Mathematisches Forschungsinstitut Oberwolfach

Report No. 24/2024

DOI: 10.4171/OWR/2024/24

Game-theoretic Statistical Inference: Optional Sampling, Universal Inference, and Multiple Testing Based on E-values

Organized by
Peter Grünwald, Amsterdam/Leiden
Aaditya Ramdas, Pittsburgh
Ruodu Wang, Waterloo
Johanna Ziegel, Zürich

5 May – 10 May 2024

Abstract. This half-size MFO workshop brings together researchers in mathematical statistics, probability theory, machine learning, medical sciences, and economics to discuss recent developments in sequential inference. New sequential inference methods that build on nonnegative martingale techniques allow us to elegantly solve prominent shortcomings of traditional statistical hypothesis tests. Instead of p-values, they are based on *e-values* which have the added benefit that their meaning is much easier to communicate to applied researchers, due to their intuitive interpretation in terms of the wealth of a gambler playing a hypothetically fair game. Significant new contributions to this fast growing research area will be presented in order to stimulate collaborations, discuss and unify notation and concepts in the fields, and tackle a variety of open problems and address current major challenges.

Mathematics Subject Classification (2020): 62L99.

Introduction by the Organizers

The workshop *Game-theoretic statistical inference*, organized by Peter Grünwald, Aaditya Ramdas, Ruodu Wang and Johanna Ziegel, was well attended and involved 23 on-site participants. The participants show a broad representation of diversity in research areas, geographic locations, ethnic and gender groups, and career stages.

The research area of e-values and game-theoretic statistical inference is currently at a very exciting stage: the first breakthrough papers appeared around 5 years ago, an initial amount of consolidation has taken place, and now, at a highly rapid rate, new results are being derived. As such, all the talks were exciting — they were all full of recent results, conjectures and ideas for novel research. This gathering brought together a highly representative sample of the most active researchers in the field, as well as some researchers working in adjacent-yet-closely-related fields. Most participants gave a talk — some talks were 30, others 45-minute length. As such we had 22 talks in total, providing a kaleidoscope of current research on e-values and their use in anytime-valid inference, decision theory and multiple testing, and indicating a plethora of open questions and possibilities for follow-up work.

The first day was reserved for, among others, what we expected to be the most 'mind-boggling' talks, namely those by Rafael Frongillo and Martin Larsson (see below). Otherwise the talks were given in no particular order. Still, one could identify several specific recurring sub-topics:

Game-Theoretic Probability and Statistics. Rafael Frongillo is in the process of writing a monograph on game-theoretic probability, connecting the foundational work by Vladimir Vovk and Glenn Shafer (both present at the workshop) to minimax theorems and minimax regret results in online learning theory. Frongillo's talk, sketching the general take of his book, provided substantial clarification on how these different concepts (game-theoretic and measure-theoretic probability, replicating prices, minimax theorems) relate. Later on, Wouter Koolen gave a talk in which he very clearly showed how supermartingales — stochastic objects — occur naturally in non-stochastic, worst-case online learning algorithms via *defensive forecasting*, which also stems from work of Vovk, Shafer, and Akimichi Takemura. Interestingly, Frongillo's and Koolen's talk did not *quite* paint the same picture, thus supplying much material for further discussion and thought.

Rather than presenting novel mathematics, pioneer Shafer gave an entertaining talk on how to *teach* game-theoretic statistics. And pioneer Vovk talked about linking game-theoretic probability with multiple testing, thereby bridging two of the prominent sub-topics in this workshop.

E-variables and Multiple Testing. This topic drew a lot of attention. Apart from Vovk, Rina Barber talked about connecting knock-offs to e-values, and Zhimei Ren and Nikos Ignatiadis both gave exceptionally clear talks on how e-values can be used generally in multiple testing, and how their power can be improved in some cases, again connecting two sub-areas:

Improving Power obtained with E-Tests. Apart from Ren and Ignatiadis, also Ruodu Wang talked explicitly about this subject, which is important from both practical and theoretical perspectives. Also Thorsten Dickhaus was implicitly about this — he presented what was essentially work in progress on modelling a series of 3×2 contingency tables with e-values, and after discussion with other participants it turned out that standard solutions to this problem such as those by Turner, Ly and Grünwald (2024) and universal inference do not provide satisfying answers. A new collaboration with Rianne De Heide and Peter Grünwald on this topic has already been initiated.

Information Projections. On the first day, Martin Larsson gave a talk about one of the most exciting novel developments: the *numeraire e-variable*, a far-reaching and unifying treatment of the idea of the reverse information projection and its use in constructing the growth-optimal e-variable that is central to the field. Some open problems (when can the numeraire e-variable be arrived at by a simple KL projection and when not?) were also discussed. Grünwald's talk about a theory for constructing information projections and e-values for exponential families also fell into this category. Wang, in the second part of his talk, considered axiomatizations of desirable e-value criteria that characterized growth optimality, and hence are intimately related to information projections.

Extensions of Ville's Inequality. Muriel Pérez' talk on Ville's inequality for potentially negative martingales as well as Johannes Ruf's talk on the proper way to define Villean theorems for composite nulls both fell in this category. Also, the part of Aaditya Ramdas' talk about anytime-valid matrix inequalities dealt with variations of this question. It is amazing to see how a result of 1939 has been revived and significantly extended in recent years!

Specific Applications and Models. Johanna Ziegel, Hongyan Shi, Parnian Kassraie, Timo Dimitriadis, Shubhhada Agrawal and Michael Lindon all talked about specific applications of e-values and anytime-valid inference to specific models, ranging from the highly parametric (regression with Gaussian noise) to nonparametric (learning a mean under moment constraints) and from the highly regular (regression) to the highly irregular (mixture models) with applications such as bandits, simple A/B-testing and forecasting. Together these talks gave a fine overview of the state-of-the-art in applying the theories to specific statistical models and situations.

Two talks that stood **on their own**, yet nevertheless were very well received, were those by Ryan Martin and Ian Waudby-Smith. Martin talked about connections between uncertainty quantification with e-values based on the *e-posterior* and Martin's own theory of *inferential models* based on *possibility contour functions*, a specific method to represent certain sets of probabilities. Waudby-Smith talked about the new ideas of asymptotic confidence sequences, distribution-uniform anytime-valid inference, and their applications to conditional independence testing.

Open Problem Discussions. On the first day, Koolen presented an open problem (or rather a conjecture) related to the *KLinf* function that is used a lot in nonparametric e-processes. Several attendees started working on it. The problem remains unsolved as yet, but some progress was made in the sense that it was shown that 'easy' solutions, which would be based on proving the conjecture via Markov's inequality on a suitably chosen random quantity, cannot work — if the conjecture is true at all it must be proven by other, considerably more complicated means. On the second day, Grünwald presented an open problem regarding whether the mixture achieving the minimum Kullback-Leibler divergence towards a point null distribution P_0 on a set $\mathcal{Q}$ of distributions separated from P_0 and with a convex complement, always has support concentrated on the boundary of $\mathcal{Q}$. Wang and co-authors have recently shown a generalization of Strassen's theorem, which, in contrast to the classical theorem, only holds in one dimension, and it is open how to achieve the same result in higher dimensions. This was posted as an open question to the participants of the workshop.

General Impression. We felt that the atmosphere during the workshop was friendly and highly inspiring. Thursday night there was an improvised musical concert; there were groups of people working together and chatting until late at night; and most of the participants joined the traditional Wednesday Schwarzwälder Kirschtortenhike. We have fond memories!

Acknowledgement: The MFO and the workshop organizers would like to thank the National Science Foundation for supporting the participation of junior researchers in the workshop by the grant DMS-2230648, "US Junior Oberwolfach Fellows".

Workshop: Game-theoretic Statistical Inference: Optional Sampling, Universal Inference, and Multiple Testing Based on E-values

Table of Contents

Abstracts

Minimax Duality in Game-Theoretic Probability and Statistics

Rafael Frongillo

Suppose Ada and Charles, circa 1850, encounter a device purporting to simulate a fair coin. Ada, skeptical, proposes a sequential bet: on round t she can gamble any amount $\beta_t \in \mathbb{R}$, after which they observe the output $y_t \in \{-1, 1\}$ of the device, and Charles pays Ada $\beta_t y_t$. Here $y_t = 1$ might represent heads, and $y_t = -1$ tails, so that $\beta_t > 0$ represents a bet on heads and $\beta_t < 0$ on tails.

Suppose Ada is correct and device fails to faithfully simulate a fair coin. If for example it has a bias toward heads, Ada can start with \$1, repeatedly bet a small fraction of her wealth on heads, and become infinitely rich. Indeed, if it has *any* sort of bias in a suitable sense, she can do something similar. Moreover, beyond simply getting rich, Ada winning a lot of money from \$1 without risking bankruptcy amounts to *evidence* that, whatever this device is doing, it is not simulating a fair coin. In particular, this evidence would remain valid even without a stochastic assumption about the device, i.e., that it is random in any traditional sense.

Now suppose Ada's friend Mary, observing the exchange between Ada and Charles, offers a deal: Mary will pay Ada \$3 now, but if the next three "flips" of the device are heads, Ada must pay Mary \$16. Should Ada accept? If the device were faithfully simulating a fair coin, Ada may well accept as the expected value of the net deal is $3 - (\frac{1}{2})^3 16 = 1$. But of course Ada is skeptical, and may be hesitant to place any assumption whatsoever on the workings of the device.

Nonetheless, Ada would accept Mary's deal. Starting with \$2 of the \$3 Mary pays her, Ada can place three all-or-nothing bets on heads, so that no matter the outcome she makes \$1! In this case we say that Ada can *replicate* the contingent security

$$X = \begin{cases} \$16 & \text{3 heads} \\ \$0 & \text{otherwise} \end{cases}$$

for \$2. In other words, \$2 is the lowest price Ada would sell X for, if she is not willing to place any assumptions on the device.

The above ideas, of evidence and replication even in the absence of stochastic assumptions, underly the field of game-theoretic probability/statistics as detailed by Shafer and Vovk [1, 2]. A convenient framework to make these ideas rigorous is via *gamble spaces*, pairs $(\Omega, \mathcal{Z})$ where Ω is a set of outcomes and $\mathcal{Z} \subseteq (\Omega \to \mathbb{R} \cup \{\infty\})$ a set of available gambles. For example, Ada's gambles on the fair coin device were each on the gamble space with $\Omega = \{-1, 1\}$ and $\mathcal{Z} = \{\omega \mapsto \beta\omega \mid \beta \in \mathbb{R}\}$.

We may define the *game-theoretic upper expectation* $\overline{\mathbf{E}}X$ of a variable $X : \Omega \to \mathbb{R}$ by its cost to replicate using gambles from $\mathcal{Z}$.

$$\overline{\mathbf{E}}X := \inf\{\alpha \in \mathbb{R} \mid \exists Z \in \mathcal{Z} \text{ s.t. } Z + \alpha \geqslant X\} \ . \qquad \text{(replication cost)}$$

We can in turn rephrase this upper expectation as a zero-sum game between two players, Gambler who chooses $Z \in \mathcal{Z}$, and Nature (or Reality) who chooses $\omega \in \Omega$,

$$= \inf_{Z\in\mathcal{Z}} \sup_{\omega\in\Omega} X(\omega) - Z(\omega) \;. \qquad \text{(zero-sum game)}$$

One can further define sequential versions of gamble spaces and upper expectations, matching the iterative nature of Ada's gambles. Defining the conditional game-theoretic upper expectation as the cost to replicate a variable on the remaining rounds gives rise to a game-theoretic notion of supermartingales.

How do these game-theoretic expectations and supermartingales relate to their measure-theoretic counterparts? It turns out that, under certain conditions, the two coincide. Specifically, if *minimax duality* holds, we can write

$$\overline{\mathsf{E}}X = \inf_{Z\in\mathcal{Z}} \sup_{\omega\in\Omega} X(\omega) - Z(\omega) = \sup_{\mu\in\Delta(\Omega)} \inf_{Z\in\mathcal{Z}} \mathsf{E}_\mu[X - Z] \;. \tag{1}$$

If we further assume that the gambles $\mathcal{Z}$ are *scalable*, meaning $Z \in \mathcal{Z}, c \geqslant 0 \implies cZ \in \mathcal{Z}$, then we can further eliminate Z in this expression,

$$= \sup_{\mu\in\Delta_0(\mathcal{Z})} \mathbb{E}_\mu X \;, \tag{2}$$

where $\Delta_0(\mathcal{Z}) := \{\mu \in \Delta(\Omega) \mid \forall Z \in \mathcal{Z}, \mathbb{E}_\mu Z \leqslant 0\}$ is the set of distributions *consistent* with the gambles. (If μ has $\mathbb{E}_\mu Z > 0$ for any $Z \in \mathcal{Z}$, then by scaling Z the infimum in the right-hand side of eq. (1) is $-\infty$.) Now any statement of the form "$\mathbb{E}_\mu X \leqslant c$ for all $\mu \in \Delta_0(\mathcal{Z})$" can be translated to a game-theoretic version, $\overline{\mathsf{E}}X \leqslant c$. Many of the results of Shafer and Vovk [1, 2] can thus be interpreted as minimax theorems for particular gamble spaces.

References

[1] G. Shafer and V. Vovk. *Probability and finance: It's Only a Game!* Wiley, 2001.

[2] G. Shafer and V. Vovk. *Game-Theoretic Foundations for Probability and Finance.* Wiley, 2019.

Bayes factors, e-values, and p-values for the simultaneous analysis of many contingency tables

Thorsten Dickhaus

Analyzing many contingency tables simultaneously is important in the context of genetic association studies. As discussed in prior work (see, e. g., [1]), computing Bayes factors for the null hypothesis of no association between the two categorical variables corresponding to the two dimensions of the contingency table is in this context often more convenient than computing p-values. However, the Bayes factors proposed in [1] are generally not e-values, especially because the null hypothesis of no association is a composite null. Recently, general approaches to computing e-values for multi-sample comparisons based on contingency table data have been presented in [2]. These e-values allow for an arbitrary number of interim analyses. Hence, they may be over-conservative in certain situations, in which the number of repeated evaluations is limited by design.

The research question that we are interested in is how to define an e-value for a $(2 \times k)$ contingency table, $k \in \{2, 3\}$, which is easy to compute and powerful. In this, we mean by "easy to compute" that resource-intensive operations like a loop over all contingency tables with given marginal counts shall be avoided. This requirement refers to the situation that hundreds of thousands of such e-values have to be computed for one and the same dataset in the case of a genome-wide association study (GWAS). By "powerful", we mean that the e-value should (with high probability) be larger than the p-value based on Fisher's exact test calibrated to the e-value scale by a "p-to-e-calibrator" in the sense of Section 2 in [3], at least if the alternative hypothesis of association is true. We do, however, not necessarily require an anytime-valid e-value, because recruiting many patients sequentially is often logistically infeasible for GWAS, and because oftentimes individual patient data are in this context unavailable to the data analyst, for confidentiality reasons.

Our primarily intended use case is a GWAS which is either multi-centric (fixed-sized and independent patient groups are recruited at different locations) or group-sequential (fixed-sized and independent patient groups are recruited at the same location at different time points). These two sampling schemes are realistic for GWAS, and e-values (which are easy to compute and powerful) would allow the data analysts to combine the evidence across centers or across time points, respectively, in a convenient manner (e. g., by multiplication or by averaging of e-values).

The author thanks all participants of the Oberwolfach Workshop 2419b on "Game-theoretic Statistical Inference: Optional Sampling, Universal Inference, and Multiple Testing Based on E-values" for discussing the aforementioned research question with him. A final solution to the problem is yet outstanding.

References

[1] T. Dickhaus, *Simultaneous Bayesian analysis of contingency tables in genetic association studies*, Statistical Applications in Genetics and Molecular Biology **14** (2015), 347–360.

[2] R. J. Turner, A. Ly, P. D. Grünwald, *Generic E-variables for exact sequential k-sample tests that allow for optional stopping*, Journal of Statistical Planning and Inference **230** (2024), Article 106116.

[3] V. Vovk, R. Wang, *E-values: Calibration, combination and applications*, The Annals of Statistics **49**, 1736–1754.

