = 1) = 1 - \delta$.

For the convex hull of P_1, P_2 and P_3, expression (11) is satisfied. This conclusion is reached by analyzing each distribution in turn. Note that lower and upper expectations for any function of a binary variable X are linearly related to lower and upper probabilities of the event $\{X = 1\}$; consequently, in this example it is enough to consider upper and lower probabilities.

For distribution P_1, $P_1(X_i = 1) = 1$ and for any $i > 1$ we have $P_1(X_i = 1 | A(X_{1:i-1})) = 1$ whenever $\underline{P}(A(X_{1:i-1})) > 0$. Note that for any event $A(X_{1:i-1})$: if $A_1 \in A$, then $P_1(A) = 1$; if $A_1 \notin A$, then $P_1(A) = 0$. For distribution P_2, $P_2(X_i = 1) = 1 - \delta$ for any $i > 0$. Additionally, for any event $A(X_{1:i-1})$ we have $P_2(X_i = 1 | A(X_{1:i-1}))$ either equal to $1 - \delta$ or 1 whenever $\underline{P}(A) > 0$: if $A_1 \notin A$, then $\underline{P}(A) = 0$ (due to P_1); so suppose $A_1 \in A$, and note that if $A_0 \in A$, then $P_2(X_i = 1 | A) = 1 - \delta$, and if $A_0 \notin A$, then $P_2(X_i = 1 | A) = 1$. For distribution P_3, we have $P_3(X_i = 1) = 1 - \delta$ and for any $i > 1$ we have $P_3(X_i = 1 | A(X_{1:i-1})) = 1 - \delta$ for any nonempty event $A(X_{1:i-1})$. Hence we have $\underline{P}(X_i = 1) = \underline{P}(X_i = 1 | A(X_{1:i-1})) = 1 - \delta$ and $\overline{P}(X_i = 1) = \overline{P}(X_i = 1 | A(X_{1:i-1})) = 1$ whenever $\underline{P}(A(X_{1:i-1})) > 0$.

The weak law of larger numbers fails because, for any $\epsilon \in (0, 1 - \delta)$,

$$\lim_{n\to\infty} \underline{P}\left(\frac{\sum_{i=1}^n \underline{E}[X_i]}{n} - \epsilon < \frac{\sum_{i=1}^n X_i}{n} < \frac{\sum_{i=1}^n \overline{E}[X_i]}{n} + \epsilon \right) = 1 - P_2(A_0) = 1 - \delta.$$

This follows from the fact that, for any integer $n > 0$, $P_1\left(\sum_{i=1}^n X_i / n = 1\right) = 1$,

[1] Example 1 of a previous publication [2] claims to convey the same message, but that example is flawed in that expression (11) does not hold.

$$\forall \epsilon > 0: \quad P_2\left((1-\delta) - \epsilon < \sum_{i=1}^{n} X_i/n < 1 + \epsilon\right) = 1 - P_2(A_0) = 1 - \delta,$$

$$\forall \epsilon > 0: \quad \lim_{n\to\infty} P_3\left((1-\delta) - \epsilon < \sum_{i=1}^{n} X_i/n < (1-\delta) + \epsilon\right) = 1.$$

3.2. Forward factorization

De Cooman and Miranda have introduced a condition called *forward factorization* for variables $X_{1:n}$ that leads to interesting laws of large numbers [4, Def. 1]. Forward factorization requires:

- for each $i \in [2, n]$, for any bounded function f of X_i and any non-negative bounded function g of $X_{1:i-1}$,

$$\underline{E}[g(X_{1:i-1})(f(X_i) - \underline{E}[f(X_i)])] \geqslant 0. \tag{12}$$

The second part of the next proposition conveys a possibly more intuitive characterization of forward factorization:

Proposition 1. *Forward factorization for variables $X_{1:n}$ is equivalent both to*

- *for each $i \in [2, n]$, for any bounded function f of X_i and any event A defined by variables $X_{1:i-1}$,*

$$\underline{E}[A(X_{1:i-1})(f(X_i) - \underline{E}[f(X_i)])] \geqslant 0,$$

and to

- *for each $i \in [2, n]$, for any bounded function f of X_i and any event A defined by variables $X_{1:i-1}$,*

$$\overline{E}^{>}[f(X_i)|A(X_{1:i-1})] \leqslant \overline{E}[f(X_i)] \quad \text{whenever } \overline{P}(A(X_{1:i-1})) > 0.$$

The proof of this proposition is in Appendix B. The proof only assumes finite additivity. Note that under countable additivity and standard conditioning (expression (3)), forward regular irrelevance implies forward factorization.

Forward factorization implies a valuable inequality that is used in Section 4:

Proposition 2. *For bounded and nonnegative functions f_i, forward factorization of $X_{1:n}$ implies*

$$\overline{E}\left[\prod_{i=1}^{n} f_i(X_i)\right] \leqslant \prod_{i=1}^{n} \overline{E}[f_i(X_i)]. \tag{13}$$

The proof of this proposition is presented in Appendix C.

4. Bounded variables

Take variables $X_1, \ldots, X_n$ such that $\sup X_i - \inf X_i \leqslant \beta_i$ for $\beta_i < \infty$. The following inequalities, proved under several assumptions in theorems to be presented, are counterparts of Hoeffding inequality [8,17]:

$$\overline{P}\left(\sum_{i=1}^{n} (X_i - \overline{E}[X_i]) \geqslant \epsilon\right) \leqslant e^{-2\epsilon^2/\gamma_n}, \tag{14}$$

$$\overline{P}\left(\sum_{i=1}^{n} (X_i - \underline{E}[X_i]) \leqslant -\epsilon\right) \leqslant e^{-2\epsilon^2/\gamma_n}. \tag{15}$$

These concentration inequalitites are similar to, but slightly tighter than, inequalities by De Cooman and Miranda [4, Remark 2]. Note that results in this section are proved under the assumption of countable additivity and definitions of conditioning and irrelevance presented earlier, while De Cooman and Miranda adopt Walley's theory; the matter is discussed in more detail in Section 6.

The next theorem assumes a factorization that is implied by forward regular irrelevance (or by forward factorization); its proof, presented in Appendix D, is remarkably similar to usual proofs of the Hoeffding inequality [8].

Theorem 1. *Suppose bounded variables $X_1, \ldots, X_n$ satisfy expression (13) for bounded and nonnegative functions f_i. If $\gamma_n \doteq \sum_{i=1}^{n} \beta_i^2 > 0$, then expressions (14) and (15) hold.*

Theorem 1 leads to simple proofs of laws of large numbers stated by De Cooman and Miranda [4]. The proof of the following theorem is presented in Appendix E. The third expression in the theorem corresponds to a finitary version of the usual strong law of large numbers [9]; because countable additivity is assumed, limits can be taken if desired (as in the last two expressions of Theorem 4).

In the next theorem and later we use

$$\underline{\mu}_n \doteq (1/n)\sum_{i=1}^{n}\underline{E}[X_i], \quad \overline{\mu}_n \doteq (1/n)\sum_{i=1}^{n}\overline{E}[X_i].$$

Theorem 2. *If bounded variables $X_1, X_2, \ldots$ are such that $X_1, \ldots, X_n$ satisfy expressions (14) and (15) for any $n > 1$, then for any $\epsilon > 0$,*

$$\forall n \geqslant 1: \quad \underline{P}\left(\underline{\mu}_n - \epsilon < \frac{\sum_{i=1}^{n} X_i}{n} < \overline{\mu}_n + \epsilon\right) \geqslant 1 - 2e^{-2n\epsilon^2/(\max_i \beta_i^2)};$$

$$\lim_{n\to\infty} \underline{P}\left(\underline{\mu}_n - \epsilon < \frac{\sum_{i=1}^{n} X_i}{n} < \overline{\mu}_n + \epsilon\right) = 1;$$

$$\exists N: \quad \forall N': \quad \underline{P}\left(\forall n \in [N, N+N']: \underline{\mu}_n - \epsilon < \frac{\sum_{i=1}^{n} X_i}{n} < \overline{\mu}_n + \epsilon\right) > 1 - 2\epsilon;$$

where n, N and N' denote positive integers.

