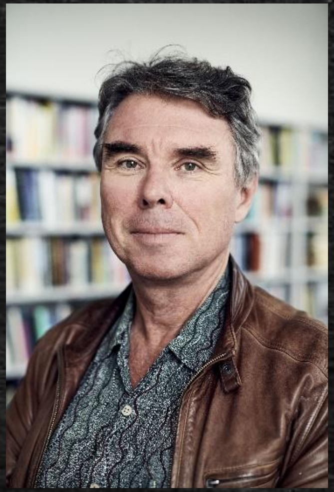
Towards a Strictly Frequentist Theory of Imprecise Probability

Christian Fröhlich, Rabanus Derr, Robert C. Williamson

The “Foundations of Machine Learning Systems” Group

http://fm.ls University of Tübingen; Tübingen AI Center



Led by Robert C. Williamson

Core research interest: better models for data in ML

Data is Imperfect

- Corruptions
- Missing data
- Dataset shift
- Data about people
  - Ethical concerns such as fairness!
  - Society in non-equilibrium

Core research interest:
better models for data in ML

Data is Imperfect

- Corruptions
- Missing data
- Dataset shift
- Data about people
  - Ethical concerns such as fairness!
  - Society in non-equilibrium

The standard **i.i.d.** assumption

- The utopia of *independent and identically distributed* data
- Just one representative example:

> We consider supervised learning with a *risk-averse learner*. The learner has a data set comprised of i.i.d. samples from an unknown distribution, i.e., $D = \{(x_1, y_1), \dots (x_N, y_N)\} \in (\mathcal{X} \times \mathcal{Y})^N \sim \mathcal{D}^N$, and her goal is to learn a function $h_\theta : \mathcal{X} \to \mathcal{R}$ that is parametrized by $\theta \in \Theta \subset \mathbb{R}^d$. The

[Curi et al., 2020]

The conceptual culprit:
The law of large numbers

The conceptual culprit: The law of large numbers

- Let $X_1, X_2, ..$ be i.i.d. random variables with finite expectation $E[X]$, then the sample average converges almost surely to the expectation

$$P\left(\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}X_i = E[X]\right) = 1$$

The conceptual culprit: The law of large numbers

- Let $X_1, X_2, ..$ be i.i.d. random variables with finite expectation $E[X]$, then the sample average converges almost surely to the expectation

$$P\left(\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}X_i = E[X]\right) = 1$$

- Theoretical criticism: holds on a set of sequences with Lebesgue measure 1..not for *all* sequences.
  - Different ways of measuring the size of this set yield different answers!

The conceptual culprit: The law of large numbers

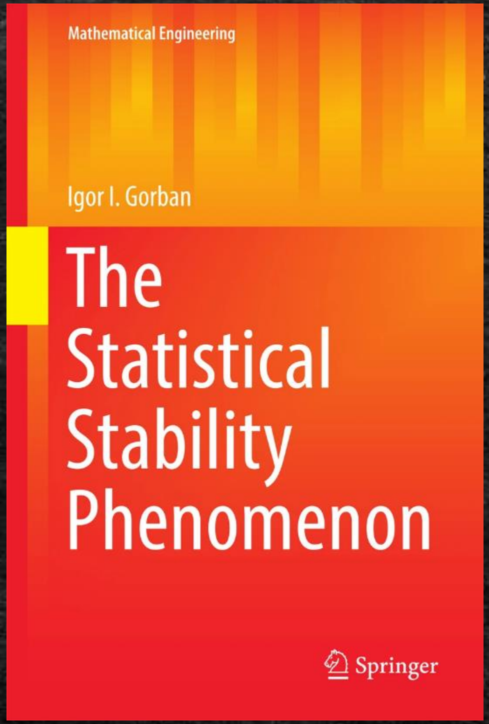
- Let $X_1, X_2, ..$ be i.i.d. random variables with finite expectation $E[X]$, then the sample average converges almost surely to the expectation

$$P\left(\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}X_i = E[X]\right) = 1$$

- Theoretical criticism: holds on a set of sequences with Lebesgue measure 1..not for *all* sequences.
  - Different ways of measuring the size of this set yield different answers!
- Serves to justify the *hypothesis of statistical stability* [Gorban, 2017]: that relative frequencies tend to stabilize in the long run