The numeraire e-variable and reverse information projection

Martin Larsson

(joint work with Aaditya Ramdas and Johannes Ruf)

We consider testing a composite null hypothesis $\mathcal{P}$ against a point alternative Q using e-variables, which are nonnegative—possibly infinite—random variables X such that $\mathbb{E}_{\mathsf{P}}[X] \leqslant 1$ for all $\mathsf{P} \in \mathcal{P}$. Here $\mathcal{P}$ is a nonempty family of probability measures on an arbitrary measurable space $(\Omega, \mathcal{F})$ and Q is a probability measure on this space. We establish that under no conditions whatsoever on $\mathcal{P}$ or Q, there exists a special e-variable X^* that we call the *numeraire e-variable* (or just the *numeraire*), which is strictly positive and satisfies $\mathbb{E}_{\mathsf{Q}}[X/X^*] \leqslant 1$ for every other e-variable X. Equivalently, X^* has the log-optimality property that $\mathbb{E}_{\mathsf{Q}}[\log(X/X^*)] \leqslant 0$ for every e-variable X. The numeraire is unique up to $\mathbb{Q}$-nullsets. The terminology derives from mathematical finance, where the *numeraire portfolio* is a central object analogous to the numeraire e-variable.

Once the numeraire X^* has been shown to exist, a satisfactory duality theory is obtained in a straightforward manner. Specifically, X^* identifies a particular sub-probability measure P^* via the density $d\mathsf{P}^*/d\mathsf{Q} = 1/X^*$. As a result, X^* can be seen as a generalized likelihood ratio of Q against $\mathcal{P}$. The measure P^* coincides with the well-known reverse information projection (RIPr) when additional assumptions are made that are required for the latter to exist. Furthermore, in the general case, P^* satisfies properties associated with the RIPr, such as minimizing relative entropy between the alternative and the null. This makes P^* a natural definition of the RIPr in the absence of any assumptions on $\mathcal{P}$ or Q.

Our theory depends crucially on the concept of the *effective null*. This is the set $\mathcal{P}^{\circ\circ}$ consisting of all sub-probabilities P such that $\mathbb{E}_{\mathsf{P}}[X] \leqslant 1$ for every e-variable X. In other words, the effective null is the set of sub-probabilities against which the e-variables for $\mathcal{P}$ are powerless, and is larger than $\mathcal{P}$ in general. The RIPr P^* belongs to the effective null and, moreover, an e-variable with nontrivial power against Q exists if and only if Q does not belong to the effective null. In the language of convex analysis, the effective null can be viewed as the bipolar of $\mathcal{P}$.

In addition to the abstract theory, we provide several tools for finding the numeraire and RIPr in concrete cases. We discuss several nonparametric examples where we can indeed identify the numeraire and RIPr, despite not having a reference measure. Our results have interpretations outside of testing in that they yield the optimal Kelly bet against $\mathcal{P}$ if we believe reality follows Q.

Finally, we develop a more general optimality theory that goes beyond the ubiquitous logarithmic utility. We focus on certain power utilities, leading to reverse Rényi projections in place of the RIPr, which also always exist.

This talk is based on the preprint [1].

References

[1] M. Larsson, A. Ramdas, J. Ruf, *The numeraire e-variable and reverse information projection*, arXiv:2402.18810

Distribution-uniform anytime-valid inference

Ian Waudby-Smith

(joint work with Edward H. Kennedy and Aaditya Ramdas)

What is asymptotic anytime-valid inference? One of the core goals of anytime-valid inference is to derive confidence sequences (CSs) — sequences of confidence intervals (CIs) that are uniformly valid for all sample sizes. This literature has historically taken a mostly nonasymptotic approach to inference so that type-I errors and coverage probabilities hold in finite samples [7]. For example, the traditional definition of a (nonasymptotic) confidence sequence with coverage $(1-\alpha)$ for a parameter $\theta \in \mathbb{R}$ is a sequence of sets $(C_n)_{n=1}^{\infty}$ so that

$$\sup_{P \in \mathcal{P}} \mathbb{P}_P\left(\exists n \geqslant 1 : \theta \notin C_n\right) \leqslant \alpha, \tag{1}$$

where $\alpha \in (0,1)$. However, nonasymptotic approaches generally require strong assumptions on the random variables such as lying in a parametric family, *a priori* known bounds on their support, or on their moments.

This work takes an *asymptotic* view of anytime-valid inference where type-I errors and coverage probabilities hold in the limit [8, 9, 1]. Following [9], the *sequence of sequences* of sets $(C_k^{(m)})_{k=m}^{\infty}$; $m = 1, 2, \ldots$ is said to have $(1-\alpha)$-coverage for a parameter $\theta \in \mathbb{R}$ if

$$\sup_{P \in \mathcal{P}} \limsup_{m \to \infty} \mathbb{P}_P\left(\exists k \geqslant m : \theta \notin C_k^{(m)}\right) \leqslant \alpha. \tag{2}$$

For example, if $(X_n)_{n=1}^{\infty}$ are i.i.d. random variables with mean θ and a finite $(2+\delta)^{\text{th}}$ moment for some $\delta > 0$, then

$$C_k^{(m)} := \frac{1}{k}\sum_{i=1}^{k} X_i \pm \widehat{\sigma}_k \cdot \sqrt{\frac{\Psi^{-1}(1-\alpha) + \log(k/m)}{k}} \tag{3}$$

satisfies (1), where $\widehat{\sigma}_k$ is the sample standard deviation, Ψ is an invertible function given by $\Psi(x) := 1 - 2\left[1 - \Phi(\sqrt{x}) + \sqrt{x}\phi(\sqrt{x})\right]$ and Φ and ϕ are the distribution and density functions of a standard Gaussian, respectively.

An advantage of the bound in (3) despite the weaker guarantee in (2) is that it can be used under substantially weaker conditions than those satisfying the nonasymptotic guarantee in (1), such as finite moment assumptions appearing in central limit theorem-based confidence intervals.

What is distribution-uniform anytime-valid inference? Notice, however, that (2) is a P-pointwise statement. In this work, we say that $C_k^{(m)}$ has $\underline{\mathcal{P}\text{-uniform } (1-\alpha)\text{-coverage}}$ if the supremum over $P \in \mathcal{P}$ and the limit supremum over $m \in \mathbb{N}$ are swapped, i.e. if

$$\limsup_{m \to \infty} \sup_{P \in \mathcal{P}} \mathbb{P}_P\left(\exists k \geqslant m : \theta \notin C_k^{(m)}\right) \leqslant \alpha, \tag{4}$$

and we show that the very same bound given in (3) satisfies (4) under the assumption that the $(2+\delta)^{\text{th}}$ absolute moment is *uniformly* bounded: $\sup_{P \in \mathcal{P}} \mathbb{E}_P|X -$

$\mathbb{E}_P X|^{2+\delta} < \infty$ and the variance is uniformly positive: $\inf_{P \in \mathcal{P}} \mathrm{Var}_P(X) > 0$. The guarantee in (4) implies the usual pointwise guarantee in (2) but highlights that asymptotic approximations hold *uniformly* within the class $\mathcal{P}$.

Showing that the bound in (3) satisfies the uniform guarantee in (4) is rather nontrivial. Prior work on asymptotic confidence sequences and coverage heavily relied on almost-sure analogues of central limit theorems called "strong Gaussian approximations" [6, 4], but this literature has thus far been entirely P-pointwise. For example, the Komlós-Major-Tusnády approximations state that if $(X_n)_{n=1}^{\infty}$ are i.i.d. on a probability space $(\Omega, \mathcal{F}, P)$ with $\mathbb{E}_P|X|^q < \infty$, then without loss of generality,[1] there exist i.i.d. Gaussians $(Y_n)_{n=1}^{\infty}$ such that

$$\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} Y_i = o\left(n^{1/q}\right) \tag{5}$$

P-almost surely. However, it is not clear what it would even mean for the guarantee in (5) to hold "uniformly" in a class of distributions $\mathcal{P}$. In this work, we give a definition of a $\mathcal{P}$-uniform strong approximation as well as a theorem stating how such an approximation can hold under uniformly bounded moment assumptions. In short, we show the following.

Theorem 1 (Distribution-uniform strong Gaussian approximation)**.** *Let* $(X_n)_{n=1}^{\infty}$ *be i.i.d. on* $(\Omega, \mathcal{F}, \mathcal{P}) \equiv (\Omega, \mathcal{F}, P)_{P \in \mathcal{P}}$ *with means* $\mu_P := \mathbb{E}_P(X)$ *and variances* $\sigma_P^2 := \mathbb{E}_P(X - \mu_P)^2$. *If* X *has* $q > 2$ *uniformly upper-bounded moments, and a uniformly positive variance, i.e.*

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P|X - \mu_P|^q < \infty \quad \textit{and} \quad \inf_{P \in \mathcal{P}} \sigma_P^2 > 0, \tag{6}$$

then there exists a construction with independent standard Gaussians $(Y_n)_{n=1}^{\infty} \sim N(0,1)$ *so that*

$$\left| \sum_{i=1}^{n} \frac{X_i - \mu_P}{\sigma_P} - \sum_{i=1}^{n} Y_i \right| = \bar{o}_{\mathcal{P}}(n^{1/q} \log^{2/q}(n)). \tag{7}$$

In Theorem 1, the convergence $\bar{o}_{\mathcal{P}}(\cdot)$ is a notion of time- and $\mathcal{P}$-uniform convergence introduced by [2] that generalizes P-almost sure convergence to a class of distributions $\mathcal{P}$. Applying Theorem 1 along with other properties of suprema of Gaussian processes, we obtain that (3) satisfies (4).

Applications to conditional independence testing. As one application of this work, we derive anytime-valid distribution-uniform tests of conditional independence without relying on Model-X assumptions. Concretely, given access to $\mathbb{R} \times \mathbb{R} \times \mathbb{R}^d$-valued i.i.d. triplets $(X_n, Y_n, Z_n)_{n=1}^{\infty}$, we want to test whether

$$H_0: \ X \perp\!\!\!\perp Y \mid Z,$$

[1] Here, "without loss of generality" means that one may need to construct a new probability space rich enough to describe Gaussian random variables.

versus the alternative that some conditional dependence exists. It is known that conditional independence testing is impossible (in a formal sense) without structural assumptions [5] which is why the existing anytime-valid tests of conditional independence rely on the so-called "Model-X" assumption where the conditional distribution of $X \mid Z$ is assumed to be known (e.g. in [3]). Moving beyond Model-X to fully nonparametric assumptions has been done in the fixed-n setting, notably by Shah & Peters [5]. They construct the "Generalized Covariance Measure" statistic GCM_n given by

$$\mathrm{GCM}_n := \frac{1}{n} \sum_{i=1}^{n} \{X_i - \widehat{\mu}_n^x(Z_i)\} \cdot \{Y_i - \widehat{\mu}_n^y(Z_i)\},$$

where $\widehat{\mu}_n^x$ and $\widehat{\mu}_n^y$ are estimates of the conditional means $\mu^x(Z) := \mathbb{E}_P(X \mid Z)$ and $\mu^y(Z) := \mathbb{E}_P(Y \mid Z)$ of X and Y given Z, respectively. They show that under certain nonparametric conditions on the estimability of μ^x and μ^y, the statistic GCM_n has a Gaussian limit and can be used to carry out distribution-uniform (fixed-n) inference.

However, the requisite statistical theory for analyzing this statistic in the anytime-valid regime did not yet exist. The aforementioned strong Gaussian approximation of Theorem 1 and the distribution-uniform coverage guarantees of (3) now allow us to obtain the following anytime-valid test of conditional independence.

Theorem 2 ($\mathcal{P}_0$-uniform type-I error control of the SeqGCM)**.** *Suppose $(X_n, Y_n, Z_n)_{n=1}^{\infty}$ are $\mathbb{R} \times \mathbb{R} \times \mathbb{R}^d$-valued triplets defined on the probability spaces $(\Omega, \mathcal{F}, \mathcal{P})$ and let $\mathcal{P}_0 \subseteq \mathcal{P}$ be a collection of distributions in $\mathcal{P}$ satisfying the conditional independence null H_0 as well as*

$$\sup_{P \in \mathcal{P}_0} \|\widehat{\mu}^x - \mu^x\|_{L_2(P)} \|\widehat{\mu}^y - \mu^y\|_{L_2(P)} = O\left(1/\sqrt{n \log^{2+\delta}(n)}\right)$$

for some $\delta > 0$, along with some other regularity conditions. Define

$$\bar{p}_{k,m}^{\mathrm{GCM}} := 1 - \Psi\left(k(\overline{\mathrm{GCM}}_k)^2 - \log(k/m)\right) \tag{8}$$

where $\overline{\mathrm{GCM}}$ is a minor modification of GCM *whose details we omit. Then $(\bar{p}_{k,m}^{\mathrm{GCM}})_{k=m}^{\infty}$ forms a $\mathcal{P}_0$-uniform anytime p-value for the conditional independence null:*

$$\lim_{m \to \infty} \sup_{P \in \mathcal{P}_0} \sup_{\alpha \in (0,1)} \left| \mathbb{P}_P\left(\exists k \geqslant m : \bar{p}_{k,m}^{\mathrm{GCM}} \leqslant \alpha\right) - \alpha \right| = 0. \tag{9}$$

A future direction. While this would be of purely probabilistic interest and would not advance the statistical goals discussed above, in future work, we aim to drop the logarithmic factor found in (7) to obtain a uniform strong Gaussian approximation theorem that generalizes the Komlós-Major-Tusnády approximations found in (5). This would require a rather different (and more sophisticated) set of proof techniques than those used to derive Theorem 1.

References

[1] Bibaut, Aurélien, Kallus, Nathan, and Lindon, Michael, "Near-Optimal Non-Parametric Sequential Tests and Confidence Sequences with Possibly Dependent Observations," *arXiv preprint arXiv:2212.14411*, 2022.

[2] Chung, Kai Lai, "The strong law of large numbers," *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, vol. 2, pp. 341–353, University of California Press, 1951.

[3] Grünwald, Peter, Henzi, Alexander, and Lardy, Tyron, "Anytime-valid tests of conditional independence under model-X," *Journal of the American Statistical Association*, pp. 1–12, Taylor & Francis, 2023.

[4] Komlós, János, Major, Péter, and Tusnády, Gábor, "An approximation of partial sums of independent RV's, and the sample DF. II," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 34, no. 1, pp. 33–58, Springer-Verlag, 1976.

[5] Shah, Rajen D. and Peters, Jonas, "The hardness of conditional independence testing and the generalised covariance measure," *The Annals of Statistics*, vol. 48, no. 3, pp. 1514–1538, 2020.

[6] Strassen, Volker, "An invariance principle for the law of the iterated logarithm," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 3, no. 3, pp. 211–226, Springer, 1964.

[7] Ramdas, Aaditya, Grünwald, Peter, Vovk, Vladimir, and Shafer, Glenn, "Game-theoretic statistics and safe anytime-valid inference," *Statistical Science*, 2023.

[8] Robbins, Herbert and Siegmund, David, "Boundary crossing probabilities for the Wiener process and sample sums," *The Annals of Mathematical Statistics*, pp. 1410–1429, JSTOR, 1970.

[9] Waudby-Smith, Ian, Arbour, David, Sinha, Ritwik, Kennedy, Edward H, and Ramdas, Aaditya, "Time-uniform central limit theory and asymptotic confidence sequences," *The Annals of Statistics (accepted)*, 2024.

Online Model Selection

Parnian Kassraie

We consider the problem of sequential inference and optimization, when the target function can be queried iteratively, but drawing samples is costly. This setting formalizes applications such as molecular design, personalized mHealth, scheduled clinical trials, and environmental monitoring, to name a few. Sequential decision-making and Bandits address such problems through algorithms that iteratively interact with the environment by drawing samples that are expected to be informative, or yield a high target value. To this end, such algorithms maintain an adaptive estimate of the target function, and use it for choosing the next sample. The statistical modeling of the target function plays a crucial role here; it is not known a priori which model is going to yield the most sample efficient algorithm, and we can only select the right model as we gather empirical evidence. This leads us to ask, can we perform online model selection, while simultaneously optimizing for the target function?