Corollary 1. *Suppose bounded variables $X_1, X_2, \ldots$ satisfy forward regular irrelevance or forward factorization. Then, for any $\epsilon > 0$, the three expressions in Theorem 2 hold.*

We move to weak forward regular irrelevance and obtain an analog of the Azuma inequality [1,7]. It is interesting to note that the proof of the following theorem, presented in Appendix F, is remarkably similar to the usual proof of the original Azuma inequality [1]. In Section 6, we comment on the similarities between the next two theorems and results by De Cooman and Miranda [4, Sec.4.1].

Theorem 3. *Suppose bounded variables $X_1, \ldots, X_n$ satisfy weak forward regular irrelevance. If $\gamma_n \doteq \sum_{i=1}^{n}\beta_i^2 > 0$, then expressions (14) and (15) hold.*

We now present laws of large numbers under weak forward regular irrelevance, that follow directly from Theorems 2 and 3. De Cooman and Miranda prove a similar pair of laws by resorting to their theory of *forward irrelevant natural extensions* [4, Sec. 4.1]; again, recall that their results do not assume countable additivity, as discussed in Section 6.

Corollary 2. *Suppose bounded variables $X_1, X_2, \ldots$ satisfy weak forward regular irrelevance. Then, for any $\epsilon > 0$, the three expressions in Theorem 2 hold.*

5. Laws of large numbers without boundedness

We now consider variables without bounds in their ranges under the assumption of weak forward regular irrelevance; the resulting laws of large numbers are the main contribution of the paper. In this section, we again assume that countable additivity holds (expression (2); that is, countable additivity of each element P of the credal set). We also assume, again, that standard (Kolmogorovian) conditioning is adopted. Thus our setup is close to the standard one; we only depart from the Kolmogorovian tradition in explicitly letting a *set* of expectation functionals to be permissible given a set of assessments.

The proof employs a sequence of variables $\{Y_n\}$ defined as follows, for a fixed P:

$$Y_n \doteq \sum_{i=1}^{n} X_i - E_P[X_i|X_{1:i-1}].$$

The key observation is that $\{Y_n\}$ is a *martingale* with respect to P. The properties of this martingale are explored in the proof of the next theorem, presented in Appendix G.

Theorem 4. *Suppose variables $X_1, X_2, \ldots$ satisfy weak forward regular irrelevance. Suppose further that $\underline{E}[X_i]$ and $\overline{E}[X_i]$ are finite quantities such that*[2] *$\overline{E}[X_i] - \underline{E}[X_i] \leqslant \delta$, and the variance of any X_i with respect to any element P of the credal set is no larger than a finite quantity σ^2. Then, for any $\epsilon > 0$,*

$$\forall n \geqslant 1: \quad \underline{P}\left(\underline{\mu}_n - \epsilon < \frac{\sum_{i=1}^{n} X_i}{n} < \overline{\mu}_n + \epsilon\right) \geqslant 1 - \frac{\sigma^2 + \delta^2}{\epsilon^2 n},$$

$$\exists N: \quad \forall N': \quad \underline{P}\left(\forall n \in [N, N+N']: \underline{\mu}_n - \epsilon < \frac{\sum_{i=1}^{n} X_i}{n} < \overline{\mu}_n + \epsilon\right) > 1 - 2\epsilon,$$

[2] As noted in Appendix G, it is possible to remove the need for δ; we thank a reviewer for providing sharper inequalities that do not require δ.

where n, N and N′ denote positive integers. Consequently,

$$\forall\epsilon > 0: \quad \lim_{n\to\infty} \underline{P}\left(\underline{\mu}_n - \epsilon < \frac{\sum_{i=1}^n X_i}{n} < \overline{\mu}_n + \epsilon\right) = 1,$$
$$\overline{P}\left(\limsup_{n\to\infty}\left(\frac{\sum_{i=1}^n X_i}{n} - \overline{\mu}_n\right) > 0\right) = 0,$$
$$\overline{P}\left(\liminf_{n\to\infty}\left(\frac{\sum_{i=1}^n X_i}{n} - \underline{\mu}_n\right) < 0\right) = 0.$$

One final question is whether it is possible to remove the condition that variances must be finite in this theorem. Even in the standard theory one finds that laws of large numbers fail if restrictions on variances are simply removed [14]. Typically when restrictions on variances are removed one requires variables to be identically distributed [13,23]. In our setting the most natural requirement would be to ask all credal sets containing marginal distributions to be identical. This is the situation where, as Epstein and Schneider aptly call, variables are *independent and indistinguishably distributed* [12, Eq. 2.2]. Alas, the following example shows that this assumption of indistinguishability fails to substitute for restrictions on variances.

Example 2. Assume countable additivity. Consider integer-valued random variables $X_1, X_2, \ldots$ that satisfy forward factorization. The only available assessment is $\underline{E}[X_i] = \overline{E}[X_i] = 0$ for every X_i (all marginal credal sets are identical). Now consider a joint distribution P that satisfies all assessments and irrelevances: P is the product measure of P_i defined as: $P_i(X_i = -i) = P_i(X_i = i) = 1/(2i)$, $P_i(X_i = 0) = 1 - 1/i$. As shown by Geller [14, Example 1] this joint distribution leads to failure of the weak law of large numbers. Consequently, the lower probability of $|\sum_{i=1}^n X_i|/n \leqslant \epsilon$ does not go to 1 as n grows without bound.

Note that if one does assume that each joint distribution in the credal set has identical marginals, then further results can be proved by combining Theorem 4 with truncation techniques [23, Sec. 4.7]. That is, by assuming forward regular irrelevance (not *weak* forward regular irrelevance), one can consider the sequence of truncated variables $X_i I_{|X_i| \leqslant i}$. Countable additivity then allows one to discard the contribution, for each joint distribution, of the variables $X_i I_{|X_i|>i}$ (because $\sum_{i=1}^\infty P(X_i \neq X_i I_{|X_i|>i}) \leqslant \sum_{i=1}^\infty P(|X_1| > i) \leqslant E[|X_1|] < \infty$ for every joint distribution, and then the Borel–Cantelli lemma guarantees that the differences are negligible). Consequently, the behavior of the original sequence can be investigated, possibly using Theorem 4, by studying the truncated sequence $\{X_i I_{|X_i| \leqslant i}\}$. The extent to which such techniques can lead to conceptually interesting results is yet to be understood, given that the assumption of identical marginals for *each* joint distribution seems to clash in spirit with regular irrelevance and the factorization conditions studied in this paper.

Finding a condition that is both stronger than indistinguishability and more intuitive than assuming identical marginals for every joint distribution, and that still leads to laws of large numbers, is an open problem.

6. A comment on Walley's theory of lower previsions

The work by De Cooman and Miranda on laws of large numbers adopts Walley's theory of lower previsions [4], and focuses on bounded variables. We now comment on the extent to which results in Section 4 apply to Walley's theory; to do so, we first review basic facts about full conditional measures.

The theory of full conditional measures, whose most vocal advocate was de Finetti [6], offers an alternative to the standard (Kolmogorovian) theory. The idea is to take the conditional expectation $E[X|A]$ as a primitive that is well defined even if the event A has zero probability. Four axioms are imposed on conditional expectations: for any nonempty event A,

(1) if $\alpha \leqslant X(\omega) \leqslant \beta$ for all $\omega \in A$, then $\alpha \leqslant E[X|A] \leqslant \beta$;
(2) $E[X+Y|A] = E[X|A] + E[Y|A]$;
(3) $E[I_A|A] = 1$;
(4) $E[I_A X|B] = E[X|A]E[I_A|B]$ whenever $A \subseteq B$.

