Lecture 1: Martingales in the foundations of statistics and probability

Vladimir Vovk

Centre for Reliable Machine Learning
Department of Computer Science
Royal Holloway, University of London

Mahalanobis Memorial Lectures 2020–21
15 March, 2021

Main points of this lecture

- This lecture is meant to be a gentle introduction into this series.
- Betting is a useful metaphor in statistics and probability.
- The role of martingales, in the form of test martingales, in online hypothesis testing.
- Their use in the foundations of probability.
- Betting as metaphor is a very old idea: some history.

Plan

1. Online hypothesis testing
2. Martingales in the foundations of probability
3. History

Betting

- The standard ways of testing statistical hypotheses are based on p-values or Bayes factors.
- Alternative: testing by betting.
- We start from a capital of 1 and are not allowed to borrow money; gamble against the null hypothesis. Our current capital: the degree to which the hypothesis has been falsified.

Glenn Shafer.
The language of betting as a strategy for statistical and scientific communication.
To appear as discussion paper in the *Journal of the Royal Statistical Society A*. Read in September 2020.

Batch vs online testing

- Traditional methods of testing statistical hypotheses: batch setting.
- Given a batch of data, we compute a measure of evidence against the null hypothesis (a p-value or Bayes factor).
- Alternative: the online setting, in which the items of data (observations) keep arriving sequentially.

How realistic is the batch setting?

- In science, we rarely have a single study if the phenomenon under study is really important.
- Usually there are follow-up studies, and we need to have a mechanism for combining the results of sequential studies.
- This will be a topic of the next lecture.

Test martingales

- The usual formalization of test martingales is that we first define a filtration (nested family of σ-algebras) in our probability space $(\Omega, \mathcal{F}, \mathbb{P})$: $\mathcal{F}_n$ is generated by the first n observations $Z_1, \ldots, Z_n$.
- A test martingale is an adapted sequence of nonnegative random variables X_n such that $X_0 = 1$ and $\mathbb{E}(X_n \mid \mathcal{F}_{n-1}) = X_{n-1}$ for all n, where $\mathbb{E}$ is the expectation w.r. to $\mathbb{P}$.
- Their use in online hypothesis testing: if X_n is big, there is a mismatch between $Z_1, \ldots, Z_n$ and $\mathbb{P}$; we are entitled to reject $\mathbb{P}$ as our null hypothesis. (Justification: Ville's inequality $\mathbb{P}(\sup_n X_n \geq 1/\epsilon) \leq \epsilon$.)
- Test martingales are easily mixable: a convex mixture of test martingales is again a test martingale.

Plan

Martingales in the foundations of probability and stochastic processes

- It is interesting that martingales do not have to be defined in terms of probability.
- We can reverse the process and start from martingales.
- Probability and expectation become derivative notions.
- Glenn Shafer and I called this approach game-theoretic probability (in a way it is dual to the standard measure-theoretic probability).

Advantages of game-theoretic probability

Some of the advantages of game-theoretic probability:

- Strategies for testing (=strategies for Sceptic, one of the players in our games). Theorems of probability are made constructive.
- Limited betting opportunities: the assumptions required for those results can be weakened.
- Strategies for Reality (another important player), also constructive.
- Strategies for Forecaster, also constructive.

Perfect-information games as an alternative foundation for probability

Example: Kolmogorov's strong law of large numbers (SLLN) as a typical limit theorem.

Forecasting protocol:

$\mathcal{K}_0 = 1$
FOR $n = 1, 2, \ldots$:
  Forecaster announces $m_n \in \mathbb{R}$ and $v_n \geq 0$
  Sceptic announces $M_n \in \mathbb{R}$ and $V_n \geq 0$
  Reality announces $y_n \in \mathbb{R}$
  $\mathcal{K}_n := \mathcal{K}_{n-1} + M_n(y_n - m_n) + V_n((y_n - m_n)^2 - v_n)$

Example: predicting temperature for tomorrow.

Player's goals

A strategy for Sceptic forces an event E if it guarantees both

- $\mathcal{K}_n \geq 0$ for all n
- either $\mathcal{K}_n \to \infty$ or E happens.

A strategy for Reality forces E if it defeats Sceptic in the sense of

- $\mathcal{K}_n < 0$ for some n, or
- $\mathcal{K}_n$ is bounded and E happens.