We detail the problem of online model selection and its challenges, e.g., handling non-i.i.d. and non-diverse data. We recover a scenario under which simultaneous model selection and optimization is possible, and propose an exponential weighting algorithm for probabilistic model aggregation. The algorithm can be stopped at any time with valid regret guarantees, and its regret has an exponentially improved dependence ($\log M$) on the number of models M. Our approach utilizes a novel time-uniform analysis of the Lasso and establishes a new connection between online learning and high-dimensional statistics. This result is presented in [1].

Open Direction (1). We tackle the problem of online model selection over classes of linear functions, however it remains open for general non-parametric model classes. Iterating back to the open problem of [2], we ask, on which classes can we perform online model selection with a regret of rate $\log M$?

Open Direction (2). Model selection seems to inherently rely on diversity of data. This requires us to mix our sampling method with uniform draws, using a mixing ratio that vanishes as more samples are acquired. However, we conjecture that such pure exploration is not required, and there is *just enough diversity* in the data that is collected for online inference as [3] might suggest.

References

[1] P. Kassraie, N. Emmenegger, A. Krause, and A. Pacchiano (2023). *Anytime Model Selection in Linear Bandits.* Proceedings of Advances in Neural Information Processing Systems.

[2] A. Agarwal, H. Luo, B. Neyshabur, and R. Schapire (2017). *Corralling a band of bandit algorithms.* In Conference on Learning Theory.

[3] D. Banerjee, A. Ghosh, S. Chowdhury, A. Gopalan (2023). *Exploration in Linear Bandits with Rich Action Sets and its Implications for Inference.*

Multiple testing in game-theoretic probability

Vladimir Vovk

The usual way of testing probability forecasts in game-theoretic probability is via construction of test martingales. The standard assumption is that all forecasts are output by the same forecaster. In my talk in Oberwolfach and paper [9] (prepared in support of the talk), I discussed possible extensions of this picture to testing probability forecasts output by several forecasters. This corresponds to multiple hypothesis testing in statistics. One interesting phenomenon is that even a slight relaxation of the requirement of family-wise validity leads to a very significant increase in the efficiency of testing procedures. The main goal of the paper and talk was to report results of preliminary simulation studies.

Game-theoretic probability, as presented in, e.g., my joint books [4] and [5] with Glenn Shafer, is based on the idea that a null hypothesis can be tested dynamically by gambling against it. More generally, we are testing a player called Forecaster, which can be a scientific theory, a computer program, a human forecaster, etc. The gambler starts from an initial capital of 1 and is required to keep his capital nonnegative. His current capital is interpreted as the degree to which the null hypothesis has been undermined.

The idea of testing via gambling goes back at least to Richard von Mises's principle of the impossibility of a gambling system (Unmöglichkeit eines Spielsystems [7, p. 58]; see also [8]), but von Mises's notion of gambling was too narrow, and it was only applicable to infinite sequences. The narrowness of von Mises's notion of gambling was demonstrated by Ville [6, Sect. II.4] (for an English translation, see [3]). Ville proposed extending von Mises's testing procedure to using nonnegative martingales [6, Chap. IV], but he is surprisingly terse when using his wider notion of testing to restate von Mises's principle of the impossibility of a gambling system, especially in the two philosophical chapters [6, preliminary chapter and Chap. 6] (even though he had been interested in the impossibility of gambling systems long before he started writing his book [6]: see [2, Sect. 5.3]). It appears that the idea of testing using nonnegative martingales emerged gradually in various fields, including the algorithmic theory of randomness.

In my paper and talk, I discussed testing several forecasters in one go, with different forecasters being tested at different steps. Testing by gambling can be studied in the usual setting of measure-theoretic probability, and this is what I did, for simplicity and as a first step. Replacing measure-theoretic probability by game-theoretic probability as mathematical foundation for our definitions and results was mentioned as one of directions of future research. For now, each forecaster was formalized as a composite null hypothesis, represented by a set of probability measures on the sample space.

In principle, we can consider two settings for testing multiple null hypotheses. In the *closed* setting, we have a fixed number K of null hypotheses. In the *open* setting, the number of null hypotheses is not known in advance and is potentially infinite. In my paper and talk I concentrated on the closed setting.

These are some possible directions of further research that I mentioned:

- The motivation behind my paper and talk was coming from game-theoretic probability and statistics, but their mathematical setting was that of measure-theoretic probability. Replacing measure-theoretic probability by purely game-theoretic probability (as developed in [5]) would simplify the exposition and lead to more natural and general definitions.
- I concentrated on simulation studies. It would be interesting to conduct empirical studies on benchmark or real-world datasets, for example ones collected in the course of statistical meta-analyses.
- The experimental results that I reported in the paper and talk established confidence regions for the numbers of true discoveries, which can be restated as results about the false discovery proportions, FDP. Are there any interesting theoretical results in this context about false discovery rates, FDR (as in [1] in the case of p-values and [10] in the case of e-values)?
- My paper and talk concentrated on the closed setting (when the number of null hypotheses K is given in advance). The open setting, where new hypotheses may appear at any moment, may be even more interesting. In this case we need, of course, to break the symmetry (which I assumed) between the null hypotheses: there is no uniform probability measure on $\{1, 2, \dots\}$.

References

[1] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, **57**:289–300, 1995.

[2] Pierre Crépel. Jean Ville remembers martingales. In Laurent Mazliak and Glenn Shafer, editors, *Splendors and Miseries of Martingales: Their History from the Casino to Mathematics*, pages 375–391. Birkhäuser, Cham, 2022.

[3] Glenn Shafer. A counterexample to Richard von Mises's theory of collectives, by Jean Ville. Available on `http://www.probabilityandfinance.com/history.html` (accessed in May 2024), 2005.

[4] Glenn Shafer and Vladimir Vovk. *Probability and Finance: It's Only a Game!* Wiley, New York, 2001.

[5] Glenn Shafer and Vladimir Vovk. *Game-Theoretic Foundations for Probability and Finance.* Wiley, Hoboken, NJ, 2019.

[6] Jean Ville. *Etude critique de la notion de collectif.* Gauthier-Villars, Paris, 1939.

[7] Richard von Mises. Grundlagen der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, **5**:52–99, 1919.

[8] Richard von Mises. *Wahrscheinlichkeit, Statistik, und Wahrheit.* Springer, Berlin, 1928. English translation: *Probability, Statistics and Truth.* William Hodge, London, 1939.

[9] Vladimir Vovk. Multiple testing in game-theoretic probability: pictures and questions. `https://arxiv.org/abs/2403.11767` March 2024.

[10] Ruodu Wang and Aaditya Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society B*, **84**:822–852, 2022.

Anytime-Valid Inference in Linear Models and Regression-Adjusted Causal Inference

Michael Lindon

Linear models are fundamental tools in statistics, econometrics and causal inference. In randomized experiments, linear models also enjoy a certain robustness property which enables inference on average treatment effects, despite the apparent model misspecification. As interest in Anytime-Valid inference continues to grow, it is increasingly asked how to perform anytime-valid inference for the linear model. In this work [1], we propose a path forward by constructing a mixture martingale, or Bayes factor, based on a multivariate Gaussian mixture over the coefficients of interest, and the right-Haar mixture over the nuisance parameters (nuisance coefficients and residual variance). The final expression

$$B_n(Y_n) = \sqrt{\frac{\det(\Phi)}{\det(\Phi + \tilde{Z}_n'\tilde{Z}_n)}}\,\frac{\left(1 + \frac{\hat{\delta}_n(Y_n)'(\tilde{Z}_n'\tilde{Z}_n - \tilde{Z}_n'\tilde{Z}_n(\Phi + \tilde{Z}_n'\tilde{Z}_n)^{-1}\tilde{Z}_n'\tilde{Z}_n)\hat{\delta}_n(Y_n)}{s_n^2(Y_n)}\right)^{-\frac{\nu_n+d}{2}}}{\left(1 + \frac{\hat{\delta}_n(Y_n)'\tilde{Z}_n'\tilde{Z}_n\hat{\delta}_n(Y_n)}{s_n^2(Y_n)}\right)^{-\frac{\nu_n+d}{2}}}$$

depends conveniently on the same statistics used in a classical fixed-n inference. In particular, $\hat{\delta}_n(Y_n)$ is the ordinary least squares (OLS) estimator of δ, $\tilde{Z}_n'\tilde{Z}_n = Z_n'(W_n(W_n'W_n)^{-1}W_n' - X_n(X_n'X_n)^{-1}X_n')Z_n = Z_n'(I_n - X_n(X_n'X_n)^{-1}X_n')Z_n$ is the precision matrix of the multivariate Gaussian sampling distribution of $\hat{\delta}_n(Y_n)/\sigma \sim$

$N(\delta/\sigma, (Z_n'Z_n)^{-1})$, $s_n^2(Y_n) = Y_n'(I_n - W_n(W_n'W_n)^{-1}W_n')Y_n/\nu_n$ is the linear model estimate of the residual variance and $\nu_n = n - p - d$ is the degrees of freedom.

The right-Haar prior exploits the group invariance structure of the linear model to result in a test statistic that is a nonnegative supermartingale, and consequently an e-process, for all values of the nuisance parameters. Group invariance arguments are common tricks to reduce composite null hypotheses to simple null hypotheses, and detailed examples can be found in [2]. Interestingly, these authors also provide a safe test for single coefficients of a linear model, with some differences. While our test handles the multivariate case of testing a collection of regression coefficients, their test only handles single coefficients. In our construction, a composite alternative is provided by taking a mixture, whereas their test uses a point alternative. Lastly, both tests use different test-statistics.

References

[1] M. Lindon, D. W. Ham, M. Tingley, and I. Bojinov, *Anytime-Valid Linear Models and Regression Adjusted Causal Inference in Randomized Experiments*, arXiv:2210.08589 [stat.ME], 2022.

[2] M. Felipe Pérez-Ortiz, T. Lardy, R. de Heide, and P. Grünwald, *E-Statistics, Group Invariance and Anytime Valid Testing*, arXiv:2208.07610 [math.ST], 2022.

Safely reliable possibilistic e-UQ

Ryan Martin

Consider a statistical model $Z \sim \mathsf{P}_\omega$ for the observable data Z, where P_ω is a probability distribution that depends on a parameter $\omega \in \mathbb{O}$. There's an uncertain "true value," Ω, and the goal is to infer a particular feature $\Theta = f(\Omega)$, taking values in $\mathbb{T} = f(\mathbb{O})$, based on observed data $Z = z$. What it means to "infer" can vary by applications, but my focus here is on *uncertainty quantification* (UQ) about Θ, given $Z = z$, which boils down to the assignment of reliable data-dependent degrees of support and/or plausibility to various hypotheses about Θ. Many have apparently fallen short of this lofty goal, most notably Fisher, but I believe there's hope of uncovering what Efron called the "Holy Grail of statistics."

It is a mathematical fact that the kind of uncertainty that arises in the context of statistical inference generally cannot be quantified reliably using ordinary probability. Indeed, the *false confidence theorem* [1] establishes that, without genuine or believable prior info, every probabilistic quantification of uncertainty—Bayes, fiducial, etc.—tends to assign high confidence, i.e., high posterior probability, to certain false hypotheses [6, 7, 10]. False confidence creates an unacceptable risk of unreliability, or "systematically misleading conclusions" [15]. Fortunately, probability theory is not the only UQ game in town. Alternatives include the Dempster–Shafer theory of belief functions [2, 16], possibility theory [3], and the theory of lower previsions [17]. *Inferential models* (IMs) [12, 13] make up a new framework for statistical reasoning near the boundary between precise and imprecise probability theory. This perspective is necessary to avoid false confidence, etc.

My current efforts [9, 11, 10] focus on the construction of *possibilistic IMs*, where the IM output takes the mathematical form of a possibility measure.[1] There are a number of reasons why I like this form, but it's worth noting here that possibility theory is among the simplest of the imprecise probability theories. It closely mimics probability theory but with a different calculus—where probability theory relies on integration, possibility theory uses optimization. The specific proposal put forward in the above references starts with the IM's contour function

$$\pi_z(\theta) = \sup_{\omega : f(\omega) = \theta} \mathsf{P}_\omega\{\rho(Z, \theta) \leqslant \rho(z, \theta)\}, \quad \theta \in \mathbb{T},$$

where $\rho(z, \theta)$ is some ranking of the parameter value θ in terms of its compatibility with z—large values indicate higher compatibility. For example, in [9, 11], I recommended taking $\rho(z, \theta)$ to be the relative profile likelihood

$$R(z, \theta) = \frac{\sup_{\omega : f(\omega) = \theta} L_z(\omega)}{\sup_\omega L_z(\omega)}, \quad \theta \in \mathbb{T},$$

with $L_z(\omega)$ the usual likelihood function. From here, the IM's possibility or upper probability output is given by

$$\overline{\Pi}_z(H) = \sup_{\theta \in H} \pi_z(\theta), \quad H \subseteq \mathbb{T}. \tag{1}$$

The interpretation is that a small $\overline{\Pi}_z(H)$ means there's strong evidence in $Z = z$ against the truthfulness of H. This possibilistic IM's output is *reliable* in the sense that there's no false confidence, i.e., no tendency to assign too small of upper probability values to true hypotheses about Θ:

$$\sup_{\omega : f(\omega) \in H} \mathsf{P}_\omega\{\overline{\Pi}_Z(H) \leqslant \alpha\} \leqslant \alpha, \quad H \subseteq \mathbb{T}.$$

Further details, both theory and applications, are given in the above references.

Statistical methods—from classical to the possibilistic IM described above—often assume the data-generating process is fully known. This can be an issue when the data are collected sequentially and the (possibly data-dependent) rule by which the collection stops isn't explicitly stated. Consider (say) an iid sequence $Z_1, Z_2, \ldots$ from P_ω, with $Z^n = (Z_1, \ldots, Z_n)$ the first n instances along the sequence, and define a filtration $\mathcal{F}_n = \sigma(Z^n)$, $n \geqslant 1$. A stopping time N is an integer-valued random variable such that $\{N \leqslant n\} \in \mathcal{F}_n$, $n \geqslant 1$. Then a realization of the data Z^N is of the form z^n, with n the observed value of N and $z^n = (z_1, \ldots, z_n)$ are the observed Z values. If the data analyst isn't privy to which stopping time N was employed, then he/she might just assume that $N \equiv n$ was fixed at the observed value in advance. But this could be very misleading, since the relevant sampling distributions of statistics based on Z^n could be drastically different that based on Z^N, thus jeopardizing the statistical method's reliability. I'll say that a statistical method is *safely reliable* if its reliability holds uniformly over stopping times. How can the above IM be made safely reliable?

[1] There's a complementary necessity measure but I won't discuss it here.

Interestingly, one can interpret N as a sort of "nuisance parameter," so the general rules in [11] suggest a safely reliable possibilistic IM with contour

$$\pi_{z^n}(\theta) = \sup_N \sup_{\omega : f(\omega)=\theta} \mathsf{P}_\omega\{\rho(Z^N, \theta) \leqslant \rho(z^n, \theta)\}, \quad \theta \in \mathbb{T},$$

where the outermost supremum is over all stopping rules. Of course, direct computation of the right-hand side above can be challenging, but there's a simple workaround. If $\rho(z^n, \theta) = \mathrm{E}(z^n, \theta)^{-1}$ is taken to be the reciprocal of an e-process [4, 14, 18] for testing "$\Theta = \theta$" based on data z^n, then an immediate consequence of Ville's inequality is that

$$\pi_{z^n}(\theta) \leqslant \pi_{z^n}^{\mathrm{E}}(\theta) := 1 \wedge \mathrm{E}(z^n, \theta)^{-1}, \quad \theta \in \mathbb{T}.$$

The upper bound corresponds to the proposed *e-posterior* in [5]. But notice that the bound itself is a possibility contour—so there's a corresponding possibilistic IM, with upper probability $\overline{\Pi}_{z^n}^{\mathrm{E}}$ defined via optimization of $\pi_{z^n}^{\mathrm{E}}$ as in (1). This provides possibilistic, e-process-based UQ, or *e-UQ*. Moreover, that $\pi_{z^n}^{\mathrm{E}}$ is an upper bound of the safely reliable π_{z^n} implies that the corresponding possibilistic IM's e-UQ is safely reliable too. Further details on this will be fleshed out elsewhere.