If a function $E[\cdot|\cdot]$ satisfies these axioms, we call it a *full conditional expectation*. We can then define a set-function $P(B|A) \doteq E[I_B|A]$ for any event B and any nonempty event A. Such P is usually called a *full conditional measure* [10,20], and it satisfies, for every nonempty event C:

(1) $P(C|C) = 1$;
(2) $P(A|C) \geqslant 0$ for all A;
(3) $P(A \cup B|C) = P(A|C) + P(B|C)$ for all disjoint A and B;
(4) $P(A \cap B|C) = P(A|B \cap C)P(B|C)$ for all A and B such that $B \cap C \neq \emptyset$.

If we are dealing with full conditional expectations, then, given two variables X and Y, the expectation $E[X|Y=y]$ is well defined for every y such that $\{Y=y\}$ is nonempty. Given a set K of full conditional expectations, we can define lower and upper conditional expectations respectively as $\underline{E}[X|Y=y] \doteq \inf_{E\in K} E[X|Y=Y]$ and $\overline{E}[X|Y=y] \doteq \sup_{E\in K} E[X|Y=y]$ for every y, without concern on whether $P(Y=y) = 0$ or not. Note that Radon–Nikodym derivatives do not always satisfy the axioms for full conditional measures when the conditioning event has probability zero [26,27]; hence there are substantial differences between full conditional measures and standard (Kolmogorovian-style) probability measures.

For a single expectation functional, *disintegrability* holds with respect to Y when $E[X] = E[E[X|Y]]$ for any X. Disintegrability may fail for a single finitely additive probability measure over an infinite space [6,10]; that is, there is a finitely additive probability measure P such that $E_P[X] > E_P[E_P[X|Y]]$. There are theories that do not adopt countable additivity but still obtain disintegrability. The theories of coherent behavior by Heath and Sudderth [16] and by Lane and Sudderth [21] axiomatize the *strategic* measures of Dubins and Savage [11], and prescribe probability measures that disintegrate appropriately along predefined partitions. It would be sufficient for our purposes to have sets of such strategic measures disintegrating over suitable partitions (note that despite the drawbacks of strategic measures [18], they do admit non-trivial laws of large numbers [19]).

For an upper expectation, define *disintegrability* with respect to Y to mean

$$\overline{E}[X] \leqslant \overline{E}[\overline{E}[X|Y]] \quad \text{for any } X. \tag{16}$$

Disintegrability without further qualification means disintegrability with respect to any Y. Walley's theory deals with full conditional measures but adds a condition of conglomerability that implies disintegrability of upper expectations [31, Sec. 6.3.5(5)].

We now return to the main purpose of this section; that is, we analyze the validity of results in Section 4 within Walley's theory of lower previsions.

Propositions 1 and 2 hold for sets of full conditional measures (without any assumption of countable additivity). Hence Theorems 1 and 2 hold for such sets (and in Walley's theory). Note that Theorem 2 presents finitary versions of the laws of large numbers that are appropriate when countable additivity is not assumed; if countable additivity is assumed, then limits can be taken as in Theorem 4. Thus the main results in De Cooman and Miranda's work are recovered, with different proofs.

Also, we have that forward irrelevance leads to forward factorization and then to the laws of large numbers in Theorem 2 (this is proved by De Cooman and Miranda using a different strategy). To see this, note that using Walley's definition of epistemic irrelevance we have: if Y is epistemically irrelevant to X, then

$$\overline{E}[f(X)|A(Y)] \leqslant \overline{E}[\overline{E}[f(X)|Y, A(Y)]|A(Y)] = \overline{E}[\overline{E}[f(X)|Y]|A(Y)] = \overline{E}[\overline{E}[f(X)]|A(Y)] = \overline{E}[f(X)]$$

for any function f of X and any event $A(Y)$ defined by Y such that $\overline{P}(A(Y)) > 0$.[3] Thus forward irrelevance implies forward factorization, using Proposition 1, and this leads to the laws of large numbers.

Theorem 3 is more delicate as the use of *elementwise* disintegrability in the proof is not really meaningful in Walley's theory. However we can derive the result by assuming only disintegrability of upper expectations and the following condition, that adapts weak forward regular irrelevance to Walley's theory:

- for each $i \in [2, n]$,

$$\overline{E}[X_i|X_{1:i-1}] \leqslant \overline{E}[X_i] \quad \text{and} \quad \underline{E}[X_i|X_{1:i-1}] \geqslant \underline{E}[X_i]. \tag{17}$$

The proof of the following theorem is given in Appendix H.

Theorem 5. *Suppose bounded variables $X_1, \ldots, X_n$ satisfy the condition given by expression (17). Assume disintegrability of upper expectations with respect to $X_{1:i-1}$ for $i \in \{2, \ldots, n\}$. If $\gamma_n \doteq \sum_{i=1}^n \beta_i^2 > 0$, then expressions (14) and (15) hold.*

Using Theorems 2 and 5:

Corollary 3. *Suppose bounded variables $X_1, X_2, \ldots$ satisfy the condition given by expression (17). Assume disintegrability of upper expectations with respect to $X_{1:n}$ for $n > 1$. Then, for any $\epsilon > 0$, the three expressions in Theorem 2 hold.*

7. Discussion

The concentration inequalities and laws of large numbers proved in this paper assume rather weak conditions of irrelevance. When compared to usual laws of large numbers, both premises and consequences are weaker: expectations are not assumed precisely known, and convergence is interval-valued.

Inequalities (14) and (15), and related theorems, slightly sharpen results in De Cooman and Miranda's seminal work [4]. The proofs of these inequalities, as presented in this paper, are rather close to well-known methods in standard probability theory. It should be noted that De Cooman and Miranda already comment on the similarity between their inequalities and Hoeffding's. Note also that De Cooman and Miranda's results generalize several previous efforts, such as by Epstein and Schneider, where credal sets are convex and closed and satisfy a condition of supermodularity [12, Sec. 4].

Theorem 4 is possibly the most valuable contribution of this paper. The strategy of the proof is to translate weak assumptions of irrelevance into facts regarding martingales, and to adapt results for martingales to this setting. This strategy keeps

[3] This derivation is not valid for general sets of full conditional measures, but it is valid for Walley's theory. Thanks to Matthias Troffaes for useful discussion about this point.

the proof close to well-known results in probability theory. The connection between lower and upper expectations and the theory of martingales seems rather natural [3,29], but the relationship between epistemic/regular irrelevance and martingales does not appear to have been explored in depth so far. We note that the basic constraint defining martingales (that is, $E[Y_n|X_{1:n-1}] = Y_{n-1}$) is preserved by convex combination of distributions; therefore, the study of martingales seems appropriate when one deals with convex sets of probability distributions – certainly it seems less contorted than the analysis through stochastic independence, as stochastic independence is *not* preserved by convex combination.

There are some open questions that call for study. First, it would be valuable to determine whether countable additivity (or at least disintegrability assumptions) are really needed in Theorems 3 and 4. Another question is whether the condition on variances in Theorem 4 can be replaced by some weaker condition; this prompts the question of whether there is some interesting notion of "identically distributed" variables in the present setting.

Acknowledgements

Thanks to Gert De Cooman and Enrique Miranda for their generous suggestions regarding content and presentation, to Paolo Vicig for discussion on finite additivity, to Matthias Troffaes for discussion concerning full conditional measures, and to Teddy Seidenfeld for sharing recent work [25] that led to the proof of Proposition 1. Thanks to a reviewer that suggested several important improvements, including the short proof of Lemma 2, the definition of irrelevance under countable additivity, and refinements for inequalities in Theorem 4.

Mike Smithson's question about restrictions on variances led to the discussion around Example 2.