Game-theoretic version of Kolmogorov's SLLN

Theorem

Sceptic can force

$$\sum_{n=1}^{\infty} \frac{v_n}{n^2} < \infty \Longrightarrow \frac{1}{n} \sum_{i=1}^{n} (y_i - m_i) \to 0.$$

Reality can force

$$\sum_{n=1}^{\infty} \frac{v_n}{n^2} = \infty \Longrightarrow \frac{1}{n} \sum_{i=1}^{n} (y_i - m_i) \not\to 0.$$

The strategies constructed in the proofs are explicit (and computable; in particular measurable).

SLLN for bounded variables

- If we bound Reality's moves, $y_n \in [A, B]$, we can drop v_n and still claim that Sceptic can force

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - m_i) \to 0.$$

- A standard interpretation for a trusted Forecaster: we do not expect Sceptic to become infinitely rich, and so expect $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}(y_i - m_i) = 0$ (calibration).
- Curious example: predictions in a prediction market either allow us to become infinitely rich or are calibrated. Which?

Connection with measure-theoretic probability

Kolmogorov's SLLN is an easy corollary of the game-theoretic version (in combination with the measurability of Sceptic's strategy): if $\xi_1, \xi_2, \ldots$ are independent random variables with expected values $\mathbb{E}(\xi_1), \mathbb{E}(\xi_2), \ldots$ and variances $\mathrm{var}(\xi_1), \mathrm{var}(\xi_2), \ldots$,

$$\sum_{n=1}^{\infty} \frac{\mathrm{var}(\xi_n)}{n^2} < \infty \Longrightarrow \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} (\xi_i - E(\xi_i)) = 0 \quad \text{a.s.}$$

Proof: If Reality follows the randomized strategy $y_n := \xi_n$, Forecaster always chooses $m_n := \mathbb{E}(\xi_n)$ and $v_n := \mathrm{var}(\xi_n)$, and Sceptic follows his winning strategy, $\mathcal{K}_n$ will be a nonnegative martingale. According to a standard result, $\mathcal{K}_n \to \infty$ with probability 0.

Strategies for Reality

- They do not have any analogues in the existing measure-theoretic probability.
- And they are often extremely simple; e.g., the one forcing Kolmogorov's SLLN is:
  - If $\mathcal{K}_{n-1} + V_n(n^2 - v_n) > 1$, set $y_n := 0$.
  - If $\mathcal{K}_{n-1} + V_n(n^2 - v_n) \leq 1$ and $M_n \leq 0$, set $y_n := n$.
  - If $\mathcal{K}_{n-1} + V_n(n^2 - v_n) \leq 1$ and $M_n > 0$, set $y_n := -n$.
- Takemura and Miyabe obtained lots of beautiful results about forcing by Reality.

Strategies for Forecaster

- This is known as defensive forecasting.
- Under weak assumptions, testing strategies can be turned into successful prediction strategies.
- Idea: Forecaster can defend Reality against a known strategy for Sceptic, making sure $\mathcal{K}_n$ does not grow.
- If Sceptic is testing calibration, Forecaster can enforce calibration.

Game-theoretic SLLN for binary observations

Binary forecasting protocol:

$\mathcal{K}_0 := 1$
FOR $n = 1, 2, \ldots$:
  Forecaster announces $p_n \in [0, 1]$
  Sceptic announces $s_n \in \mathbb{R}$
  Reality announces $y_n \in \{0, 1\}$
  $\mathcal{K}_n := \mathcal{K}_{n-1} + s_n(y_n - p_n)$

Proposition (game-theoretic SLLN)

As we know:

Sceptic has a strategy which guarantees that

- $\mathcal{K}_n$ is never negative
- either

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (y_i - p_i) = 0$$

or

$$\lim_{n\to\infty} \mathcal{K}_n = \infty.$$

Caveat

- Remember that the measure-theoretic SLLN follows easily.
- Reality need not be oblivious (or even follow a strategy).
- Forecaster need not ignore Sceptic (this is what makes defensive forecasting possible).
- Caveat: we assumed that Sceptic's strategy was measurable. Fact of life: for all kinds of limit theorems, Sceptic's strategy we construct is measurable; moreover, it is continuous.