One notable observation is that my proposed possibilistic e-UQ can be used for safely reliable decision-making. Let $\ell_a(\theta) \geqslant 0$ denote the loss associated with taking action $a \in \mathbb{A}$ when the world is in state θ. Then the possibilistic IM's upper expected loss associated with an action a is a Choquet integral, which is given by

$$\overline{\Pi}_{z^n}^{\mathrm{E}} \ell_a = \int_0^1 \sup_{\theta : \pi_{z^n}^{\mathrm{E}}(\theta) > s} \ell_a(\theta)\, ds.$$

The claimed reliability of my IM's upper expected loss corresponds to the following extension of the results in [5, 8]:

$$\sup_N \sup_{\omega \in \mathbb{O}} \mathsf{E}\left\{ \sup_{a \in \mathbb{A}} \frac{\ell_a(f(\omega))}{\overline{\Pi}_{Z^N}^{\mathrm{E}} \ell_a} \right\} \leqslant 1.$$

In words, the IM's assessment of an action a won't be more optimistic than that of an oracle who knows the true $\theta = f(\omega)$, and this holds uniformly over actions, true states of the world, and stopping rules. More details will be presented elsewhere.

Open questions include how to incorporate incomplete or partial prior information about $\Theta = f(\Omega)$ and/or about the stopping rule N. Indeed, there's an opportunity for efficiency gain if we don't need to be safe relative to *all* stopping rules. With the inclusion of partial prior information, one could try to generalize the above upper expected loss result and, moreover, try to prove a version of the classical Bayesian result, namely, that the possibilistic IM's optimal action—the one that minimizes the upper expected loss $a \mapsto \overline{\Pi}_{z^n}^{\mathrm{E}} \ell_a$—is admissible. It's relatively straightforward to incorporate partial prior information coming in the form of a possibility measure, and I'll report on this elsewhere. But given the connection to e-values, game-theoretic probability, etc., it's of interest to consider prior information that takes the form of gambles about the uncertain value of Θ that are *a prior* acceptable to the data analyst, i.e., gambles that aren't expected to multiply an opponent's capital by a large factor.

References

[1] M. S. Balch, R. Martin, and S. Ferson. Satellite conjunction analysis and the false confidence theorem. *Proc. Royal Soc. A*, **475** (2019), 2018.0565.

[2] A. P. Dempster. The Dempster–Shafer calculus for statisticians. *Internat. J. Approx. Reason.*, **48** (2008), 365–377.

[3] D. Dubois and H. Prade. *Possibility Theory.* Plenum Press, New York (1988).

[4] P. Grünwald, R. de Heide, and W. Koolen. Safe testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, to appear; `arXiv:1906.07801` (2024).

[5] P. Grünwald. The e-posterior. *Philos. Trans. Roy. Soc. A*, **381** (2023), 2022.0146.

[6] R. Martin. False confidence, non-additive beliefs, and valid statistical inference. *Internat. J. Approx. Reason.*, **113** (2019), 39–73.

[7] R. Martin. An imprecise-probabilistic characterization of frequentist statistical inference. `arXiv:2112.10904` (2021).

[8] R. Martin. Inferential models and the decision-theoretic implications of the validity property. `arXiv:2112.13247` (2021).

[9] R. Martin. Valid and efficient imprecise-probabilistic inference with partial priors, II. General framework. `arXiv:2211.14567` (2022).

[10] R. Martin. A possibility-theoretic solution to Basu's Bayesian–frequentist via media. *Sankhya A*, to appear, `arXiv:2303.17425` (2023).

[11] R. Martin. Valid and efficient imprecise-probabilistic inference with partial priors, III. Marginalization. `arXiv:2309.13454` (2023).

[12] R. Martin and C. Liu. Inferential models: a framework for prior-free posterior probabilistic inference. *J. Amer. Statist. Assoc.* **108**, (2013), 301–313.

[13] R. Martin and C. Liu. *Inferential Models*, volume 147 of *Monographs on Statistics and Applied Probability.* CRC Press, Boca Raton, (2015).

[14] A. Ramdas, P. Grünwald, V. Vovk, and G. Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statist. Sci.* **38**, (2023), 576–601.

[15] N. Reid and D. R. Cox. On some principles of statistical inference. *Int. Stat. Rev.* **83**, (2015), 293–308.

[16] G. Shafer. *A Mathematical Theory of Evidence.* Princeton University Press, Princeton, (1976).

[17] P. Walley. *Statistical Reasoning with Imprecise Probabilities*, volume 42 of *Monographs on Statistics and Applied Probability.* Chapman & Hall Ltd., London, (1991).

[18] L. Wasserman, A. Ramdas, and S. Balakrishnan. Universal inference. *Proc. Natl. Acad. Sci. U.S.A.* **117**, (2020), 16880–16890.

Continuous Monitoring of Systemic Risks

Timo Dimitriadis
(joint work with Yannick Hoga)

In the wake of numerous instances of financial market turmoil in recent decades, increasingly more attention has been paid to *systemic* risks, that is, spillovers of risk from one bank to another, opposed to managing risks of each individual institution in isolation; see for example [1]. This holds for regulators (who became more concerned with the interconnectedness of banks in the financial system) as well as individual financial institutions (in avoiding joint distress across trading desks or business units).

To fix ideas, given absolutely continuous random variables X_t and Y_{kt}, $k = 1, \ldots, K$ with (invertible) distribution functions F_{X_t} and $F_{Y_{kt}}$ representing negative financial log-returns, the Value-at-Risk (VaR) of X_t at level $\beta \in (0,1)$ is defined as the β-quantile $F_{X_t}^{-1}(\beta)$ of X_t. The Conditional Value-at-Risk (CoVaR) as the most prominent systemic risk measure is given by the α-quantile of Y_{kt} given that X_t violates its conditional β-quantile, $\{X_t \geqslant F_{X_t}^{-1}(\beta)\}$, where $\alpha, \beta \in (0,1)$ are values often chosen close to one.

Given sequences of forecasts v_t for the VaR and c_{kt} for the CoVaR of institution k for the trading days $t \in \mathbb{N}$, we propose *continuous monitoring* schemes for these risk measures, which allow to detect changes in systemic risk in an "online" fashion by continuously monitoring (re-testing) every trading day without inflating type I errors. Building on recent forecast evaluation results in [2], [3] and [4], one can show that the typical condition of conditional forecast calibration (also known as forecast optimality or forecast rationality) is equivalent to

$$\begin{aligned} I_t &:= \mathbb{1}_{\{X_t > v_t\}} \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(1-\beta), \\ I_{kt} &:= \mathbb{1}_{\{X_t > v_t,\ Y_{kt} > c_{kt}\}} \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}\big((1-\alpha)(1-\beta)\big), \\ \mathrm{Cov}(I_t, I_{kt}) &= (1-\alpha)\beta(1-\beta). \end{aligned} \tag{1}$$

As the probabilistic structure of these indicator functions is (almost) fully known under the null hypothesis in (1), we can monitor the correctness of (1) through necessary conditions in the form of a moving sum detector and the Gini coefficient in an online fashion by simulating critical values based on a maximum length of the monitoring procedure of $n \in \mathbb{N}$ trading days. This method holds size by construction, such that the null of correct systemic risk assessments is only rejected during the monitoring period with at most a pre-specified probability. The monitoring procedures further allows multiple (i.e., K) series at once to be monitored, thus increasing the likelihood and the speed with which early signs may be picked up.

Such procedures are vital in taking timely countermeasures to avoid financial distress. An empirical application to US banks during multiple crises demonstrates the usefulness of our monitoring schemes for both regulators and financial institutions. Open questions include the relaxation of the finite monitoring length through the use of e-processes akin to [5] and the most suitable generalizations to other systemic risk measures such as the Conditional or Marginal Expected Shortfall.

References

[1] T. Adrian and M. Brunnermeier (2016) *CoVaR*, American Economic Review **106**: 1705–1741.

[2] T. Fissler and Y. Hoga (2024) *Backtesting systemic risk forecasts using multi-objective elicitability*, Journal of Business & Economic Statistics **42**: 485–498.

[3] T. Gneiting, D. Wolffram, J. Resin, K. Kraus, J. Bracher, T. Dimitriadis, V. Hagenmeyer, A.I. Jordan, S. Lerch, K. Phipps and M. Schienle (2023) *Model diagnostics and forecast evaluation for quantiles*, Annual Review of Statistics and Its Application **10**: 597–621.

[4] Y. Hoga and M. Demetrescu (2023) *Monitoring value-at-risk and expected shortfall forecasts*, Management Science **69**: 120–140.

[5] Q. Wang, R. Wang and J. Ziegel (2024) *E-backtesting*, Preprint, `https://arxiv.org/abs/2209.00991`

A generalization of Ville's inequality to possibly negative martingales

Muriel Pérez

(joint work with Tyron Lardy and Wouter Koolen)

We introduce an extension of Ville's inequality to possibly negative martingales. Ville's inequality, a maximal inequality for a certain family of martingales, is at the center of current statistical guarantees for anytime-valid tests. Indeed, if $n \mapsto M_n$ is a nonnegative martingale under a distribution $\mathbf{P}$ with expected value equal to one—a test martingale—, Ville's inequality states that

$$\mathbf{P}\{\exists n : M_n \geqslant 1/\alpha\} \leqslant \alpha.$$

This inequality ensures that a statistical test that monitors M_n continuously and rejects $\mathbf{P}$ as soon as M_n crosses the threshold $1/\alpha$ has type-I error below α uniformly over time—a test with such guarantee is called anytime valid. A large effort has been put into designing test martingales M_n that result in powerful and flexible tests against specific alternatives and extensions of this inequality are known for composite null distributions (see Ruf et al. [3] and Johannes Ruf's extended abstract in this repport) and possibly nonintegrable martingales [4]. A large portion of the design principles used to to build these tests hinge on interpretations of these martingales as betting strategies or as sequential information-theoretical coding schemes. In this interpretations, the fact that test martingales are positive is crucial. The present talk, with its focus on possibly negative martingales, deviates from this this wisdom.

We present an extension of Ville's inequality that holds for martingales that are possibly negative, but still bounded from below by some function. The main result is the following theorem.

Theorem 1. *Let $n \mapsto M_n$ be a supermartingale and let $n \mapsto f_n$ be nondecreasing and such that $n \mapsto M_n + f_n$ is nonnegative. Let $n \mapsto g_n$ be nondecreasing, and such that $f_0 + g_0 \geqslant 0$ and $\mathbf{E}_{\mathbf{P}}[M_0] \in [-f_0, g_0]$. Then*

$$\mathbf{P}\{\exists n : M_n \geqslant g_n\} \leqslant 1 - \frac{g_0 - \mathbf{E}_{\mathbf{P}}[M_0]}{g_0 + f_0} \prod_{i=1}^{\infty} \frac{g_i + f_{i-1}}{g_i + f_i}.$$

Furthermore, if f and g are not only defined on $\mathbf{N}_{\geqslant 0}$ but on $\mathbf{R}_{\geqslant 0}$ and they are both differentiable, the r.h.s. is bounded by $1 - \frac{g_0 - \mathbf{E}_{\mathbf{P}}[M_0]}{g_0 + f_0} \exp\left(-\int_0^\infty \frac{f_t'}{f_t + g_t} \mathrm{d}t\right)$.

This inequality is sharp two senses: firstly, when M_n is nonnegative ($f_n = 0$ in that case), Ville's inequality is recovered; secondly, there exists a martingale that attains equality. At this point it may not be clear why such a generalization is useful, but such martingales are suggested by the analysis of online-learning algorithms (see Wouter Koolen's extended abstract in this report) and they can be used to derive tight time-uniform probabilistic bounds directly. This last feature is best illustrated with an example. Indeed, if $X_1, X_2, \dots$ is a sequence of i.i.d. standard normal random variables, the analysis of Squint [1] suggests a functional form similar to $M_n^\star = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \left(\mathrm{e}^{\eta \sum_{i \leqslant n} X_i - \eta^2 n/2} - 1 \right) \mathrm{e}^{-\eta^2/2} \frac{\mathrm{d}\eta}{|\eta|}$ —it can be shown that $M_n^\star$ bounded by $\ln(1+n)/\sqrt{2\pi}$ from below. Using this particular martingale, we show a choice of g_n for which Theorem 1 implies a finite-time law of the iterated logarithm for $n \mapsto \sum_{i \leqslant n} X_i$ with sharp constants, reproducing one of the fundamental results in this branch of sequential analysis with a more direct proof than with standard methods [2]. This direct proof becomes possible, on the technical side, because $M_n^\star$ includes an "improper prior" $\mathrm{d}\eta/|\eta|$ on possibly negative martingales, while standard methods use approximations of $\mathrm{d}\eta/|\eta|$ with "proper priors" on nonnegative martingales. It is known that improperness is necessary to derive laws of the iterated logarithm, but the standard method of mixtures excludes it.

This result opens the door to use a larger family of martingales for anytime-valid inference by lifting the nonnegativity restriction. It remains to be seen whether these objects have a game-theoretic, betting or information-theoretic interpretation; to explore more deeply the connection with online learning; and to see if unknown concentration results become accessible with these techniques.

References

[1] Wouter M. Koolen and Tim van Erven. Second-order Quantile Methods for Experts and Combinatorial Games. In *Proceedings of The 28th Conference on Learning Theory*, pages 1155–1175. PMLR, June 2015. URL `https://proceedings.mlr.press/v40/Koolen15a.html`. ISSN: 1938-7228.

[2] Herbert Robbins and David Siegmund. Boundary Crossing Probabilities for the Wiener Process and Sample Sums. *The Annals of Mathematical Statistics*, 41(5):1410–1429, October 1970. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177696787. Publisher: Institute of Mathematical Statistics.

[3] Johannes Ruf, Martin Larsson, Wouter M. Koolen, and Aaditya Ramdas. A composite generalization of Ville's martingale theorem, May 2023. URL `http://arxiv.org/abs/2203.04485`. arXiv:2203.04485 [cs, math, stat].

[4] Hongjian Wang and Aaditya Ramdas. The extended Ville's inequality for nonintegrable nonnegative supermartingales, April 2024. URL `http://arxiv.org/abs/2304.01163`. arXiv:2304.01163 [math, stat].

E-values as unnormalized weights in multiple testing

Nikolaos Ignatiadis
(joint work with Ruodu Wang, Aaditya Ramdas)

We study how to combine p-values and e-values, and design multiple testing procedures where both p-values and e-values are available for every hypothesis [1].

To describe the basic setting, let $H_1, \dots, H_K$ be K hypotheses, and write $\mathcal{K} = \{1, \dots, K\}$. Denote the true (unknown) data-generating probability measure by $\mathbb{P}$. For each $k \in \mathcal{K}$, we may think of hypothesis H_k as defining a set of joint probability measures, and H_k is called a true null hypothesis if $\mathbb{P} \in H_k$. A *p-value* P for a hypothesis H is a random variable that satisfies $Q(P \leqslant t) \leqslant t$ for all $t \in [0,1]$ and all $Q \in H$. In other words, a p-value is stochastically larger (or equal) than a uniform random variable $\mathrm{U}(0,1)$. An *e-value* E for a hypothesis H is a $[0,\infty]$-valued random variable satisfying $\mathbb{E}^Q(E) \leqslant 1$ for all $Q \in H$. Let $\mathcal{N} \subseteq \mathcal{K}$ be the unknown index set of true null hypotheses.

Two settings of testing multiple hypotheses were considered in [2]. In the first setting, for each $k \in \mathcal{K}$, P_k is a p-value for H_k. In the second setting, for each $k \in \mathcal{K}$, E_k is an e-value for H_k. Here we consider the setting where both P_k and E_k are available for each H_k. Since we are testing whether $\mathbb{P} \in H_k$ for each k, we will only use the following condition: if $k \in \mathcal{N}$, then $\mathbb{P}(P_k \leqslant t) \leqslant t$ for all $t \in [0,1]$ and $\mathbb{E}^{\mathbb{P}}(E_k) \leqslant 1$. We impose no restrictions on P_k and E_k if $k \notin \mathcal{N}$.