Appendix A. Regular conditioning

Using an earlier proposal by Walley himself [30, Sec. 7], we can define regular conditioning without any reference to individual probability measures:

$$\overline{E}^{>}[X|A] \doteq \inf(\alpha : \overline{E}[A(X-\alpha)] \leqslant 0) \quad \text{if } \overline{P}(A) > 0, \qquad (18)$$
$$\overline{E}^{>}[X|A] \text{ undefined} \quad \text{if } \overline{P}(A) = 0,$$

Lemma 1. *If* $\overline{P}(A) > 0$,

$$\overline{E}^{>}[X|A] = \sup_{E\in K:P(A)>0} E[X|A] = \inf(\alpha : \overline{E}[A(X-\alpha)] \leqslant 0).$$

Proof. We have:

$$\inf(\alpha : \overline{E}[A(X-\alpha)] \leqslant 0) = \inf\left(\alpha : \sup_{E\in K}(E[AX] - \alpha E[A]) \leqslant 0\right)$$
$$= \inf\left(\alpha : \sup_{E\in K:P(A)>0}(E[AX] - \alpha E[A]) \leqslant 0\right) = \inf\left(\alpha : \sup_{E\in K:P(A)>0} E[X|A] \leqslant \alpha\right) = \sup_{E\in K:P(A)>0} E[X|A]. \qquad \square$$

Appendix B. Proof of Proposition 1

We divide the proof in two steps.

Lemma 2. *Forward factorization for variables* $X_{1:n}$ *is equivalent to: for each* $i \in [2, n]$, *for any bounded function* f *of* X_i *and any event* A *defined by variables* $X_{1:i-1}$,

$$\underline{E}[A(X_{1:i-1})(f(X_i) - \underline{E}[f(X_i)])] \geqslant 0. \qquad (19)$$

Proof. If condition (12) holds, then by selecting $g(X_{1:i-1}) = A(X_{1:i-1})$ we obtain expression (19). Now assume conversely that expression (19) holds. For a fixed i, define $Y = g(X_{1:i-1})$ and construct a sequence of simple functions indexed by $j \geqslant 1$:

$$Y_j^{\geqslant} \doteq \sum_{k=1}^{2^j+1} \frac{M_Y}{2^j} k A_{j,k},$$

where, for $k \in \{0, 1, 2, \ldots, 2^j + 1\}$, $A_{j,k}$ is the indicator function of

$$\left\{\omega : \frac{M_Y}{2^j}(k-1) \leqslant Y(\omega) < \frac{M_Y}{2^j}k\right\}.$$

For any P of interest, expression (19) implies

$$E_P[Y_j^{\geqslant}(f(X_i) - \underline{E}[f(X_i)])] \geqslant 0,$$

because $Y_j^{\geqslant}$ is a weighted sum of indicator functions, where weights are all non-negative. As j grows, the simple functions $Y_j^{\geqslant}$ converge uniformly to Y. Uniform convergence of $\{Y_j^{\geqslant}\}$ and boundedness of f imply uniform convergence of $Y_j^{\geqslant}(f(X_i) - \underline{E}[f(X_i)])$ to $Y(f(X_i) - \underline{E}[f(X_i)])$. Consequently, using expression (1),

$$\lim_{j\to\infty} E_P[Y_j^{\geqslant}(f(X_i) - \underline{E}[f(X_i)])] = E_P[Y(f(X_i) - \underline{E}[f(X_i)])],$$

and then $E_P[Y(f(X_i) - \underline{E}[f(X_i)])] \geqslant 0$ for every P of interest, as desired. □

Lemma 3. *Forward factorization for variables $X_{1:n}$ is equivalent to: for each $i \in [2, n]$, for any bounded function f of X_i and any event A defined by variables $X_{1:i-1}$,*

$$\overline{E}^{>}[f(X_i)|A(X_{1:i-1})] \leqslant \overline{E}[f(X_i)] \quad \textit{whenever } \overline{P}(A(X_{1:i-1})) > 0.$$

Proof. Denote $f(X_i)$ by X and $A(X_{1:i-1})$ by A. Using Lemma 2 and expression (18), it is enough to show that $\overline{E}[A(X - \overline{E}[X])] \leqslant 0$ is equivalent to $\inf(\alpha : \overline{E}[A(X - \alpha)] \leqslant 0) \leqslant \overline{E}[X]$ whenever $\overline{P}(A) > 0$. Clearly if $\overline{E}[A(X - \overline{E}[X])] \leqslant 0$ then $\inf(\alpha : \overline{E}[A(X - \alpha)] \leqslant 0) \leqslant \overline{E}[X]$ (just take $\alpha = \overline{E}[X]$). And because $\overline{E}[A(X - \alpha)]$ is decreasing in α, $\inf(\alpha : \overline{E}[A(X - \alpha)] \leqslant 0) \leqslant \overline{E}[X]$ implies $\overline{E}[A(X - \overline{E}[X])] \leqslant 0$. □

Appendix C. Proof of Proposition 2

Proof. For any X, Y, we have $\overline{E}[X] - \overline{E}[Y] \leqslant \overline{E}[X - Y]$ because

$$\overline{E}[X] = \overline{E}[X - Y + Y] \leqslant \overline{E}[X - Y] + \overline{E}[Y].$$

Define $f^i \doteq \prod_{j=1}^{i} f_j(X_j)$; then:

$$\begin{aligned}
\overline{E}[f^n] - \overline{E}[f^{n-1}\overline{E}[f_n(X_n)]] &\leqslant \overline{E}[f^n - f^{n-1}\overline{E}[f_n(X_n)]] \\
&= \overline{E}[f^{n-1}(f_n(X_n) - \overline{E}[f_n(X_n)])] \\
&= -\underline{E}[f^{n-1}(-f_n(X_n) - \underline{E}[-f_n(X_n)])] \\
&\leqslant 0 \quad (\text{using expression (12)}).
\end{aligned}$$

Hence $\overline{E}[f^n] \leqslant \overline{E}[f^{n-1}\overline{E}[f_n(X_n)]]$, and because $f_n(X_n) \geqslant 0$ (thus $\overline{E}[f_n(X_n)] \geqslant 0$), we have $\overline{E}[f^n] \leqslant \overline{E}[f^{n-1}]\overline{E}[f_n(X_n)]$. We obtain the desired result by iterating this reasoning. □

Appendix D. Proof of Theorem 1

Proof. If $X \geqslant 0$, then $I_{\{X \geqslant \epsilon\}} \leqslant X/\epsilon$ for any $\epsilon > 0$; using the fact that if $X \leqslant Y$ then $\overline{E}[X] \leqslant \overline{E}[Y]$, we obtain $\overline{P}(X \geqslant \epsilon) \leqslant \overline{E}[X]/\epsilon$ (a Markov inequality). Consequently, for $s > 0$, any variable X satisfies

$$\overline{P}(X \geqslant \epsilon) = \overline{P}(e^{sX} \geqslant e^{s\epsilon}) \leqslant e^{-s\epsilon}\overline{E}[\exp(sX)].$$

Using this inequality and Proposition 2:

$$\overline{P}\left(\sum_{i=1}^{n}(X_i - \overline{E}[X_i]) \geqslant \epsilon\right) \leqslant e^{-s\epsilon}\overline{E}\left[\exp\left(\sum_{i=1}^{n} s(X_i - \overline{E}[X_i])\right)\right] \leqslant e^{-s\epsilon}\prod_{i=1}^{n}\overline{E}[\exp(s(X_i - \overline{E}[X_i]))].$$