Observation

- Interesting observation (first made by Foster and Vohra in a different context): the game-theoretic approach to probability can be used for designing learning algorithms.
- For any continuous strategy for Sceptic there exists a strategy for Forecaster that does not allow Sceptic's capital to grow.

Modified protocol

$\mathcal{K}_0 := 1$
FOR $n = 1, 2, \dots$:
  Sceptic announces continuous $S_n : [0,1] \to \mathbb{R}$
  Forecaster announces $p_n \in [0,1]$
  Reality announces $y_n \in \{0,1\}$
  $\mathcal{K}_n := \mathcal{K}_{n-1} + S_n(p_n)(y_n - p_n)$

Theorem (Takemura) Forecaster has a strategy that ensures $\mathcal{K}_0 \ge \mathcal{K}_1 \ge \mathcal{K}_2 \cdots$.

Proof

- choose p_n so that $S_n(p_n) = 0$
- if the equation $S_n(p) = 0$ has no roots (in which case S_n never changes sign),

$$p_n := \begin{cases} 1 & \text{if } S_n > 0 \\ 0 & \text{if } S_n < 0 \end{cases}$$

QED
Has been greatly extended; Intermediate Value Theorem $\mapsto$ Ky Fan's fixed point theorem.

Research programme (forecasting)

- Open a probability textbook and decide which property (such as LLN, CLT, LIL, Hoeffding's inequality,. . .) you want Forecaster's moves to satisfy.
- Prove the corresponding game-theoretic result.
- Apply Takemura's theorem.

What does it give in the case of SLLN?

In this special case, nothing interesting: Forecaster performs his task too well. E.g., he can choose

$$p_n := \begin{cases} 1/2 & \text{if } n = 1 \\ y_{n-1} & \text{otherwise,} \end{cases}$$

ensuring

$$\left| \sum_{i=1}^{n} (y_i - p_i) \right| \leq 1/2$$

for all n (much better than using the true probabilities).

Forecasting without statistical assumptions

- But the law of probability can be much more demanding!
- There are areas of machine learning (such as prediction with expert advice) where people design prediction algorithms that have various performance guarantees without making any statistical assumptions about the observations.
- Defensive forecasting is a way of obtaining such results.

Game-theoretic probability 1

Sceptic's part of the game-theoretic SLLN can be restated as:

$$\sum_{n=1}^{\infty} \frac{v_n}{n^2} < \infty \Longrightarrow \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (x_i - m_i) = 0 \qquad (1)$$

holds with lower probability one, for a new notion of probability.

Or: the complement of (1) holds with upper probability zero.

Game-theoretic probability 2

For an event E,

$$\overline{\mathbb{P}}(E) := \inf\{\epsilon : \exists \text{ allowed strategy for Sceptic such that } \sup_n \mathcal{K}_n \geq 1/\epsilon \text{ on the event } E\}$$

(upper game-theoretic probability; equivalent definition with $\limsup$ or $\liminf$ in place of $\sup$) and

$$\underline{\mathbb{P}}(E) := 1 - \overline{\mathbb{P}}(E^c)$$

(lower game-theoretic probability).

For many interesting sets E, $\overline{\mathbb{P}}(E) = \underline{\mathbb{P}}(E)$ (in continuous time) and $\overline{\mathbb{P}}(E) \approx \underline{\mathbb{P}}(E)$ (in discrete time).

Other game-theoretic results

For example:

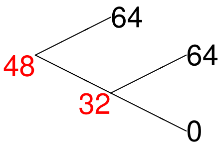
- law of the iterated logarithm [lower probability one]
- zero-one law [lower probability one]
- central limit theorem [general game-theoretic probability]
- Itô calculus [lower probability one]
- ...

Imply their measure-theoretic counterparts.

Plan

The problem of points and Pascal's solution

- The two approaches to probability go back to Pascal and Fermat and earlier!
- Two gamblers play three throws; each puts 32 pistoles at stake. The first has two (points) and the other one, and they have to stop the game. How should they divide the 64 pistoles?
- Pascal to Fermat on 29 July, 1654:



This is a martingale.