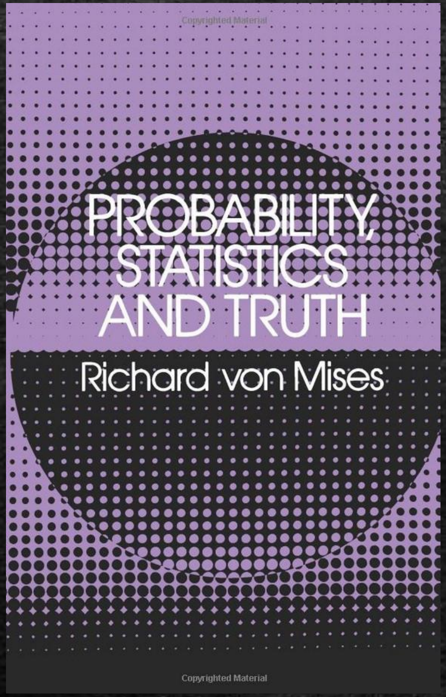
Challenges to statistical stability

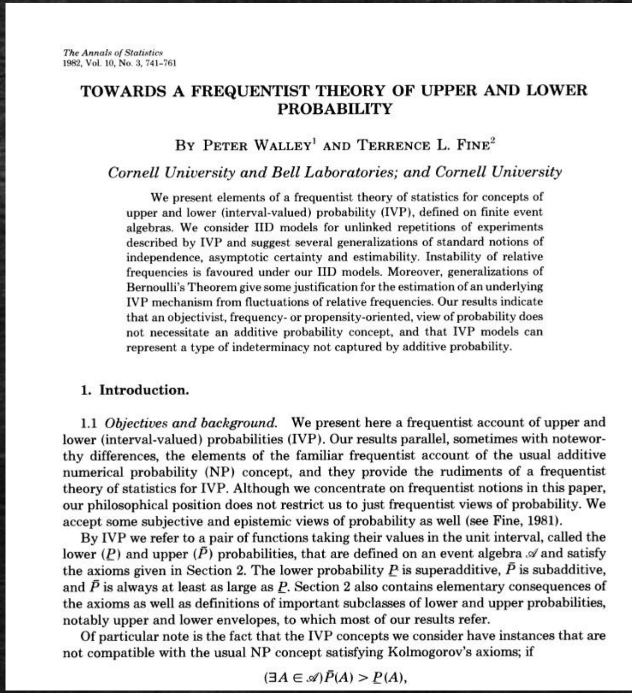
- From machine learning: dataset shift etc.
- Many interesting examples for instability [Gorban, 2017]
  - Mains voltage in a city
  - Earth's magnetic field
  - …



Our aim:

Find a principled way to model statistically unstable phenomena





The Annals of Statistics
1982, Vol. 10, No. 3, 741–761

TOWARDS A FREQUENTIST THEORY OF UPPER AND LOWER PROBABILITY

By Peter Walley[1] and Terrence L. Fine[2]

Cornell University and Bell Laboratories; and Cornell University

We present elements of a frequentist theory of statistics for concepts of upper and lower (interval-valued) probability (IVP), defined on finite event algebras. We consider IID models for unlinked repetitions of experiments described by IVP and suggest several generalizations of standard notions of independence, asymptotic certainty and estimability. Instability of relative frequencies is favoured under our IID models. Moreover, generalizations of Bernoulli's Theorem give some justification for the estimation of an underlying IVP mechanism from fluctuations of relative frequencies. Our results indicate that an objectivist, frequency- or propensity-oriented, view of probability does not necessitate an additive probability concept, and that IVP models can represent a type of indeterminacy not captured by additive probability.