We now use the variant of Hoeffding's result given by expression (23): If variable X satisfies $a \leqslant X \leqslant b$ and $\overline{E}[X] \leqslant 0$, then $\overline{E}[\exp(sX)] \leqslant \exp(s^2(b-a)^2/8)$ for any $s > 0$. Hence $\overline{E}[\exp(s(X_i - \overline{E}[X_i]))] \leqslant \exp(s^2\beta_i^2/8)$ and

$$\overline{P}\left(\sum_{i=1}^{n}(X_i - \overline{E}[X_i]) \geqslant \epsilon\right) \leqslant e^{-s\epsilon}e^{s^2\gamma_n/8} \leqslant e^{-2\epsilon^2/\gamma_n},$$

where the last inequality is obtained by taking $s = 4\epsilon/\gamma_n$. This proves the first inequality in the theorem; the second inequality is proved by considering the upper probability $\overline{P}(\sum_{i=1}^n((-X_i) - \overline{E}[-X_i]) \geqslant \epsilon)$ and noting that $\underline{E}[X_i] = -\overline{E}[-X_i]$. □

Appendix E. Proof of Theorem 2

Proof. Define $\beta^2 \doteq \max_i \beta_i^2$. Noting that $\underline{P}(A) = 1 - \overline{P}(A^c)$ for any event A, using subadditivity of upper probability, and then expressions (14) and (15):

$$\underline{P}\left(\underline{\mu}_n - \epsilon < \frac{\sum_{i=1}^n X_i}{n} < \overline{\mu}_n + \epsilon\right) = 1 - \overline{P}\left(\left\{\sum_{i=1}^n X_i - n\overline{\mu}_n \geqslant n\epsilon\right\} \cup \left\{\sum_{i=1}^n X_i - n\underline{\mu}_n \leqslant n\epsilon\right\}\right)$$
$$\geqslant 1 - \overline{P}\left(\sum_{i=1}^n X_i - n\overline{\mu}_n \geqslant n\epsilon\right) - \overline{P}\left(\sum_{i=1}^n X_i - n\underline{\mu}_n \leqslant n\epsilon\right) \geqslant 1 - e^{-\frac{2n\epsilon^2}{\beta^2}} - e^{-\frac{2n\epsilon^2}{\beta^2}}$$
$$= 1 - 2e^{-\frac{2n\epsilon^2}{\beta^2}}.$$

By taking the limit as n grows without bound, we obtain that the lower probability goes to one. Now consider the strong law of large numbers. For any $\epsilon > 0$, $N > 0$ and $N' > 0$,

$$\overline{P}\left(\exists n \in [N, N+N'] : \frac{\sum_{i=1}^n X_i}{n} \geqslant \overline{\mu}_n + \epsilon\right) \leqslant \sum_{n=N}^{N+N'} \overline{P}\left(\frac{\sum_{i=1}^n X_i}{n} \geqslant \overline{\mu}_n + \epsilon\right) \leqslant \sum_{n=N}^{N+N'} e^{-2n\epsilon^2/\beta^2} = (e^{-2N\epsilon^2/\beta^2}) \sum_{n=0}^{N'} e^{-2n\epsilon^2/\beta^2}$$
$$= (e^{-2N\epsilon^2/\beta^2}) \frac{1 - e^{-2(N'+1)\epsilon^2/\beta^2}}{1 - e^{-2\epsilon^2/\beta^2}} < \frac{e^{-2N\epsilon^2/\beta^2}}{1 - e^{-2\epsilon^2/\beta^2}}.$$

Consequently,

$$\overline{P}\left(\exists n \in [N, N+N'] : \frac{\sum_{i=1}^n X_i}{n} \geqslant \overline{\mu}_n + \epsilon\right) < \epsilon,$$

provided that N is a positive integer such that $N > -(\beta^2/(2\epsilon^2)) \ln \epsilon (1 - e^{-2\epsilon^2/\beta^2})$. An analogous argument leads to

$$\overline{P}\left(\exists n \in [N, N+N'] : \frac{\sum_{i=1}^n X_i}{n} \leqslant \underline{\mu}_n - \epsilon\right) < \epsilon.$$

By superadditivity of upper probability: for any $\epsilon > 0$, there is N such that for any N',

$$\underline{P}\left(\forall n \in [N, N+N'] : \underline{\mu}_n - \epsilon < \frac{\sum_{i=1}^n X_i}{n} < \overline{\mu}_n + \epsilon\right) > 1 - 2\epsilon,$$

as desired. □

Appendix F. Proof of Theorem 3

Proof. Using both Markov's inequality (as in the proof of Theorem 1) and elementwise disintegrability, for any $s > 0$,

$$\overline{P}\left(\sum_{i=1}^n (X_i - \overline{E}[X_i]) \geqslant \epsilon\right) \leqslant e^{-s\epsilon}\overline{E}\left[\exp\left(\sum_{i=1}^n s(X_i - \overline{E}[X_i])\right)\right] = e^{-s\epsilon} \sup_P E_P\left[E_P\left[\exp\left(\sum_{i=1}^n s(X_i - \overline{E}[X_i])\right)\middle| X_{1:n-1}\right]\right]$$
$$= e^{-s\epsilon} \sup_P E_P\left[\exp\left(\sum_{i=1}^{n-1} s(X_i - \overline{E}[X_i])\right) h_P(X_{1:n-1})\right],$$

where $h_P(X_{1:n-1}) = E_P[\exp(s(X_n - \overline{E}[X_n]))|X_{1:n-1}]$. Due to weak forward regular irrelevance:

$$E_P[X_n|X_{1:n-1}] \leqslant \overline{E}[X_n|X_{1:n-1}] \leqslant \overline{E}[X_n],$$

whenever the event defined by $X_{1:n-1}$ has nonzero probability with respect to P. For these events we now apply expression (23); other events have probability zero and do not matter when the outer expectation is calculated. So, for events of interest,

$$E_P[X_n - \overline{E}[X_n]|X_{1:n-1}] \leqslant 0.$$

We apply expression (23) to P in conditional form (that is: if variable X satisfies $a \leqslant X \leqslant b$ and $E_P[X|A] \leqslant 0$, then $E_P[\exp(sX)|A] \leqslant \exp(s^2(b-a)^2/8)$ for any $s > 0$). Then:

$$h_P(X_{1:n-1}) = E_P[\exp(s(X_n - \overline{E}[X_n]))|X_{1:n-1}] \leqslant \exp(s^2\beta_n^2/8). \tag{20}$$

Given this inequality,

$$\overline{P}\left(\sum_{i=1}^{n}(X_i - \overline{E}[X_i]) \geqslant \epsilon\right) \leqslant e^{-s\epsilon}\sup_P E_P\left[\exp\left(\sum_{i=1}^{n-1}s(X_i - \overline{E}[X_i])\right)e^{s^2\beta_n^2/8}\right] \leqslant e^{-s\epsilon}e^{s^2\beta_n^2/8}\sup_P E_P\left[\exp\left(\sum_{i=1}^{n-1}s(X_i - \overline{E}[X_i])\right)\right].$$

These inequalities can be iterated to produce:

$$\overline{P}\left(\sum_{i=1}^{n}(X_i - \overline{E}[X_i]) \geqslant \epsilon\right) \leqslant e^{-s\epsilon}\exp\left(s^2\sum_{i=1}^{n}\beta_i^2/8\right) = e^{-s\epsilon}e^{s^2\gamma_n/8}.$$

Finally, by selecting $s = 4\epsilon/\gamma_n$,

$$\overline{P}\left(\sum_{i=1}^{n}(X_i - \overline{E}[X_i]) \geqslant \epsilon\right) \leqslant e^{-2\epsilon^2/\gamma_n}.$$

The second inequality in the theorem is proved by noting that weak forward factorization of $X_1,\ldots,X_n$ implies weak forward factorization of $-X_1,\ldots,-X_n$ (as $\underline{E}[X_i] = -\overline{E}[-X_i]$), and by focusing on $\overline{P}(\sum_{i=1}^{n}((-X_i) - \overline{E}[-X_i]) \geqslant \epsilon)$. □

Appendix G. Proof of Theorem 4

As noted before the statement of Theorem 4, we use the sequence $\{Y_n\}$ defined as

$$Y_n \doteq \sum_{i=1}^{n} X_i - E_P[X_i|X_{1:i-1}].$$

This sequence is a function of all variables $X_{1:n}$ such that

$$E_P[Y_n|X_{1:n-1}] = \left(\sum_{i=1}^{n-1}X_i - E_P[X_i|X_{1:i-1}]\right) + E_P[X_n - E_P[X_n|X_{1:n-1}]|X_{1:n-1}] = Y_{n-1} + E_P[X_n|X_{1:n-1}] - E_P[X_n|X_{1:n-1}] = Y_{n-1};$$

so, $\{Y_n\}$ is a *martingale* with respect to P.