We also write $\mathcal{D}$ for a multiple testing procedure, that is, a Borel mapping that produces a subset of $\mathcal{K}$ representing the indices of rejected hypotheses based on p-values (we write p-$\mathcal{D}$ to denote a procedure $\mathcal{D}$ that is based only on p-values), e-values, or a combination of both as the input. Below we construct new procedures $\mathcal{D}$ that take as input both p-values and e-values.

Our starting point is the following: If P_k and E_k are independent, how should we combine them to form a new p-value P_k^*? To be more formal, we call a function $f : [0,1] \times \overline{\mathbb{R}}_+ \to [0,1]$ an i-pe/p combiner if $f(P,E)$ is an e-value for any independent p-value P and e-value E, and $(p,e) \mapsto f(p,e)$ is non-decreasing in p and non-increasing in e.

We have the following result:

Theorem 1. *Consider the function Q defined by $Q(p,e) := (p/e) \wedge 1$. This function is an admissible i-pe/p combiner.*

We call the function above the "Q-combiner." We describe its properties as a general-purpose method for meta-analysis from two studies, where a primary dataset is used to compute p-values, and an independent secondary dataset is used to compute e-values. Furthermore, the Q-combiner is used as the main building block for our main proposal, which turns multiple testing procedures based only on p-values (p-$\mathcal{D}$) into procedures that can use both p-values and e-values:

Definition 1 (e-weighted p-value procedure (ep-$\mathcal{D}$))**.** *Let p-$\mathcal{D}$ be a multiple testing procedure based on p-values. Given p-values $(P_1, \dots, P_K)$ and e-values $(E_1, \dots, E_K)$, we define the e-weighted p-value procedure ep-$\mathcal{D}$ which proceeds as follows: for*

$k \in \mathcal{K}$, compute the Q-combiner $P_k^* := Q(P_k, E_k) = (P_k/E_k) \wedge 1$, and then supply $(P_1^*, ..., P_K^*)$ to p-$\mathcal{D}$.

We describe type-I error control guarantees for several different ep-$\mathcal{D}$ procedures including the ep-BH (Benjamini-Hochberg) and ep-Bonferroni procedures. These procedures may be interpreted in two ways: first, they are p-value based procedures applied to the p-value vector $(P_1^*, \ldots, P_K^*)$, and second, they are weighted p-value based procedures with p-value vector $(P_1, \ldots, P_K)$ and weight vector $(E_1, \ldots, E_K)$. We build on both perspectives to derive guarantees on the control of generalized type-I error rates. Our guarantees depend on the dependence of the e-values $(E_1, \ldots, E_K)$, the p-values $(P_1, \ldots, P_K)$, as well as the cross-dependence between p-values and e-values (for example, to guarantee that P_k^* is indeed a p-value, we require that E_k is independent of P_k for all $k \in \mathcal{N}$).

The perspective in terms of weighted multiple testing is important as our results provide a new perspective on multiple testing with data-driven weights: while standard weighted multiple testing methods require the weights to deterministically add up to the number of hypotheses being tested, we show that this normalization is not required when the weights are e-values that are independent of the p-values. Our procedures can result in a substantial increase in power, especially if the nonnull hypotheses have e-values much larger than one.

Finally, the weighted multiple testing perspective demonstrates that for several of our guarantees, e.g., for false discovery rate control of ep-BH or family-wise error rate control of ep-Bonferroni, it suffices to relax the notion of e-value. Instead, it suffices that $(E_1, \ldots, E_K)$ are compound e-values as defined below.

Definition 2 (Compound e-values [1]). *We say that $(E_1, \ldots, E_K)$ are compound e-values, if the following holds:*

$$\sum_{k \in \mathcal{N}} \mathbb{E}^{\mathbb{P}}[E_k] \leqslant K.$$

The importance of this notion for multiple testing has been convincingly demonstrated in prior work [2, 3]. Here, we provide a name for this property (inspired by Herbert Robbins' compound decision theory) and show its versatility and usefulness in the context of multiple testing with p-values and e-values.

References

[1] Nikolaos Ignatiadis, Ruodu Wang, Aaditya Ramdas, *E-values as unnormalized weights in multiple testing*, Biometrika **111(2)** (2024), 417–439.

[2] Ruodu Wang, Aaditya Ramdas, *False discovery rate control with e-values*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **84(3)** (2022), 822–852.

[3] Zhimei Ren, Rina Foygel Barber, *Derandomised knockoffs: leveraging e-values for false discovery rate control*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **86(1)** (2024), 122–154.

Supermartingales in Online Learning

Wouter M. Koolen

We investigate one fascinating connection between supermartingales and online learning algorithms. We start by seeing how we can look at the classic Hedge algorithm in this way. We will then discuss the design of two sophisticated online learning algorithms: Squint and Muscada. In each, we go over the desiderata, review the construction of the supermartingale, what its design achieves, and why certain things don't work. On the way, we highlight connections to e-processes, deviation inequalities and to defensive forecasting.

1. Introduction

We investigate connections between sequential testing and online learning. The reason to revisit these connections is the following. Martingales are popular tools for sequential testing that have recently gotten new attention by the explosion of interest in e-values. It has been known since [1] that so-called test supermartingales can be converted into learning algorithms by a method called Defensive Forecasting.

2. Online Learning

We revisit the classic Hedge setting with K experts. Here for rounds $t = 1, 2, \dots$

- The Learner plays weights $\boldsymbol{w}_t \in \triangle_K$ from the probability simplex.
- The Adversary picks a bounded loss vector $\boldsymbol{\ell}_t \in [0,1]^K$.

After T rounds, the *regret* w.r.t expert k is defined as

$$R_T^k \;:=\; \sum_{t=1}^{T} \left(\boldsymbol{w}_t^\top \boldsymbol{\ell}_t - \ell_t^k \right).$$

The goal from the online learning perspective is to develop a strategy for Learner keeping all R_T^k small, against any Adversary.

To test whether a given Learner is doing a good job, an external observer (customarily called the Skeptic) may engage in a sequence of bets that cost 0 and pay off $\boldsymbol{w}_t^\top \boldsymbol{\ell}_t - \ell_t^k$. In other words, Skeptic may construct a supermartingale by sequentially multiplying conditional e-values of the form $1 + \sum_{k=1}^K \eta_t^k \left(\boldsymbol{w}_t^\top \boldsymbol{\ell}_t - \ell_t^k \right)$ for η_t^k positive yet small enough to avoid bankruptcy. One simple and effective way to construct such a supermartingale uses mixing over experts k, repeating a fixed bet $\eta > 0$ over and over, and invoking the convenient bound

$$1 + \eta \left(\boldsymbol{w}_t^\top \boldsymbol{\ell}_t - \ell_t^k \right) \;\geqslant\; e^{\eta \left(\boldsymbol{w}_t^\top \boldsymbol{\ell}_t - \ell_t^k \right) - \eta^2/2}.$$

With that, we have motivated the choice of the *Hedge supermartingale*

$$\Phi_T \;:=\; \sum_{k=1}^{K} \frac{1}{K} \prod_{t=1}^{T} e^{\eta \left(\boldsymbol{w}_t^\top \boldsymbol{\ell}_t - \ell_t^k \right) - \eta^2/2} \;=\; \sum_{k=1}^{K} \frac{1}{K} e^{\eta R_T^k - T\eta^2/2}.$$

One can check that the choice $\eta = \sqrt{\frac{2 \ln \frac{K}{\alpha}}{T}}$ ensures that at time T the supermartingale Φ_T is below the threshold $\frac{1}{\alpha}$ only if the regret $R_T^k \leqslant \sqrt{2T \ln \frac{K}{\alpha}}$ for every expert k. In other words, Φ_T gets big when Learner is incurring too much regret.

Why is this interesting? Because Defensive Forecasting allows us to take the supermartingale Φ_T as the *specification of a good Learner* and synthesise a learning algorithm from it. In our context, this means choosing the weights $\boldsymbol{w}_t \in \triangle_K$ so that even for the worst-case losses $\boldsymbol{\ell}_t \in [0,1]^K$ still $\Phi_t \leqslant \Phi_{t-1}$. Solving this min-max requirement leads to the solution

$$w_{T+1}^k = \frac{e^{\eta R_T^k}}{\sum_{j=1}^K e^{\eta R_T^j}}$$

which is the well-known Hedge algorithm introduced and analysed by [2]. We next review two extensions of these ideas that have been developed in the online learning world, and interpret them as supermartingales.

3. The Squint Supermartingale

For our first interesting online learning supermartingale we turn to a strategy called Squint [3]. Let us introduce the abbreviation $r_t^k := \boldsymbol{w}_t^\top \boldsymbol{\ell}_t - \ell_t^k$ for the instantaneous regret w.r.t. expert k in round t. The *Squint supermartingale* is given by

$$\Phi_T := \sum_{k=1}^K \pi_k \int_0^{\frac{1}{2}} \frac{e^{\eta R_T^k - \eta^2 V_T^k} - 1}{\eta} \, \mathrm{d}\eta \qquad \text{where} \qquad V_T^k = \sum_{t=1}^T (r_t^k)^2$$

where $\boldsymbol{\pi} \in \triangle_K$ is any prior distribution on K experts chosen by the user.[1] Keeping this supermartingale from growing is possible (this requirement essentially fixes the learning algorithm), and has desirable consequences in terms of online learning: no tuning parameter, anytime regret guarantees, stochastic luckiness, quantile bounds and adaptivity to the complexity of the comparator. Moreover, this comes at no increase in computational cost due to available closed-form expressions for the integral. But in terms of testing by betting something unexpected happened. Note that the density $\frac{1}{\eta}$ w.r.t. with which we are integrating is *improper*, as it packs too much mass close to $\eta = 0$. Yet since we are mixing centred (by subtracting -1 in the numerator) supermartingales, this does not make the value diverge immediately. However, something gives: this supermartingale is not guaranteed to be non-negative. Yet not all is lost, as one can show that $\Phi_T \geqslant -\ln(1+T)$. Within the testing community, we need to learn to harvest the benefits of working with supermartingales that are bounded below by a function of time. We refer to the contribution of Muriel Pérez Ortiz (also in this workshop report) for more details on first steps in that regard.

[1] In fact, here we can tolerate countably many experts too.

4. The Muscada Supermartingale

For our second exhibit we turn to multi-scale online learning and the Muscada strategy [4]. The setup is the standard expert setting, with one new ingredient. We fix a vector $\boldsymbol{\sigma} \in (0, \infty)^K$ of positive loss ranges, and we constrain the losses $\boldsymbol{\ell}_t$ to be such that $\ell_t^k \in [\pm\sigma_k]$. So the losses are still bounded, but the ranges differ per expert. Here the *Muscada supermartingale* is constructed as follows:

$$\Phi_T \;:=\; \max_{\boldsymbol{w} \in \triangle(K)} \langle \boldsymbol{w}, \boldsymbol{R}_T - \boldsymbol{\mu}_T \rangle - D_{\boldsymbol{\eta}_T}(\boldsymbol{w}, \boldsymbol{u}). \tag{1}$$

where for $\boldsymbol{w}, \boldsymbol{u} \in \triangle_K$ the relative entropy at multi-scale $\boldsymbol{\eta} \in (0, \infty)^K$ is

$$D_{\boldsymbol{\eta}}(\boldsymbol{w}, \boldsymbol{u}) \;=\; \sum_{k=1}^{K} \frac{w_k \ln(w_k/u_k) - w_k + u_k}{\eta_k}$$

and we define $\boldsymbol{\mu}_T$ by $\mu_T^k := \sigma_k \sqrt{T \ln K}$ and set $\eta_T^k \approx \frac{1}{\sigma_k}\sqrt{\frac{\ln K}{T}}$ (we refer to [4] for full details). With this notation, the online learning goal is to ensure $R_T^k \leqslant \mu_T^k$. And indeed, we find that Defensive Forecasting results in a strategy for choosing the weights $\boldsymbol{w}_T$ that guarantees

$$R_T^k \;\leqslant\; \sigma_k \sqrt{T \ln K}.$$

One could argue that the supermartingale above is constructed to make the Defensive Forecasting proof go through. Most interestingly here is that Φ_T is not of mixture-over-experts form. While that form worked for the same-scale Hedge case, it apparently does not work multi-scale. The reason being that we do not control the range of the instantaneous loss $r_t^k = \boldsymbol{w}_t^\top \boldsymbol{\ell}_t - \ell_t^k$. While the range of ℓ_t^k is $[\pm\sigma_k]$, the range of $\boldsymbol{w}_t^\top \boldsymbol{\ell}_t$ can still be $[\pm \max_k \sigma_k]$.

Yet the optimization-based and mixture-based supermartingales are related, for the same-scale case, by duality for KL. That is for any $\boldsymbol{\pi} \in \triangle_K$ and $\boldsymbol{X} \in \mathbb{R}^K$,

$$\ln \sum_k \pi_k e^{X_k} \;=\; \max_{\boldsymbol{w} \in \triangle_K} \langle \boldsymbol{w}, \boldsymbol{X} \rangle - \mathrm{KL}(\boldsymbol{w} \| \boldsymbol{\pi}).$$

The proof that a certain choice of weights $\boldsymbol{w}_t$, namely the minimiser of Φ_{t-1} in (1), keeps $\Phi_t \leqslant 0$ can be found in [4]. What we find interesting is that the form of the supermartingale [4] makes it possible to use the knowledge about the per-expert loss ranges effectively.

5. Conclusion

It is our belief that testing by betting can learn some cool techniques from online learning. The success criterion in online learning, small regret, makes that supermartingales appearing there have a certain from. That form, perhaps, made it possible to invent new techniques. Yet these are not online learning specific; instead, they can often be translated back to pure testing-by-betting of whether data conforms to a probabilistic forecast. We brought two such techniques to the attention of the testing community, and we are looking forward to further discussions.

References

[1] A. Chernov and V. Vovk. Prediction with expert evaluators' advice. In *ALT Proceedings*, volume 5809 of *Lecture Notes in Computer Science*, pages 8–22, 2009.

[2] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139, 1997.

[3] W. M. Koolen and T. van Erven. Second-order quantile methods for experts and combinatorial games. In P. Grünwald, E. Hazan, and S. Kale, editors, *Proceedings of the 28th Annual Conference on Learning Theory (COLT)*, pages 1155–1175, June 2015.

[4] M.F. Pérez-Ortiz and W. M. Koolen. Luckiness in multiscale online learning. In *Advances in Neural Information Processing Systems (NeurIPS) 35*, Dec. 2022.

$\mathrm{KL}_{\mathrm{inf}}$ and Optimality in Multi-Armed Bandits

Shubhada Agrawal

(joint work with Sandeep Juneja, Wouter Koolen, Peter Glynn)

Real-world machine learning applications often involve making decisions sequentially through dynamic interactions with the environment, based on limited feedback. While tech giants like Google and Meta use machine learning algorithms to generate billions in revenue via online advertising, their adoption in safety-critical domains is limited. This limitation arises mainly due to a lack of understanding of their performance in diverse practical environments. Thus, it is crucial to develop a foundational understanding of the performance of these algorithms and their associated statistical limitations.

The multi-armed bandit (MAB) problem is a simple and elegant statistical model for interactive learning in uncertain environments. Originating from Thompson's work in the 1930s on adaptive clinical trials [3], the MAB framework involves an algorithm interacting with a fixed set of unknown and independent probability distributions, or arms. In each iteration, the algorithm selects an arm and receives a sample from the underlying distribution, considering prior actions and outcomes. The objective is to choose arms to optimize a certain objective.

Classical MAB settings have been extensively studied. Instance-dependent lower bounds for these problems, as well as optimal algorithms that match these lower bounds even in the multiplicative constants on every bandit instance, have been developed under minimal assumptions on the uncertainty distributions. A naturally occurring quantity in these lower bounds is the infimum of KL-divergences between probability measures, denoted as $\mathrm{KL}_{\mathrm{inf}}$. Moreover, the optimal algorithms that match these lower bounds, even in the multiplicative constants, rely on the empirical version of $\mathrm{KL}_{\mathrm{inf}}$.