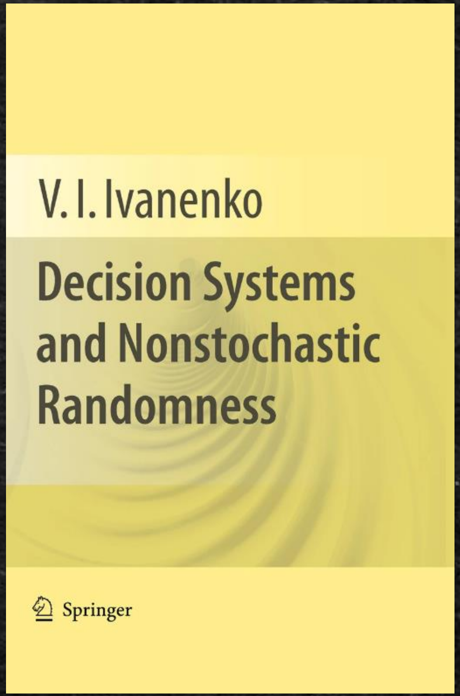
1. Introduction.

1.1 *Objectives and background.* We present here a frequentist account of upper and lower (interval-valued) probabilities (IVP). Our results parallel, sometimes with noteworthy differences, the elements of the familiar frequentist account of the usual additive numerical probability (NP) concept, and they provide the rudiments of a frequentist theory of statistics for IVP. Although we concentrate on frequentist notions in this paper, our philosophical position does not restrict us to just frequentist views of probability. We accept some subjective and epistemic views of probability as well (see Fine, 1981).

By IVP we refer to a pair of functions taking their values in the unit interval, called the lower ($\underline{P}$) and upper ($\bar{P}$) probabilities, that are defined on an event algebra $\mathscr{A}$ and satisfy the axioms given in Section 2. The lower probability $\underline{P}$ is superadditive, $\bar{P}$ is subadditive, and $\bar{P}$ is always at least as large as $\underline{P}$. Section 2 also contains elementary consequences of the axioms as well as definitions of important subclasses of lower and upper probabilities, notably upper and lower envelopes, to which most of our results refer.

Of particular note is the fact that the IVP concepts we consider have instances that are not compatible with the usual NP concept satisfying Kolmogorov's axioms; if

$$(\exists A \in \mathscr{A})\bar{P}(A) > \underline{P}(A),$$



How:

Tying together threads from von Mises, Walley & Fine, and Ivanenko

A strictly frequentist approach

A strictly frequentist approach

- Take the notion of a sequence as the primitive
  - In the spirit of von Mises and Ivanenko
  - Start with the data!

A strictly frequentist approach

- Take the notion of a sequence as the primitive
  - In the spirit of von Mises and Ivanenko
  - Start with the data!
- Assume an arbitrary possibility set Ω and $L^\infty = \{X\colon \Omega \to \mathbb{R}\colon \sup_{\omega \in \Omega} |X(\omega)| < \infty\}$

A strictly frequentist approach

- Take the notion of a sequence as the primitive
  - In the spirit of von Mises and Ivanenko
  - Start with the data!
- Assume an arbitrary possibility set Ω and $L^\infty = \{X: \Omega \to \mathbb{R}: \sup_{\omega \in \Omega} |X(\omega)| < \infty\}$
- Assume a sequence $\vec{\Omega}: \mathbb{N} \to \Omega$ of elementary outcomes

A strictly frequentist approach

- Take the notion of a sequence as the primitive
  - In the spirit of von Mises and Ivanenko
  - Start with the data!
- Assume an arbitrary possibility set Ω and $L^\infty = \{X\colon \Omega \to \mathbb{R}\colon \sup_{\omega\in\Omega}|X(\omega)| < \infty\}$
- Assume a sequence $\Omega^\rightarrow\colon \mathbb{N} \to \Omega$ of elementary outcomes
- For any event $A \subseteq \Omega$, consider the limiting relative frequency:

$$P(A) \coloneqq \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \chi_A\left(\Omega^\rightarrow(i)\right)$$

But the limit may or may not exist! Relative frequencies could *diverge*.