We now manipulate a number of standard conditional expectations, where the conditioning events that have positive probability with respect to P are the ones that matter. We have:

$$\begin{aligned} E_P[(Y_n - Y_{n-1})^2|X_{1:n-1}] &= E_P[Y_n^2|X_{1:n-1}] - 2E_P[Y_{n-1}Y_n|X_{1:n-1}] + Y_{n-1}^2 = E_P[Y_n^2|X_{1:n-1}] - 2Y_{n-1}E_P[Y_n|X_{1:n-1}] + Y_{n-1}^2 \\ &= E_P[Y_n^2|X_{1:n-1}] - 2Y_{n-1}Y_{n-1} + Y_{n-1}^2 = E_P[Y_n^2|X_{1:n-1}] - Y_{n-1}^2. \end{aligned}$$

And by taking expectations on both sides we obtain the following martingale property (note the use of elementwise disintegrability):

$$E_P[Y_n^2] = E_P[(Y_n - Y_{n-1})^2] + E_P[Y_{n-1}^2]. \tag{21}$$

Elementwise disintegrability also leads to

$$E_P[Y_n] = \sum_{i=1}^{n} E_P[X_i] - E_P[E_P[X_i|X_{1:i-1}]] = \sum_{i=1}^{n} E_P[X_i] - E_P[X_i] = 0.$$

Proof. We start with expression (21) for a fixed P. Because $Y_n - Y_{n-1} = X_n - E_P[X_n|X_{1:n-1}]$,

$$E_P[Y_n^2] = E_P[(X_n - E_P[X_n|X_{1:n-1}])^2] + E_P[Y_{n-1}^2].$$

Iterating the last expression, and denoting $E_P[X_i] - E_P[X_i|X_{1:i-1}]$ by Δ_i:

$$\begin{aligned} E_P[Y_n^2] &= \sum_{i=1}^{n} E_P[(X_i - E_P[X_i|X_{1:i-1}])^2] = \sum_{i=1}^{n} E_P[((X_i - E_P[X_i]) + (E_P[X_i] - E_P[X_i|X_{1:i-1}]))^2] \\ &= \sum_{i=1}^{n} E_P[(X_i - E_P[X_i])^2] + 2\Delta_i E_P[X_i - E_P[X_i]] + \Delta_i^2 = \sum_{i=1}^{n} E_P[(X_i - E_P[X_i])^2] + \Delta_i^2 \leqslant \sum_{i=1}^{n} \sigma^2 + \delta^2 = n(\sigma^2 + \delta^2), \end{aligned} \tag{22}$$

using weak forward regular irrelevance to conclude that $\Delta_i^2 \leqslant \delta^2$.[4]

After these preliminaries on the sequence $\{Y_n\}$, note that for any $\epsilon > 0$,

$$P\left(\underline{\mu}_n - \epsilon < \frac{\sum_{i=1}^n X_i}{n} < \overline{\mu}_n + \epsilon\right) = P\left(\sum_{i=1}^n \underline{E}[X_i] - \epsilon n < \sum_{i=1}^n X_i < \sum_{i=1}^n \overline{E}[X_i] + \epsilon n\right)$$
$$\geqslant P\left(\sum_{i=1}^n E_P[X_i|X_{1:i-1}] - \epsilon n < \sum_{i=1}^n X_i < \sum_{i=1}^n E_P[X_i|X_{1:i-1}] + \epsilon n\right),$$

using weak forward regular irrelevance. The last expression is equal to

$$P\left(-\epsilon < \frac{\sum_{i=1}^n X_i - E_P[X_i|X_{1:i-1}]}{n} < \epsilon\right) = P(|Y_n/n| < \epsilon).$$

By Chebyshev's inequality and expression (22),

$$P(|Y_n/n| < \epsilon) = 1 - P(|Y_n/n| \geqslant \epsilon) \geqslant 1 - \frac{E_P[Y_n^2]}{\epsilon^2 n^2} \geqslant 1 - \frac{\sigma^2 + \delta^2}{\epsilon^2 n}.$$

By combining these inequalities for any P of interest, the first inequality in the theorem is proved. By taking the limit as n grows without bound, we obtain

$$\lim_{n\to\infty} \underline{P}\left(\underline{\mu}_n - \epsilon < \frac{\sum_{i=1}^n X_i}{n} < \overline{\mu}_n + \epsilon\right) = 1.$$

The proof of the strong law of large numbers uses the same strategy, but replaces the appeal to Chebyshev's inequality by an appeal to the Kolmogorov–Hajek–Renyi inequality (expression (24)), as in the proof of the strong law of large numbers by Whittle [34, Thm. 14.2.3]. So, for a fixed P and for any $\epsilon > 0$, we proceed as previously to obtain:

$$P\left(\forall n \in [N, N+N'] : \underline{\mu}_n - \epsilon < \frac{\sum_{i=1}^n X_i}{n} < \overline{\mu}_n + \epsilon\right) \geqslant P\left(\forall n \in [N, N+N'] : -\epsilon < \frac{Y_n}{n} < \epsilon\right) = P(\forall n \in [N, N+N'] : |Y_n/n| < \epsilon).$$

As $\{0, Y_N, Y_{N+1}, \ldots, Y_{N+N'}\}$ forms a martingale, we use the Kolmogorov–Hajek–Renyi inequality (expression (24)) to produce:

$$P(\forall n \in [N, N+N'] : |Y_n/n| < \epsilon) \geqslant 1 - \frac{E_P[Y_N^2]}{\epsilon^2 N^2} - \sum_{i=N+1}^{N+N'} \frac{E_P(Y_i - Y_{i-1})^2}{\epsilon^2 i^2}.$$

Hence:

$$P(\forall n \in [N, N+N'] : |Y_n/n| < \epsilon) \geqslant 1 - \frac{\sum_{i=1}^N E_P[(X_i - E_P[X_i|X_{1:i-1}])^2]}{\epsilon^2 N^2} - \sum_{i=N+1}^{N+N'} \frac{E_P[(X_i - E_P[X_i|X_{1:i-1}])^2]}{\epsilon^2 i^2}$$
$$\geqslant 1 - \frac{\sigma^2 + \delta^2}{\epsilon^2 N} - \sum_{i=N+1}^{N+N'} \frac{\sigma^2 + \delta^2}{\epsilon^2 i^2} \quad \text{(using expression (22))}$$
$$\geqslant 1 - \frac{\sigma^2 + \delta^2}{\epsilon^2 N} - \sum_{i=N+1}^{\infty} \frac{\sigma^2 + \delta^2}{\epsilon^2 i^2}$$
$$\geqslant 1 - \frac{\sigma^2 + \delta^2}{\epsilon^2}\left(\frac{1}{N} + \int_N^\infty 1/i^2\, di\right) = 1 - \frac{\sigma^2 + \delta^2}{\epsilon^2}\left(\frac{1}{N} + \frac{1}{N}\right)$$
$$= 1 - 2\frac{\sigma^2 + \delta^2}{\epsilon^2 N}.$$