In this talk, we will first consider a simple 1-armed bandit problem, or a specific sequential hypothesis testing problem (with composite null and point alternative). We will derive a lower bound on the average number of samples required by any algorithm to ensure δ-correctness in this setup. This lower bound will involve $\mathrm{KL}_{\mathrm{inf}}$. We will then examine a natural first algorithm and demonstrate its sub-optimality. Subsequently, we will explore modifications to arrive at an exactly-optimal algorithm, which will notably rely on $\mathrm{KL}_{\mathrm{inf}}$. We will also present a concentration result for the empirical $\mathrm{KL}_{\mathrm{inf}}$ statistic in a non-parametric setting that allows for heavy-tailed distributions. Finally, we will discuss an extension to the multi-armed bandit setting and conclude with several structural and topological properties of $\mathrm{KL}_{\mathrm{inf}}$.

This talk is based on [1] and [2], which developed optimal algorithms for bandits with heavy-tailed distributions and the theory of $\mathrm{KL}_{\mathrm{inf}}$ for general classes of distributions.

References

[1] S. Agrawal, S. Juneja, P. Glynn *Optimal delta-correct best-arm selection for heavy-tailed distributions*, In Proceedings of the 31st International Conference on Algorithmic Learning Theory, PMLR, **117** (2020), 61–110.

[2] S. Agrawal, S. Juneja, W.M. Koolen *Regret minimization in heavy-tailed bandits*, In Proceedings of Thirty Fourth Conference on Learning Theory, PMLR, **134** (2021), 26–62.

[3] W.R. Thompson, *On the likelihood that one unknown probability exceeds another in view of the evidence of two samples*, Biometrika **25** (1933), 285–294.

On universal inference in Gaussian mixture models

Hongjian Shi

(joint work with Mathias Drton)

Let $\{\mathrm{P}_\theta : \theta \in \Theta\}$ be a parametric statistical model, with parameter space $\Theta \subseteq \mathbb{R}^d$. The distributions P_θ are assumed to be dominated by a common measure ν, and have probability densities f_θ with respect to ν. Suppose that the observations $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.) according to an unknown distribution P_θ in the model. We are now interested in a testing problem

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta \backslash \Theta_0 \tag{1}$$

for a subset $\Theta_0 \subsetneq \Theta$.

Let $\ell(\theta) = \sum_{i=1}^{n} \log f_\theta(X_i)$ be the log-likelihood function. The (classical) likelihood ratio statistic for (1) is given by

$$\lambda_n := 2\Big\{ \sup_{\theta\in\Theta} \ell(\theta) - \sup_{\theta\in\Theta_0} \ell(\theta) \Big\}.$$

For regular problems, asymptotically valid likelihood ratio tests (LRTs) may be constructed via Wilks' theorem, i.e., the fact that the distribution of λ_n asymptotically converges to the chi-squared distribution χ^2_m under the null hypothesis. However, when regularity conditions fail, it can be difficult to provide theoretical insights on the distribution of likelihood ratios and standard bootstrapping is not necessarily valid; see, e.g., [1]. These issues are particularly pressing for mixture models.

Recent work on game-theoretic statistics and safe anytime-valid inference (SAVI) (see [4] for a comprehensive review) provides new tools for statistical inference without assuming any regularity conditions. In particular, the framework of universal inference proposed by [5] offers new solutions by modifying the likelihood ratio test in a data-splitting scheme. The data are divided into two parts, D_0

for inference and D_1 for estimation. For this split, choose a fraction $m_0 \in (0,1)$ and partition the data into two disjoint subsets $D_0 = \{X_{1,0}, \dots, X_{\lfloor m_0 n \rfloor,0}\}$ and $D_1 = \{X_{1,1}, \dots, X_{\lceil m_1 n \rceil,1}\}$, where $m_1 := 1 - m_0$. We will write n_0 for $\lfloor m_0 n \rfloor$ and n_1 for $\lceil m_1 n \rceil$ to shorten notation. Let $\ell_k(\theta) = \sum_{i=1}^{n_k} \log f_\theta(X_{i,k}),\ k = 0, 1$, be the likelihood functions based on D_0 and D_1, respectively. Let $\widehat{\theta}_{n,1} := \arg\max_{\theta \in \Theta} \ell_1(\theta)$ be the maximum likelihood estimator (MLE) of θ under the full model and based on D_1, and let $\widehat{\theta}_{n,0} := \arg\max_{\theta \in \Theta_0} \ell_0(\theta)$ be the MLE of θ under H_0 and based on D_0. Now the *split likelihood ratio statistic* is defined as

$$\lambda_n^{\text{split}} := 2\Big\{\ell_0(\widehat{\theta}_{n,1}) - \ell_0(\widehat{\theta}_{n,0})\Big\}. \tag{2}$$

As shown in [5], under the null hypothesis $H_0 : \theta \in \Theta_0$, it holds for any positive integer n that

$$\mathrm{E}_\theta[\exp(\lambda_n^{\text{split}}/2)] \leqslant 1. \tag{3}$$

An application of Markov's inequality yields for any $\alpha \in (0,1)$ and any positive integer n,

$$\mathrm{P}_\theta(\lambda_n^{\text{split}} > -2\log\alpha) \leqslant \alpha. \tag{4}$$

Accordingly, the split likelihood ratio test (split LRT) given by $\mathbb{1}(\lambda_n^{\text{split}} > -2\log\alpha)$ is finite-sample-valid at significance level α.

In this paper, we study the performance of the resulting split likelihood ratio test under the Gaussian mixture model

$$f_{p,t}(x) = (1-p)\phi(x; 0, 1) + p\phi(x; t, 1) \tag{5}$$

where the mixture weight $p \in [0,1]$ and the mean $t \in \mathbb{R}$ are unknown parameters. We consider the homogeneity testing problem

$$H_0 : p = 0 \text{ or } t = 0 \text{ against } H_1 : p \in (0,1), t \in \mathbb{R}\backslash\{0\}. \tag{6}$$

This model is a canonical example of models in which classical regularity conditions fail to hold. In particular, the likelihood ratio statistic diverges to $+\infty$ in probability at the order of $O(\log\log n)$ as proven in [3].

Proposition 1 (Theorem 2 in [3])**.** *The likelihood ratio statistic λ_n for testing homogeneity in the Gaussian mixture model (5) satisfies*

$$\lim_{n\to\infty} \mathrm{P}_{H_0}\{\lambda_n - \log\log n + \log(2\pi^2) \leqslant x\} = \exp\{-\exp(-x/2)\}, \quad x \in \mathbb{R}.$$

Consequently,

$$\lim_{n\to\infty} \mathrm{P}_{H_0}\{\lambda_n > c_{n,\alpha}\} = \alpha,$$

where the critical value is defined as

$$c_{n,\alpha} = \log\log n - \log(2\pi^2) - 2\log\log(1-\alpha)^{-1}. \tag{7}$$

We first establish that under the null hypothesis, the split likelihood ratio statistic is asymptotically normal with increasing mean and variance.

Theorem 2. *Suppose that $X_1, \dots, X_n$ are i.i.d. standard normal random variables. The asymptotic null distribution of the split LRT is obtained as*

$$\frac{\lambda_n^{\text{split}} + \frac{m_0}{m_1} \log\log n}{2\sqrt{\frac{m_0}{m_1} \log\log n}} \xrightarrow{\mathsf{d}} N(0,1). \tag{8}$$

As a direct corollary of Theorem 2, if we adopt the asymptotic critical point from the asymptotic null distribution (8), namely,

$$c_{n,\alpha}^{\text{split}} := 2\sqrt{\frac{m_0}{m_1} \log\log n} \times \Phi^{-1}(1-\alpha) - \frac{m_0}{m_1} \log\log n, \tag{9}$$

then the split LRT will have the asymptotic size of α:

$$\lim_{n\to\infty} \mathrm{P}_{H_0}\{\lambda_n^{\text{split}} > c_{n,\alpha}^{\text{split}}\} = \alpha.$$

Moreover, contradicting the usual belief that the flexibility of SAVI and universal methods comes at the price of a significant loss of power, we are able to prove that universal inference surprisingly achieves the same detection rate $(n^{-1} \log\log n)^{1/2}$ as the classical likelihood ratio test. In detail, we consider the following sequence of local alternative hypotheses:

$$H_{1,n}^{\#} : p = q_n,\, t = \mu_n, \text{ with } q_n\mu_n = \gamma(n^{-1}\log\log n)^{1/2},\, \mu_n = O\{(\log n)^{-1/2}\} \tag{10}$$

of the model $f_{p,t}(x) = (1-p)\phi(x;0,1) + p\phi(x;t,1)$. The following result, due to [2], shows that the LRT can distinguish the null hypothesis from the local alternative at the rate $(n^{-1}\log\log n)^{1/2}$. In addition, the rate $(n^{-1}\log\log n)^{1/2}$ is optimal in the sense that there is a dramatic change in the power of the LRT at $|\gamma| = 1$.

Proposition 3 (Theorem 2.1 in [2])**.** *Under the sequence of local alternative hypotheses $H_{1,n}^{\#}$ given in* (10)*, the asymptotic local power of the LRT is given by*

$$\lim_{n\to\infty} \mathrm{P}_{H_{1,n}^{\#}}\{\lambda_n > c_{n,\alpha}\} = \begin{cases} \alpha, & \textit{if } |\gamma| < 1, \\ (1+\alpha)/2, & \textit{if } |\gamma| = 1, \\ 1, & \textit{if } |\gamma| > 1. \end{cases}$$

A similar phenomenon can also be found in the split likelihood ratio test, which is summarized in Theorem 4 below.

Theorem 4. *Under the sequence of local alternative hypotheses $H_{1,n}^{\#}$ given in* (10)*, the asymptotic local power of the split LRT is given by*

$$\lim_{n\to\infty} \mathrm{P}_{H_{1,n}^{\#}}\{\lambda_n^{\text{split}} > -2\log(\alpha)\} = \begin{cases} 0, & \textit{if } |\gamma| < m_1^{-1/2}, \\ 1/2, & \textit{if } |\gamma| = m_1^{-1/2}, \\ 1, & \textit{if } |\gamma| > m_1^{-1/2}, \end{cases}$$

and if the asymptotic critical point $c_{n,\alpha}^{\text{split}}$ defined in (9) *is adapted, then the split LRT will have very similar asymptotic local power as the LRT with a shifted threshold:*

$$\lim_{n\to\infty} \mathrm{P}_{H_{1,n}^{\#}}\{\lambda_n^{\text{split}} > c_{n,\alpha}^{\text{split}}\} = \begin{cases} \alpha, & \textit{if } |\gamma| < m_1^{-1/2}, \\ (1+\alpha)/2, & \textit{if } |\gamma| = m_1^{-1/2}, \\ 1, & \textit{if } |\gamma| > m_1^{-1/2}. \end{cases}$$

References

[1] Mathias Drton and Benjamin Williams, *Quantifying the failure of bootstrap likelihood ratio tests*, Biometrika **98** (2011), no. 4, 919–934. MR 2860333

[2] Peter Hall and Michael Stewart, *Theoretical analysis of power in a two-component normal mixture model*, J. Statist. Plann. Inference **134** (2005), no. 1, 158–179. MR 2146091

[3] Xin Liu and Yongzhao Shao, *Asymptotics for the likelihood ratio test in a two-component normal mixture model*, J. Statist. Plann. Inference **123** (2004), no. 1, 61–81. MR 2058122

[4] Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer, *Game-theoretic statistics and safe anytime-valid inference*, Statist. Sci. **38** (2023), no. 4, 576–601. MR 4665027

[5] Larry Wasserman, Aaditya Ramdas, and Sivaraman Balakrishnan, *Universal inference*, Proc. Natl. Acad. Sci. USA **117** (2020), no. 29, 16880–16890. MR 4242731

Game-theoretic statistical modelling

Glenn Shafer

Statistical modelling is an art, not a mathematical exercise. One first chooses or invents a *statistical model*, which is a mathematical object. Then, as Paul R. Rosenbaum has explained [3, p. 46]:

> . . . a separate argument, not always a particularly clear or compelling argument, is invoked to connect this convenient but rather technical model to the scientific problem at hand. The arguments that connect statistical models to important scientific questions—these connectivity arguments—are often most compelling to people who do not understand them, and least compelling to people who do.

Perhaps because they are not clear and compelling, statisticians often leave their connectivity arguments implicit.

Many standard connectivity arguments suppose that the model represents a "true" state of affairs and relates this supposition to the use of hypothetical repetitions in assessing conclusions.[1] A game-theoretic connectivity argument, in contrast, interprets probabilities as forecasts. The statistician argues that the model is a good forecaster, in the sense that an opponent will not multiply their money by a large factor betting against its probability forecasts. This formulation may still involve repetition, perhaps actual, perhaps hypothetical, because a single

[1] See for example the well known textbook by David R. Cox and David V. Hinkley [1, p. 45].

bet may offer limited opportunity for such multiplication.[2] But there is no need for "true" in quotation marks.

The shift from truth to good forecasting often makes a statistician's argument neither more nor less compelling. But it may make the argument more understandable and its limitations clearer, especially for those who are not experts in probability and statistics.

Game-theoretic statistical modelling usually uses a game imagined by a statistician (call her *Statistician*, with a capital S), which makes explicit some of the steps she takes in analyzing data. Chapter 10 of [7] gives a few examples. Here I offer some further examples. In each example, we find some or all of these players:

(1) Player I (call him Oracle), who announces information that Statistician does not know, information that provides a full or partial answer to the scientific or practical question.
(2) Player II (call him Informant), who announces information Statistician does know.
(3) Player III (call him Forecaster), who announces bets on what Informant will announce.
(4) Player IV (call him Skeptic), who announces a bet chosen from Forecaster's announcement.

Skeptic begins with unit capital. In this extended abstract, we consider only one round of play, but some of our examples can be extended to multiple rounds.

Before each round, Statistician

- tells Oracle what question to answer,
- tells Informant what to say,
- assigns to Forecaster a strategy that uses what Forecaster has heard so far to tell him what to announce, and
- assigns to Skeptic a strategy that uses what Skeptic has heard so far to announce a bet that does not risk his cumulative capital becoming negative.

In general, the games used in game-theoretic probability are *perfect information* games. This means that each player hears the other players' moves as they are made. This concept is not relevant here, however, because the players are not free. Statistician tells them what to do.

After each round, Statistician can calculate Skeptic's capital. When Oracle is not in the game, and the capital is large enough, Statistician will conclude that the strategy she assigned to Forecaster has been discredited. When Oracle is in the game, Statistician can calculate Skeptic's capital only as a function of Oracle's announcement. She can draw conclusions about Oracle's announcement from this function, the possible announcements that would have produced large

[2] Attempts to give precise mathematical meaning to "true" probabilities or "good" forecaster lead to imagined infinities: unlimited repetition where some event has probability one, or unlimited play where the nonnegative capital process of the forecaster's opponent remains bounded [7]. Unfortunately, imagined infinities may do little to make a statistician's data analysis more compelling.

capital being discredited. These conclusions can be expressed as warranty sets or confidence sets [5].

The strategies Statistician assigns to the players may not use all the information that she has at the outset or acquires in the course of play. This corresponds to "shrinking the filtration" or "forgetting" in measure-theoretic statistics. Moreover, the strategy she assigns to Skeptic may use more information than the strategy she assigns to Forecaster. This is one way game-theoretic statistics can make statistical arguments more explicit and thus clarify aspects that may otherwise seem puzzling or counter-intuitive.

Example 1. Gaussian measurement model. Statistician plans to measure an unknown quantity μ. Calling her measurement Y and its value y, she imagines this protocol.

Oracle announces μ.
Forecaster announces a probability distribution P for Y with $\mathbf{E}_P(Y) = \mu$.
Skeptic announces $Z \geqslant 0$ with $\mathbf{E}_P(Z(Y)) \leqslant 1$.
Informant announces y.
$\mathcal{K} := Z(y)$.

Forecaster's announcement is interpreted as an offer to sell for 1 any Z satisfying $Z \geqslant 0$ and $\mathbf{E}_P(Z(Y)) \leqslant 1$. The quantity $\mathcal{K}$ is Skeptic's return on his investment of 1. If $\mathcal{K}$ large, Skeptic has multiplied his capital by a lot and discredited Forecaster.