Fermat's solution

Imagine (some of Pascal's and Fermat's contemporaries protested!) that the two gamblers make two more throws. The first's expected win is

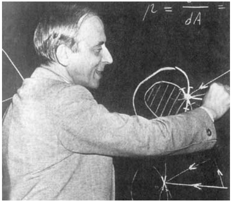
$$64 \times \frac{3}{4} + 0 \times \frac{1}{4} = 48.$$

The same answer but the methods are different: Pascal's can be regarded as a precursor of game-theoretic probability, and Fermat's as a precursor of measure-theoretic probability (in this case, the measure assigned to each pair of outcomes is $1/4$).

Game-theoretic vs measure-theoretic foundations of probability

Hilbert's sixth problem: to treat axiomatically, after the model of geometry, those parts of physics in which mathematics already played an outstanding role, especially probability and mechanics.

- Richard von Mises (1919, 1928, 1931): probability is a derivative notion, based on the notion of a gambling system.
- Andrei Kolmogorov (1931, 1933): probability is axiomatized directly as a special case of measure.



von Mises (1883–1957)

Kolmogorov (1903–1987)

Limitations of von Mises's concept of gambling

The claim that von Mises's gambling systems provide a satisfactory foundation for probability was refuted by Jean Ville (1939): they are not sufficient to derive the law of the iterated logarithm.

Von Mises's gambling systems choose a subsequence of trials on which to bet. Ville: more sophisticated gambling systems which also vary the amount of the bet and the outcome on which to bet. He called the capital processes of such strategies martingales.

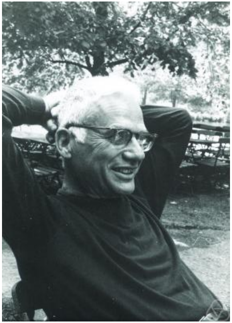
Jean Ville (1910–1988)

Further developments

Joseph Doob reviewed Ville's 1939 book about martingales for Mathematical Reviews. He then translated the notion of martingale into measure-theoretic probability and developed it greatly.

Now the measure-theoretic theory of martingales is the centrepiece of many areas of probability theory. A prominent role in developing the theory of stochastic processes based on martingales was played by Kiyosi Itô (his integration and calculus).

Martingales flourish in the mathematics (but not the foundations or philosophy) of probability.



Doob (1910–2004)

Itô (1915–2008)

Reviving betting in the foundations of probability

Kolmogorov writing in India in 1962 during his visit to Mahalanobis and Indian Statistical Institute:

> *The set theoretic axioms of the calculus of probability, in formulating which I had the opportunity of playing some part (Kolmogorov, 1950), had solved the majority of formal difficulties in the construction of a mathematical apparatus which is useful for a very large number of applications of probabilistic methods, so successfully that the problem of finding the basis of real applications of the results of the mathematical theory of probability became rather secondary to many investigators.*

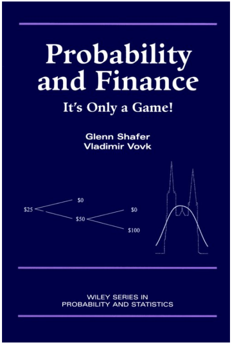
(*Sankhyā*, 1963).

Developments of Kolmogorov’s programme

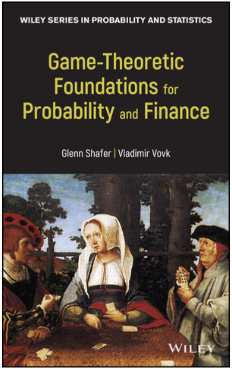
- Kolmogorov saw the “basis of real applications of the results of the mathematical theory of probability” in von Mises’s ideas.
- For that, he introduced the notion of Kolmogorov complexity, already mentioned in his 1963 *Sankhyā* paper.
- His followers (Leonid Levin and Claus Schnorr) made a step similar to Ville’s and replaced von Mises-type betting schemes by martingales.

Moving away from Kolmogorov's axioms

- A. Philip Dawid: prequential (predictive sequential) statistics. Calls for evaluating a probability forecaster only using his actual forecasts. (He might not even have a strategy.)
- Sits uneasily with Kolmogorov's axioms of probability.
- Glenn Shafer's and my books: systematization.



Our 2001 book



Our 2019 book

Bibliography

Glenn Shafer and Vladimir Vovk.
Game-Theoretic Foundations for Probability and Finance
Hoboken, NJ: Wiley, 2019.

A growing list of working papers (now $58 + 20$):

`http://www.probabilityandfinance.com`

Thank you for your attention!