Consequently, for integer $N > (\sigma^2 + \delta^2)/\epsilon^3$, we obtain the desired inequality

$$\underline{P}\left(\forall n \in [N, N+N'] : \underline{\mu}_n - \epsilon < \frac{\sum_{i=1}^n X_i}{n} < \overline{\mu}_n + \epsilon\right) > 1 - 2\epsilon.$$

Using the Kolmogorov–Hajek–Renyi without an upper bound on n,

$$P\left(\forall n \geqslant N : \frac{\sum_{i=1}^n X_i}{n} < \overline{\mu}_n + \epsilon\right) \geqslant P\left(\forall n \geqslant N : \underline{\mu}_n - \epsilon \leqslant \frac{\sum_{i=1}^n X_i}{n} \leqslant \overline{\mu}_n + \epsilon\right) > 1 - 2\epsilon.$$

[4] A reviewer generously suggested a derivation that shows $E_P[Y_n^2] \leqslant n\sigma^2$, thus obtaining a sharper inequality and removing the need for δ. The strategy is to recall that the set of square-integrable functions is an Hilbert space; hence $E_P[\cdot|X_{1:n}]$ is the orthogonal projection onto the set of square-integrable $X_{1:n}$-measurable functions. Consequently, from the properties of Hilbert spaces, $E_P[(X - E_P[X|X_{1:n}])^2] \leqslant E_P[(X - E_P[X])^2]$ and then $E_P[Y_n^2] = \sum_{i=1}^n E_P[(X - E_P[X|X_{1:n}])^2] \leqslant \sum_{i=1}^n E_P[(X - E_P[X])^2] \leqslant n\sigma^2$.

Consequently,

$$P\left(\exists n \geqslant N : \frac{\sum_{i=1}^n X_i}{n} \geqslant \overline{\mu}_n + \epsilon\right) < 2\epsilon.$$

This is almost exactly the inequality obtained by De Cooman and Miranda [4, Thm. 7] for bounded variables. We now copy their reasoning [4, A.8] to obtain probabilities over lim sup and lim inf. Event $\Lambda = \{\omega : \limsup_n (1/n)\sum_{i=1}^n (X_i - \overline{E}[X_i]) > 0\}$ is equal to $\cap_{m\geqslant 1}\cap_{N\geqslant 1}\cup_{n\geqslant N}\Lambda_{m,n}$, where

$$\Lambda_{m,n} = \left\{\omega : (1/n)\sum_{i=1}^n (X_i(\omega) - \overline{E}[X_i]) \geqslant 1/m\right\}.$$

Using countable additivity, $P(\Lambda) = \inf_{m\geqslant 1}\inf_{N\geqslant 1} P(\cup_{n\geqslant N}\Lambda_{m,n})$ for every P. And using the previous inequality, for every m there is some $N^* \geqslant 1$ such that

$$\inf_{m\geqslant 1}\inf_{N\geqslant 1} P(\cup_{n\geqslant N}\Lambda_{m,n}) \leqslant \inf_{m\geqslant 1} P(\cup_{n\geqslant N^*}\Lambda_{m,n}) \leqslant \inf_{m\geqslant 1} 2/m = 0;$$

consequently, $P(\Lambda) = 0$ for any P of interest, as desired.

The last expression in the theorem is proved from

$$P\left(\forall n \geqslant N : \frac{\sum_{i=1}^n X_i}{n} > \underline{\mu}_n - \epsilon\right) > 1 - 2\epsilon,$$

by a similar argument. □

Appendix H. Proof of Theorem 5

Proof. Using both Markov's inequality (as in the proof of Theorem 1) and disintegrability, for any $s > 0$ we get

$$\overline{P}\left(\sum_{i=1}^n (X_i - \overline{E}[X_i]) \geqslant \epsilon\right) \leqslant e^{-s\epsilon}\overline{E}\left[\exp\left(\sum_{i=1}^n s(X_i - \overline{E}[X_i])\right)\right] \leqslant e^{-s\epsilon}\overline{E}\left[\overline{E}\left[\exp\left(\sum_{i=1}^n s(X_i - \overline{E}[X_i])\right)|X_{1:n-1}\right]\right]$$
$$= e^{-s\epsilon}\overline{E}\left[\exp\left(\sum_{i=1}^{n-1} s(X_i - \overline{E}[X_i])\right)\overline{h}(X_{1:n-1})\right],$$

where $\overline{h}(X_{1:n-1}) = \overline{E}[\exp(s(X_n - \overline{E}[X_n]))|X_{1:n-1}]$. Due to condition (17),

$$\overline{E}[X_n|X_{1:n-1}] \leqslant \overline{E}[X_n]; \quad \text{thus } \overline{E}[X_n - \overline{E}[X_n]|X_{1:n-1}] \leqslant 0.$$

We now apply expression (23) (if variable X satisfies $a \leqslant X \leqslant b$ and $\overline{E}[X] \leqslant 0$, then $\overline{E}[\exp(sX)] \leqslant \exp(s^2(b-a)^2/8)$ for any $s > 0$), conditional on $X_{1:n-1}$:

$$\overline{h}(X_{1:n-1}) = \overline{E}[\exp(s(X_n - \overline{E}[X_n]))|X_{1:n-1}] \leqslant \exp(s^2\beta_n^2/8).$$

We have reached an analog of expression (20), and the proof of the theorem can be produced by copying the steps after that expression. □

Appendix I. Two auxiliary inequalities

The following inequality is a simple extension of a basic result by Hoeffding [8,17]: If variable X satisfies $a \leqslant X \leqslant b$ and $\overline{E}[X] \leqslant 0$, then for any $s > 0$,

$$\overline{E}[\exp(sX)] \leqslant \exp(s^2(b-a)^2/8). \tag{23}$$

Proof. First, the inequality is clearly valid if $a = b$ or if $b < 0$. From now on, suppose $b \geqslant 0 \geqslant a$. By convexity of the exponential function,

$$\exp(sx) \leqslant \frac{x-a}{b-a}e^{sb} + \frac{b-x}{b-a}e^{sa} \quad \text{for } x \in [a,b].$$

Given monotonicity of upper expectations,

$$\overline{E}[\exp(sX)] \leqslant \overline{E}\left[\frac{X-a}{b-a}e^{sb} + \frac{b-X}{b-a}e^{sa}\right].$$

Because $(e^{sb} - e^{sa})(b-a) > 0$, and using $\overline{E}[X] \leqslant 0$,

$$\overline{E}[\exp(sX)] \leqslant \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb} + \frac{e^{sb}-e^{sa}}{b-a}\overline{E}[X] \leqslant \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb}.$$

Now if $a = 0$, then $\overline{E}[\exp(sX)] \leqslant 1 \leqslant \exp(s^2b^2/8)$ and the theorem is valid. Thus suppose $b \geqslant 0 > a$. By rearranging terms, we obtain:

$$\overline{E}[\exp(sX)] \leqslant \exp(\phi(s(b-a))),$$

for $\phi(u) = -pu + \log(1 - p + pe^u)$ with $p = -a/(b-a)$ (note that $p \in (0,1]$). Given that $\phi(0) = \phi'(0) = 0$ and $\phi''(u) \leqslant 1/4$ for $u > 0$ (as the maximum of $\phi''(u)$ is 1/4, attained at $e^u = (1-p)/p$), we can use Taylor's theorem as follows. For some $v \in (0,u)$, $\phi(u) = \phi(0) + u\phi'(0) + (u^2/2)\phi''(v) \leqslant (u^2/8)$ and consequently $\phi(s(b-a)) \leqslant s^2(b-a)^2/8$. By putting together these inequalities, we obtain expression (23). □