Statistician and her public believe that the standard normal is a good forecaster of the error of her measuring instrument. So she assigns to Forecaster the strategy that announces $\mathcal{N}_{\mu,1}$ when Oracle announces μ. Once she has done this, a strategy for Skeptic is a mapping $\mathcal{Z}$ that assigns to each μ a variable $\mathcal{Z}_\mu$ satisfying $\mathcal{Z}_\mu \geqslant 0$ and $\mathbf{E}_{\mathcal{N}_{\mu,1}}(\mathcal{Z}_\mu(Y)) \leqslant 1$.

We can distinguish two fairly distinct classes of strategies for Skeptic.

(1) Statistician might choose each $\mathcal{Z}_\mu$ to make $\mathcal{K}$ larger the farther μ is from y. An extreme choice is to make $\mathcal{Z}_\mu$ an all-or-nothing bet. For example: $\mathcal{Z}_\mu(y) := 20$ when $|y - \mu| > 1.96$ and $\mathcal{Z}_\mu(y) := 0$ when $|y - \mu| \leqslant 1.96$. This makes the interval $y \pm 1.96$ a 20-fold warranty interval and a 95% confidence interval for μ [5].

(2) Statistician might assess a subjective probability distribution for μ, use it to average over the normal distributions with variance 1, and then use the resulting distribution for Y as the alternative in a Kelly bet against each $\mathcal{N}_{\mu,1}$. This also produces warranty and confidence sets. Here Statistician is assigning Forecaster the task of assuring validity (based on the consensus in favor of standard normal errors) and Skeptic the task of assuring efficiency (relative to her subjective opinion) [6].

Example 2. Linear regression. Statistician realizes that the temperature may be influencing her measurements of μ. The previous experience that had justified her forecasting the errors with the $\mathcal{N}_{0,1}$ distribution was based on measurements at 0° Celsius. To test whether temperature is making a difference and take it into account, she imagines this protocol:

Oracle announces μ, β.
Informant announces x, the temperature on the Celsius scale.
Forecaster announces a distribution P for Y with $\mathbf{E}_P(Y) = \mu + \beta x$.
Skeptic announces $Z \geqslant 0$ with $\mathbf{E}_P(Z(Y)) = 1$.
Informant announces $y \in \mathbb{R}$.
$\mathcal{K} := Z(y)$.

Statistician replaces Forecaster with the strategy that announces $\mathcal{N}_{\mu+\beta x,1}$. A strategy for Skeptic is then a mapping $\mathcal{Z}$ that assigns to each triplet (μ, β, x) a function $\mathcal{Z}_{\mu,\beta,x}$ satisfying $\mathcal{Z}_{\mu,\beta,x} \geqslant 0$ and $\mathbf{E}_{\mathcal{N}_{\mu+\beta x,1}}(\mathcal{Z}_{\mu,\beta,x}(Y)) \leqslant 1$.

As in Example 1, Statistician has a wide variety of possible strategies for Skeptic. Whatever strategy she assigns to Skeptic, she will obtain warranty sets for the pair (μ, β). If there is a values of μ such that $(\mu, 0)$ is in the 4-fold warranty set, say, then she may decide that it is unnecessary to take temperature into account and return to the model in Example 1.

Example 3. Fisher's exact test. In a clinical trial, N patients are chosen randomly from a group of $2N$ to receive treatment A; the other N receive treatment B. The outcome is to survive or not. Statistician writes s for the observed number of survivors, Y for the number of these survivors receiving treatment A, and y for Y's value. Statistician imagines this protocol.

Informant announces s.
Forecaster announces a probability distribution P for Y.
Skeptic announces $Z \geqslant 0$ with $\mathbf{E}_P(Z(Y)) \leqslant 1$.
Informant announces y.
$\mathcal{K} := Z(y)$.

Statistician assigns to Forecaster the strategy that announces the distribution P_s for Y obtained by conditioning the randomization probabilities on s. A strategy for Skeptic is now a mapping $\mathcal{Z}$ that assigns to each s a variable $\mathcal{Z}_s$ satisfying $\mathcal{Z}_s \geqslant 0$ and $\mathbf{E}_{P_s}(\mathcal{Z}_s(Y)) \leqslant 1$.

As in Example 1, there are many different possibilities for $\mathcal{Z}$. Fisher recommended all-or-nothing bets that pay off for large enough y [2]. Another possibility is Kelly bets that use as their alternative the distribution for Y obtained by conditioning on s Statistician's subjective probabilities for the outcomes assessed using everything she knows before observing any results of the experiment. Here again she assigns Forecaster the task of assuring validity and Skeptic the task of assuring efficiency relative to her subjective opinion. This is one way of resolving the Bayesian discomfort with randomization tests discussed by L. J. Savage [4, p. 34].

Example 4. Conformal prediction. Here we use concepts and terminology defined in [8], limited to a single prediction. Statistician observes pairs of the form $z = (x, y)$. The xs, called *objects*, are from a space $\mathbf{X}$. The ys, called *labels*, are from a space $\mathbf{Y}$. The product $\mathbf{Z} := \mathbf{X} \times \mathbf{Y}$ is called the *example space*.

Statistician has observed $n - 1$ examples and an additional object x, and she wants to "predict" x's label y. (We use quotation marks, because she may never observe or otherwise learn y's exact value.) Her basis for prediction is the opinion that (x, y) is not particularly different from previous examples. She makes this into a forecast using the concept of a *smoothed conformal transducer* [8, p. 55].

To define a smoothed conformal transducer f, Statistician begins with a way of scoring how different one example is from $n-1$ others. For each possible value y, she puts (x, y) in a bag with the $n-1$ examples she has observed and scores how different each is from the others. She writes $n_<$ the number of examples in the bag that score less different than (x, y) and $n_=$ for the number that tie with (x, y). Then she draws a pseudo-random number τ from $[0, 1]$ and sets $f(y) := \frac{n_< + \tau n_=}{n}$. Her forecast, which combines her expectation that (x, y) will not be particularly different with her opinion that τ is random, is the uniform probability distribution for $f(y)$.

Using these ideas, Statistician imagines this very simple protocol:

Forecaster announces a smoothed conformal transducer f.
Skeptic announces a probability density q on $[0, 1]$.
Oracle announces y.
$\mathcal{K} := q(f(y))$.

Because the uniform probability density for $f(y)$ is implicit in the forecast, Skeptic's bet is a Kelly bet against f.

If Statistician actually knew y (in this case Informant, not Oracle, would announce it), then $\mathcal{K}$ would be a betting score for testing Statistician's supposition that that the new example is not particularly different from the old ones [8, Pt. III]. But here she is using $\mathcal{K}$ to predict y. As in our other examples, her K-fold warranty interval will consist of the values of y for which $\mathcal{K}$ is less than K.

Statistician's choice of q will depend on her purpose. If she expects to see y, she will have no reason to make future bets on y using other information, and she may make the all-or-nothing bet usually prescribed for conformal prediction. If she does expect to use future evidence to continue betting on y, then she may use a Kelly bet that maximizes the expected logarithm of $\mathcal{K}$ with respect to a probability distribution for y based on evidence she already has about y.

References

[1] D.R. Cox and D.V. Hinkley, *Theoretical Statistics*, Chapman and Hall (1974).

[2] R.A. Fisher, *Design of Experiments*, Oliver & Boyd (1935).

[3] P.R. Rosenbaum, *Observation & Experiment: An Introduction to Causal Inference*, Harvard University Press (2017).

[4] L.J. Savage et al., *The foundations of statistical inference: A discussion opened by Professor L. J. Savage*, Methuen (1962).

[5] G. Shafer, *Testing by betting: A strategy for statistical and scientific communication (with discussion)*, Journal of the Royal Statistical Society: Series A **184(2)** (2021), 407–478.

[6] G. Shafer, *How the game-theoretic foundation for probability resolves the Bayesian vs. frequentist standoff*, pp. 264–275 of Handbook of Bayesian, Fiducial, and Frequentist Inference, J. Berger, X.-L. Meng, N. Reid, and M. Xie (eds.), Chapman & Hall (2024).

[7] G. Shafer and V. Vovk, *Game-Theoretic Foundations for Probability and Finance*, Wiley (2019).

[8] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*, 2nd ed., Springer (2022).

[9] V. Vovk and R. Wang, *E-values: Calibration, combination, and applications*, Annals of Statistics, **49(3)** (2021), 1736–1754.

A composite generalization of Ville's martingale theorem using e-processes

Johannes Ruf

(joint work with Martin Larsson, Wouter M. Koolen, Aaditya Ramdas)

We provide a composite version of Ville's theorem that an event has zero measure if and only if there exists a nonnegative martingale which explodes to infinity when that event occurs. This is a classic result connecting measure-theoretic probability to the sequence-by-sequence game-theoretic probability, recently developed by Shafer and Vovk. Our extension of Ville's result involves appropriate composite generalizations of nonnegative martingales and measure-zero events: these are respectively provided by "e-processes", and a new inverse capital outer measure. We then develop a novel line-crossing inequality for sums of random variables which are only required to have a finite first moment, which we use to prove a composite version of the strong law of large numbers (SLLN). This allows us to show that violation of the SLLN is an event of outer measure zero and that our e-process explodes to infinity on every such violating sequence, while this is provably not achievable with a nonnegative (super)martingale. This presentation is based on [1].

References

[1] J. Ruf, M. Larsson, W.M. Koolen, A. Ramdas, *A composite generalization of Ville's martingale theorem using e-processes*, Electronic Journal of Probability **28**(1) (2023), 1–21.

E-Values for Exponential Families

Peter Grünwald

(joint work with Tyron Lardy, Yunda Hao, Shaul K. Bar-Lev, Martijn de Jong)

We provide a general theory for constructing various types of e-variables, including optimal ones (in the GRO sense), when the null hypothesis $\mathcal{P}$ is a composite, multivariate exponential family.

An e-value is the value taken by an e-variable, which is a test statistic that, in contrast to the p-value, is suitable for experiments with a flexible design; see e.g. Ramdas et al. [5] for a comprehensive overview. The most straightforward example of e-variables are likelihood ratios between simple alternatives and simple null hypotheses. E-variables for composite nulls, and in particular 'good' e-variables, are generally more complicated. However, e-variables in the form of a likelihood ratio with a single, special element of the null representing the full, composite null sometimes still exist. We refer to such e-variables as 'simple' e-variables.

Simple e-variables, if they exist, can easily be computed, and are known to be optimal in an expected-log-optimality sense [3, 1]. That is, if we combine evidence from a repeated experiment where data is collected using a fixed stopping rule, then using the simple e-variable will asymptotically result in the most evidence against the null, among all e-variables. As such, it is desirable to find out whether or not simple e-variables exist in specific settings. Our first main result provides a condition – we will call it *Condition Simple* on the family $\mathcal{P}$ and any postulated simple alternative Q, under which simple e-variables exist for exponential family nulls. We then consider the case of the *Anti-Simple* condition. Under this condition, simple e-variables *cannot* exist but there is still a beautiful mathematical structure in the problem that allows us to determine, at least asymptotically, GRO and other (especially *conditional*) e-variables.

We briefly and superficially describe the results here, assuming prior knowledge on e-variables and exponential families. We fix a regular multivariate exponential family null $\mathcal{P}$ for data U with some sufficient statistic vector $X = t(U)$ and a distribution Q for U, outside of $\mathcal{P}$, and with density q. As our most important regularity condition, we assume that Q has a moment generating function and that there exists $P_{\vec{\mu}*} \in \mathcal{P}$ with the same mean of X, say $\vec{\mu}^*$, as Q. It is known that $P_{\vec{\mu}*}$ is the *Reverse Information Projection (RIPr)* of Q onto $\mathcal{P}$ [4], that is, it achieves $\min_{P \in \mathcal{P}} D(Q \| P)$. Denoting the density of $P_{\vec{\mu}*}$ by $p_{\vec{\mu}*}$, it follows by Theorem 1 of [1] that $q(U)/p_{\vec{\mu}*}(U)$ would be an e-variable in case $\inf_{P \in \text{CONV}(\mathcal{P})} D(Q \| P) = \min_{P \in \mathcal{P}} D(Q \| P)$. Our theorem establishes a sufficient condition for when this is actually the case. It is based on constructing a second exponential family $\mathcal{Q}$ with densities proportional to $\exp(\vec{\beta}^T t(U))q(U)$ for varying $\vec{\beta}$: $\mathcal{Q}$ contains Q and has the same sufficient statistic as $\mathcal{P}$. In some cases, but not all, $\mathcal{Q}$ may be thought of as the composite alternative we are interested in.

Condition Simple. *(this part is based on arXiv paper [2])* Letting $\Sigma_p(\vec{\mu})$ and $\Sigma_q(\vec{\mu})$ denote the covariance matrices of the $P_{\vec{\mu}} \in \mathcal{P}$ and $Q_{\vec{\mu}} \in \mathcal{Q}$ with mean $\vec{\mu}$, our first main result implies the following: under a further regularity condition on the parameter spaces of $\mathcal{P}$ and $\mathcal{Q}$, simple e-variables exist whenever $\Sigma_p(\vec{\mu}) - \Sigma_q(\vec{\mu})$ is positive semidefinite for all $\vec{\mu}$ in the mean-value parameter space of $\mathcal{Q}$. We call this *Condition Simple* (additionally, three equivalent rephrasings of the condition are given). In this case, we may further conclude that for *every* element $Q_{\vec{\mu}'}$ of the constructed $\mathcal{Q}$, the likelihood ratio $q_{\vec{\mu}'}(U)/p_{\vec{\mu}'}(U)$ is an e-variable, where $P_{\vec{\mu}}$ is the element of $\mathcal{P}$ to which $Q_{\vec{\mu}}$ is projected. An example pair $(Q, \mathcal{P})$ to which the theorem applies is when, under Q, $U \sim N(m, s^2)$ for fixed m, s^2 and $\mathcal{P} = \{N(0, \sigma^2) : \sigma^2 > 0\}$ is the univariate (scale) family of normal distributions. This situation is illustrated in Figure 1. The proof of this result is based on convex duality properties of exponential families.

Condition Anti-Simple. *(this part has not been published yet)* If, broadly speaking, $\Sigma_p(\vec{\mu}) - \Sigma_q(\vec{\mu})$ is *negative* semidefinite for all $\vec{\mu}$ in the mean-value parameter space

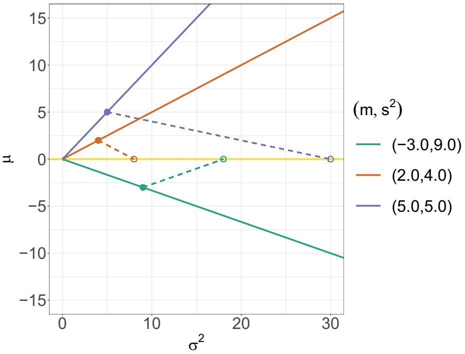

FIGURE 1. The family $\mathcal{Q}$ for various (m, s^2). The coordinate grid represents the parameters of the full Gaussian family, the horizontal line shows the parameter space of $\mathcal{P}$, the sloped lines show the parameters of the distributions in $\mathcal{Q}$, and the dashed lines show the projection of (m, s^2) onto the parameter space of $\mathcal{P}$. For example, we may start out with Q expressing $U \sim N(m, s^2)$ with $m = -3.0, s^2 = 9.0$, represented as the green dot on the green line. Its RIPr onto $\mathcal{P}$ is the green point on the yellow line. The corresponding family $\mathcal{Q}$, constructed in terms of Q and $\mathcal{P}$, is depicted by the green solid line. Our first main theorem implies that the likelihood ratio between any point on the green line and its RIPr onto the yellow line is an e-variable; similarly for the red and blue lines.

of $\mathcal{Q}$, we say that *Condition Anti-Simple* holds. Note that the Simple and Anti-Simple conditions cannot both hold at the same time, but there are cases in which neither holds.

In our second main result, we show that under the anti-simple condition, if we extend the hypotheses to n outcomes by assuming independence, then the Reverse Information Projection (RIPr) of Q onto $\mathcal{P}$ is obtained by adopting a specific Gaussian prior W on the mean-value parameters $\vec{\mu}$, with a covariance matrix that scales as order $1/n$. The GRO e-variable is then given by $q(U)/p_W(U)$, p_W being the marginal density of U obtained if we equip $\mathcal{P}$ with prior W, and it (nontrivially) happens to be equal to the so-called *conditional* e-variable. This holds exactly if $\mathcal{Q}$ and $\mathcal{P}$ are both multivariate Gaussian location families, differing merely in their covariance matrices. For other exponential families, the result holds in an asymptotic sense.