Cluster points -> coherent upper prevision

Cluster points -> coherent upper prevision

- Define a sequence of linear previsions as:

$$E^{\rightarrow}(n) \coloneqq X \mapsto \frac{1}{n}\sum_{i=1}^{n} X\left(\Omega^{\rightarrow}(i)\right), \quad X \in L^{\infty}$$

Cluster points -> coherent upper prevision

- Define a sequence of linear previsions as:

$$E^{\rightarrow}(n) \coloneqq X \mapsto \frac{1}{n}\sum_{i=1}^{n} X\big(\Omega^{\rightarrow}(i)\big), \quad X \in L^{\infty}$$

- From this, induce coherent upper prevision:

$$\bar{R}(X) \coloneqq \sup\{E(X) : E \in \mathrm{CP}(E^{\rightarrow})\}, \quad X \in L^{\infty}$$

  - Where $\mathrm{CP}(\cdot)$ is the set of cluster points with respect to the weak* topology on the dual of L^{∞}
  - This set is guaranteed to be non-empty!

Cluster points -> coherent upper prevision

- Define a sequence of linear previsions as:

$$E^{\to}(n) \coloneqq X \mapsto \frac{1}{n}\sum_{i=1}^{n} X\big(\Omega^{\to}(i)\big), \quad X \in L^{\infty}$$

- From this, induce coherent upper prevision:

$$\bar{R}(X) \coloneqq \sup\{E(X): E \in \mathrm{CP}(E^{\to})\}, \quad X \in L^{\infty}$$

  - Where $\mathrm{CP}(\cdot)$ is the set of cluster points with respect to the weak* topology on the dual of L^{∞}
  - This set is guaranteed to be non-empty!

- Proposition: $\bar{R}(X) = \limsup_{n \to \infty} \frac{1}{n}\sum_{i=1}^{n} X\big(\Omega^{\to}(i)\big)$

Limit *is to* Expectation
as
Cluster Points *to* Upper Prevision

Take-away:

Even if "the" probability does not exist, we always have a coherent upper prevision.

Also: can recover the generalized Bayes rule and formulate independence.

The converse direction

- **Theorem:** for finite Ω and arbitrary coherent upper prevision $\overline{R}$ on Ω, can always construct a corresponding sequence $\Omega^{\rightarrow}$(such that $\Omega^{\rightarrow}$ induces $\overline{R}$)
- Proof is constructive!

The converse direction

- **Theorem:** for finite Ω and arbitrary coherent upper prevision $\bar{R}$ on Ω, can always construct a corresponding sequence $\Omega^{\rightarrow}$(such that $\Omega^{\rightarrow}$ induces $\bar{R}$)
- Proof is constructive!

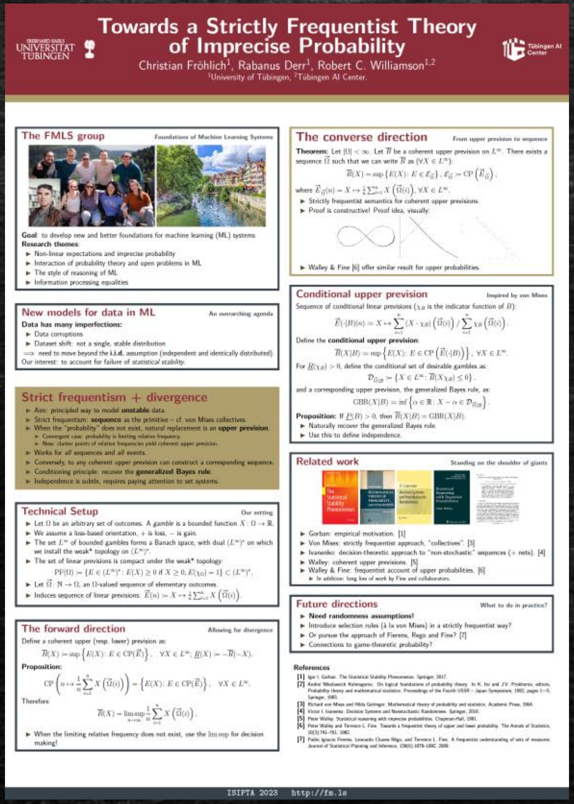
=> Strictly frequentist re-interpretation of subjective upper previsions.

Future Directions – What to do in practice?

The finite data setting

- Need randomness assumptions
- Connections to game-theoretic randomness?
- How is this related to work of Fierens, Rêgo and Fine?
  - "A frequentist understanding of sets of measures"

Come visit our poster ☺



Happy to chat about machine learning and IP