We now review the Kolmogorov–Hajek–Renyi inequality, almost exactly as proved by Whittle [34, Thm. 14.2.2]; this is presented just to indicate the role of elementwise disintegrability in the derivation. Let $\{X_i\}$ be a martingale with $X_0 = 0$, and let $\{\epsilon_i\}$ be a sequence $0 < \epsilon_1 \leqslant \epsilon_2 \leqslant \cdots$; the inequality is

$$P(\forall j \in [1,n] : |X_j| < \epsilon_j) \geqslant 1 - \sum_{i=1}^{n} \frac{E[(X_i - X_{i-1})^2]}{\epsilon_i^2}. \tag{24}$$

Proof. Define $\epsilon_0 \doteq \epsilon_1$ and $A_n \doteq \{\forall j \in [0,n] : |X_j| < \epsilon_j\}$. Using $\xi_i = X_i - X_{i-1}$, and denoting the indicator function of some events by the events themselves,

$$\begin{aligned}
P(A_n) &= E_P[A_n] = E_P[A_{n-1}I_{\{|X_n|<\epsilon_n\}}] \\
&\geqslant E_P[A_{n-1}(1 - X_n^2/\epsilon_n^2)] \quad (\text{as } I_{\{|X|<\epsilon\}} \geqslant 1 - X^2/\epsilon^2) \\
&= E_P[A_{n-1}(1 - (X_{n-1}^2 + \xi_n^2)/\epsilon_n^2)] \\
&\qquad (\text{by the martingale property, expression } (21)) \\
&\geqslant E_P[A_{n-2}(1 - X_{n-1}^2/\epsilon_{n-1}^2)] - E_P[\xi_n^2/\epsilon_n^2] \\
&\qquad (\text{as } \epsilon_{n-1} \leqslant \epsilon_n \text{ and } I_{\{|X|<\epsilon\}}(1 - X^2/\epsilon^2) \geqslant (1 - X^2/\epsilon^2)).
\end{aligned}$$

Iteration of the last inequality yields the result. Note that disintegrability for each P was used when applying the martingale property. □

It should be noted that the inequality proved by Whittle is slightly different: $P(\forall j \in [1,n] : |X_j| \leqslant \epsilon_j) \geqslant 1 - \sum_{i=1}^n E[(X_i - X_{i-1})^2]/\epsilon_i^2$ (under the same conditions). The proof only changes by replacing indicator functions ($I_{\{|X|<\epsilon\}}$ by $I_{\{|X|\leqslant\epsilon\}}$).

References

[1] F. Chung, L. Lu, Concentration inequalities and martingale inequalities: a survey, Internet Mathematics 3 (1) (2006) 79–127.
[2] F.G. Cozman, Concentration inequalities and laws of large numbers under epistemic irrelevance, in: International Symposium on Imprecise Probability: Theories and Applications, Durham, United Kingdom, 2009, pp. 109–118.
[3] G. De Cooman, F. Hermans, Imprecise probability trees: bridging two theories of imprecise probability, Artificial Intelligence 172 (2008) 1400–1427.
[4] G. De Cooman, E. Miranda, Weak and strong laws of large numbers for coherent lower previsions, Journal of Statistical Planning and Inference 138 (8) (2008) 2409–2432.
[5] G. De Cooman, E. Miranda, Forward irrelevance, Journal of Statistical Planning and Inference 139 (2009) 256–276.
[6] B. de Finetti, Theory of Probability, vols. 1–2, Wiley, New York, 1974.
[7] L. Devroye, Exponential inequalities in nonparametric estimation, in: G. Roussas (Ed.), Nonparametric Functional Estimation and Related Topics, NATO ASI Series, Kluwer Academic Publishers, Dordrecht, 1991, pp. 31–44.
[8] L. Devroye, L. Gyorfi, G. Lugosi, A Probabilistic Theory of Pattern Recognition, Springer-Verlag, New York, 1996.
[9] L.E. Dubins, On Lebesgue-like extensions of finitely additive measures, Annals of Probability 2 (3) (1974) 456–463.
[10] L.E. Dubins, Finitely additive conditional probability, conglomerability and disintegrations, Annals of Probability 3 (1) (1975) 89–99.
[11] L.E. Dubins, L.J. Savage, How to Gamble If You Must: Inequalities for Stochastic Processes, McGraw-Hill, New York, 1965.
[12] L.G. Epstein, M. Schneider, IID: independently and indistinguishably distributed, Journal of Economic Theory 113 (1) (2003) 32–50.
[13] N. Etemadi, An elementary proof of the strong law of large numbers, Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 55 (1981) 119–122.
[14] N. Geller, Some examples of the weak and strong laws of large numbers for averages of mutually independent random variables, The American Statistician 32 (1) (1978) 34–36.
[15] F.J. Giron, S. Rios, Quasi-Bayesian behaviour: a more realistic approach to decision making?, in: J.M. Bernardo, J.H. DeGroot, D.V. Lindley, A.F.M. Smith (Eds.), Bayesian Statistics, University Press, Valencia, Spain, 1980, pp. 17–38.
[16] D. Heath, W. Sudderth, On finitely additive priors, coherence, and extended admissibility, Annals of Mathematical Statistics 43 (1978) 2072–2077.
[17] W. Hoeffding, Probability inequalities for sums of bounded random variables, Journal of the American Statistical Association 58 (1963) 13–30.

[18] J.B. Kadane, M.J. Schervish, T. Seidenfeld, Reasoning to a foregone conclusion, Journal of the American Statistical Association 91 (435) (1996) 1228–1235.
[19] R.L. Karandikar, A general principle for limit theorems in finitely additive probability, Transactions of the American Mathematical Society 273 (2) (1982) 541–550.
[20] P. Krauss, Representation of conditional probability measures on Boolean algebras, Acta Mathematica Academiae Scientiarum Hungaricae 19 (3–4) (1968) 229–241.
[21] D.A. Lane, W.D. Sudderth, Coherent predictions are strategic, Annals of Statistics 13 (3) (1985) 1244–1248.
[22] I. Levi, The Enterprise of Knowledge, The MIT Press, Massachusetts, 1980.
[23] D. Pollard, A User's Guide to Measure Theoretic Probability, Cambridge University Press, 2002.
[24] M. Schervish, Theory of Statistics, Springer-Verlag, New York, 1995.
[25] M. Schervish, T. Seidenfeld, J.B. Kadane, On the equivalence of conglomerability and disintegrability for unbounded random variables, Unpublished manuscript, 2009.
[26] T. Seidenfeld, Remarks on the theory of conditional probability: some issues of finite versus countable additivity, in: Statistics – Philosophy, Recent History, and Relations to Science, Kluwer Academic, 2001, pp. 167–177.
[27] T. Seidenfeld, M.J. Schervish, J.B. Kadane, Improper regular conditional distributions, Annals of Probability 29 (4) (2001) 1612–1624.
[28] T. Seidenfeld, M.J. Schervish, J.B. Kadane, Coherent choice functions under uncertainty, in: International Symposium on Imprecise Probability: Theories and Applications, Prague, Czech Republic, 2007.
[29] G. Shafer, V. Vovk, Probability and Finance: It's Only a Game!, Wiley, 2001
[30] P. Walley, Coherent Lower (and Upper) Probabilities, Statistics Research Report 22, University of Warwick, Coventry, 1981.
[31] P. Walley, Statistical Reasoning with Imprecise Probabilities, Chapman and Hall, London, 1991.
[32] K. Weichselberger, The theory of interval-probability as a unifying concept for uncertainty, International Journal of Approximate Reasoning 24 (2–3) (2000) 149–170.
[33] K. Weichselberger, T. Augustin (assistant), A. Wallner (assistant), Elementare Grundbegriffe einer Allgemeineren Wahrscheinlichkeitsrechnung. I: Intervallwahrscheinlichkeit als Umfassendes Konzept, Physica-Verlag, Heidelberg, 2001.
[34] P. Whittle, Probability via Expectation, third ed., Springer-Verlag, 1992.