Repercussions. These results have repercussions for the difference in growth optimality between various types of e-variables and e-processes, including GRO e-variables, sequential-local GRO e-variables, conditional e-variables and *Universal Inference* GRO e-processes. We discussed these repercussions in our talk.

References

[1] P. Grünwald, Rianne de Heide, and Wouter Koolen. Safe testing. *Journal of the Royal Statistical Society, Series B*, 2024. with discussion.

[2] Peter Grünwald, Tyron Lardy, Yunda Hao, Shaul K. Bar Lev, and Martijn de Jong. Optimal e-values for exponential families: the simple case. *arXiv preprint arXiv:2404.19465*, 2024.

[3] W. Koolen and P. Grünwald. Log-optimal anytime-valid e-values. *International Journal of Approximate Reasoning*, 2021. Festschrift for G. Shafer's 75th Birthday.

[4] Qiang Jonathan Li. *Estimation of mixture models*. Yale University, 1999.

[5] Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 2023.

E-power and improvements for e-tests

Ruodu Wang

The simplest criterion to quantify the power of an e-variable E is through its growth rate under an alternative probability measure Q, defined as $\mathbb{E}^Q[\log E]$. This idea goes back to [2], and it is studied by [4], [1] and [7] in detail. The quantity $\mathbb{E}^Q[\log E]$ is called the *e-power* of E by [5].

The main justification behind e-power is built on the fact that e-variables for sequential data are often multiplicative, and hence $\mathbb{E}^Q[\log E_1]$ measures the asymptotic growth rate of the partial product of iid e-variables $E_1, E_2, \dots$ under Q, due to the Law of Large Numbers.

It is notable that this definition of e-power also has some deficiencies, such as being not well-defined for some e-variables, and inconsistency with natural intuition for some e-processes in special cases.

To identify justified notion of e-power, we can use an axiomatic approach from decision theory. Let $\mathcal{X}$ be the set of all bounded nonnegative random variables, representing potential e-variables (the fact that we need to work with bounded e-values is a limitation of the e-power). For a function $\Pi : \mathcal{X} \to [-\infty, \infty]$, the following five properties are relevant.

P1 Law-invariance: $\Pi(E)$ is determined by the distribution of E under Q.

P2 Strict monotonicity: $\Pi(E_1) \leqslant \Pi(E_2)$ if $E_1 \leqslant E_2$, and $\Pi(E_1) < \Pi(E_2)$ if $Q(E_1 < E_2) = 1$.

P3 Multiplicative invariance: $\Pi(E_1) > \Pi(E_2) \implies \Pi(EE_1) > \Pi(EE_2)$ for E independent of E_1, E_2 under Q.

P4 Consistency: For $E_1, E_2, \dots$, iid under Q with $\Pi(E_1) > 0$,

$$Q\left(\prod_{k=1}^{n} E_k > \frac{1}{\alpha}\right) \to 1 \text{ as } n \to \infty \text{ for all } \alpha \in (0,1).$$

P5 Symmetry: $\Pi(E^{-1}) = -\Pi(E)$ if $E^{-1} \in \mathcal{X}$.

It can be shown, based on a recent result of [3], that a function $\Pi : \mathcal{X} \to [-\infty, \infty]$ satisfies P1-P5 if and only if there exists a strictly increasing and symmetric function f such that

$$\Pi(E) = f(\mathbb{E}^Q[\log E]) \text{ for all } E \in \mathcal{X}.$$

Therefore, the e-power is uniquely determined by the five properties introduced above. Whether these five properties are natural and desirable, or whether there are better alternatives, can be debated, and should be carefully studied.

Another important issue related to the power of e-values in hypothesis testing is the choice of the rejection threshold t. That is, we reject the null hypothesis when $E \geqslant t$ is observed for an e-variable E. This threshold is by default set to $t = 1/\alpha$ for a type-I error control at α, and this is based on Markov's inequality. This threshold can be wasteful in practical applications but cannot be improved without further assumptions. We show how this threshold can be improved under additional distributional assumptions on the e-values. For instance, the threshold can be improved to $t = 1/(2\alpha)$ when we know that the e-variable has a decreasing or unimodal density; the factor of 2 in case of the decreasing density was earlier obtained by [6] in a slightly weaker result than our statement. It can be approximately improved to $t = 1/(e\alpha)$ when we know that the log-transformed e-variable has a decreasing or unimodal–symmetric density. These improvements can help to enhance the power of testing via e-values, if good knowledge about the distribution of the e-variable is available. We can also find methods to boost e-values in the e-BH procedure, leading to potentially more discoveries under some assumptions while controlling the false discovery rate.

References

[1] Grünwald, P., de Heide, R. and Koolen, W. M. (2024). Safe testing. *Journal of the Royal Statistical Society, Series B*, forthcoming.

[2] Kelly, J. L. (1956). A new interpretation of information rate. *Bell System Technical Journal*, **35**(4), 917–926.

[3] Mu, X., Pomatto, L., Strack, P. and Tamuz, O. (2024). Monotone additive statistics. *Econometrica*, forthcoming.

[4] Shafer, G. (2021). The language of betting as a strategy for statistical and scientific communication. *Journal of the Royal Statistical Society, Series A*, **184**(2), 407–431.

[5] Vovk, V. and Wang, R. (2024). Nonparametric e-tests of symmetry. *New England Journal of Statistics in Data Science*, forthcoming.

[6] Wang, R. (2024). Testing with p*-values: Between p-values, mid p-values, and e-values. *Bernoulli*, **30**(2), 1313–1346.

[7] Waudby-Smith, I. and Ramdas, A. (2024). Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B*, **86**(1), 1–27.

Three new directions in e-statistics

Aaditya Ramdas

This talk described three new and underexplored directions in game-theoretic statistics. These are (A) the role of randomization, (B) the utility of matrix e-processes and (C) the problem of sequential change detection.

(A) The first advance that used external randomization came in the form of the "uniformly randomized Markov's inequality" [1], which uniformly improves on Markov's inequality. Since (deterministic) tests based on e-values always employ Markov's inequality (or its generalization to nonnegative supermartingales, Ville's inequality), the randomized improvement yields strictly more powerful tests. However, it is not the case that one must move from e-values to tests in order to observe the benefits of randomization. [2] describe the concept of "stochastic rounding of e-values to a grid", which transforms e-values to (randomized) e-values that only take values on a predefined grid. While stochastic rounding hurts e-power, it improves power when employed with other multiple testing procedures like the e-Benjamini-Hochberg procedure [3], and thus can be seen as a tool to trade e-power for power.

(B) The concept of a composite nonnegative supermartingale (and their generalization, e-processes) has been particularly central for game-theoretic statistics. One can ask whether an appropriate generalization exists beyond the scalar setting. For matrices, [4] studies the concept of a positive semidefinite (psd) supermartingale, which are a sequence of square symmetric psd matrices that satisfy the supermartingale constraint in the usual psd sense. While matrix martingales have been studied, these often reduce to an elementwise martingale claim, but with the psd constraint, supermartingales have a richer structure. We prove optional stopping theorems, a matrix Ville's inequality, and demonstrate promising applications to matrix testing problems like sequential covariance testing.

(C) While the e-statistics literature has focused primarily on testing and estimation, it is of interest to extend the developed techniques to other related areas such as change detection. In a first attempt to do this systematically, [5] developed the notion of an e-detector, which controls the "average run length" (frequency of false alarms, an analog of type-1 error) nonasymptotically at a predefined error level, for composite classes of pre-change distributions. E-detectors can be built by summing e-processes started at consecutive times. Thus, e-detectors are really a sophisticated reduction from sequential change detection to sequential testing, generalizing and improving an old reduction by Lorden. We demonstrated some new practical applications, as well as nontrivial bounds on the detection delay.

References

[1] A. Ramdas, T. Manole, *Randomized and exchangeable improvements of Markov's, Chebyshev's and Chernoff's inequalities*, arXiv:2304.02611 (2023).

[2] Z. Xu, A. Ramdas, *More powerful multiple testing under dependence via randomization*, arXiv:2305.11126 (2023).

[3] R. Wang, A. Ramdas, *False discovery rate control with e-values*, Journal of the Royal Statistical Society, Series B (2022).

[4] H. Wang, A. Ramdas, *Positive semidefinite supermartingales and randomized matrix concentration inequalities*, arXiv:2401.15567 (2024).

[5] J. Shin, A. Rinaldo, A. Ramdas, *E-detectors: a nonparametric framework for sequential change detection*, New England Journal on Statistics and Data Science (2023).

Sequential model confidence sets

Johanna Ziegel

(joint work with Sebastian Arnold, Georgios Gavrilopoulos, Benedikt Schulz)

In most prediction and estimation situations, scientists consider various statistical models for the same problem, and naturally want to select amongst the best. Hansen, Lunde and Nason [3] provide a powerful solution to this problem by the so-called model confidence set (MCS), a subset of the original set of available models that contains the best models with a given level of confidence. More precisely, given a finite set of models $\mathcal{M}_0 = \{1, \ldots, m\}$, and a confidence level $\alpha \in (0,1)$, they construct a set $\widehat{\mathcal{M}} \subset \mathcal{M}_0$ with

$$\mathsf{P}(\mathcal{M}_\star \subset \widehat{\mathcal{M}}) \geqslant 1 - \alpha,$$

where $\mathcal{M}_\star$ is the set of superior models with respect to a chosen expected loss L. Importantly, model confidence sets respect the underlying selection uncertainty by being flexible in size.

However, the MCS construction presuppose a fixed sample size which stands in contrast to the fact that model comparison and forecast evaluation are often inherently sequential tasks where new data is collected sequentially and where the decision to continue or conclude a study or analysis may depend on the previous outcomes.

We extend model confidence sets sequentially over time by relying on sequential testing methods. This is challenging since the set of superior objects will also become time dependent, and there are several natural ways to define it. To this end, let $d_{ij,t} = L_{i,t} - L_{j,t}$, $i, j \in \mathcal{M}_0$ denote the loss difference of model i and j at time point t, and define $\mu_{ij,t} = \mathsf{E}(d_{ij,t} \mid \mathcal{F}_{t-1})$, where $\mathcal{F}_{t-1}$ is the σ-algebra of available information at time point $t-1$. Finally, let $\Delta_{ij,t} = (1/t)\sum_{s=1}^{t} \mu_{ij,s}$.

We follow [5, 4, 2] by defining the superior models in terms of the (average) conditional expected loss differences. Specifically, we consider *strongly superior objects*

$$\mathcal{M}_t^{\mathrm{s},\star} = \left\{ i \in \mathcal{M}_0 \mid \mu_{ij,r} \leqslant 0 \text{ for all } r \leqslant t, \text{ for all } j \in \mathcal{M}_0 \right\},$$

uniformly weakly superior objects

$$\mathcal{M}_t^{\mathrm{uw},\star} = \left\{ i \in \mathcal{M}_0 \mid \Delta_{ij,r} \leqslant 0 \text{ for all } j \in \mathcal{M}_0, \text{ for all } r \leqslant t \right\},$$

and *weakly superior objects*

$$\mathcal{M}_t^{\mathrm{w},\star} = \left\{ i \in \mathcal{M}_0 \mid \Delta_{ij,t} \leqslant 0 \text{ for all } j \in \mathcal{M}_0 \right\}.$$

Clearly, $\mathcal{M}_t^{\mathrm{s},\star} \subseteq \mathcal{M}_t^{\mathrm{uw},\star} \subseteq \mathcal{M}_t^{\mathrm{w},\star}$. Furthermore, $\mathcal{M}_t^{\mathrm{w},\star} \neq \varnothing$ for all t, whereas $\mathcal{M}_t^{\mathrm{s},\star}$ and $\mathcal{M}_t^{\mathrm{uw},\star}$ are decreasing sequences of sets and may become empty from some time point on.

Let $(\mathcal{M}_t^\star)_t \subseteq \mathcal{M}_0$ be the targeted sequence of superior objects. We call $(\widehat{\mathcal{M}}_t)_t \subseteq \mathcal{M}_0$ *sequential model confidence sets* at level α if

$$\mathsf{P}\left(\forall t : \mathcal{M}_t^\star \subseteq \widehat{\mathcal{M}}_t\right) \geqslant 1 - \alpha.$$

In [1], we provide possible constructions of valid sequential model confidence sets building on the work of [4, 2]. Furthermore, the performance of our proposals is investigated in simulation studies and their usefulness is illustrated in two case studies on predictions of Covid-19 related deaths, and of wind gusts over Germany.

References

[1] S. Arnold, G. Gavrilopoulos, B. Schulz and J. Ziegel, *Sequential model confidence sets*, Preprint, arXiv:2404.18678.

[2] Y. J. Choe and A. Ramdas, *Comparing sequential forecasters*, Operations Research, to appear.

[3] P. R. Hansen, A. Lunde, and J. M. Nason, *The model confidence set*, Econometrica **79** (2011), 453–497.

[4] A. Henzi and J. F. Ziegel, *Valid sequential inference on probability forecast performance*, Biometrika **109** (2022), 647–663.

[5] T. L. Lai, S. T. Gross and D. B. Shen, *Evaluating probability forecasts*, Annals of Statistics **39** (2011), 2356–2382.

Reporter: Muriel F. Pérez-Ortiz

Participants

Dr. Shubhada Agrawal
H. Milton Stewart School of Industrial
and Systems Engineering,
Georgia Institute of Technology
765 Ferst Drive NW
Atlanta, GA 30332-0205
UNITED STATES

Dr. Rianne de Heide
Vrije Universiteit Amsterdam
Boelelaan 1111
1081 HV Amsterdam
NETHERLANDS

Prof. Dr. Thorsten Dickhaus
Fachbereich 3
Mathematik und Informatik
Universität Bremen
Bibliothekstr. 5
Postfach 330440
28334 Bremen
GERMANY

Prof. Dr. Timo Dimitriadis
Alfred-Weber-Institute for Economics
Universität Heidelberg
Bergheimer Str. 58
69115 Heidelberg
GERMANY

Prof. Dr. Rina Foygel Barber
Department of Statistics
The University of Chicago
5747 S. Ellis Avenue
Chicago, IL 60637-1514
UNITED STATES

Dr. Rafael Frongillo
Dept. of Computer Science
University of Colorado at Boulder
Boulder, CO 80309-0526
UNITED STATES

Dr. Peter Grünwald
Centrum Wiskunde & Informatica
Postbus 94079
1098 XG Amsterdam
NETHERLANDS

Dr. Nikolaos Ignatiadis
Department of Statistics & Data Science
Institute, University of Chicago
5735 South Ellis Avenue
60637 Chicago
UNITED STATES

Parnian Kassraie
Department of Computer Science
ETH Zürich
8050 Zürich
SWITZERLAND

Prof. Dr. Wouter Koolen
Centrum Wiskunde & Informatica
Science Park 123
1098 XG Amsterdam
NETHERLANDS

Prof. Dr. Martin Larsson
Department of Mathematical Sciences
Carnegie Mellon University
Pittsburgh, PA 15213
UNITED STATES

Michael Lindon
Netflix
888 Broadway
New York, NY 10003
UNITED STATES

Prof. Dr. Ryan Martin
Department of Statistics
North Carolina State University
Raleigh, NC 27695-8205
UNITED STATES

Dr. Muriel Pérez
Eindhoven University of Technology
De Zaale
P.O. Box 513
5600 MB Eindhoven
NETHERLANDS

Dr. Aaditya Ramdas
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
UNITED STATES

Dr. Zhimei Ren
Department of Statistics and Data
Science,
University of Pennsylvania
Philadelphia, PA 19104
UNITED STATES

Prof. Dr. Johannes Ruf
Department of Mathematics
The London School of Economics
and Political Science
10 Houghton Street
London WC2A 2AE
UNITED KINGDOM

Prof. Dr. Glenn Shafer
Rutgers University
1 Washington Park
07102 Newark, NJ 07102
UNITED STATES

Dr. Hongjian Shi
Technische Universität München
TUM School of Computation,
Information and Technology
Boltzmannstraße 3
85748 Garching bei München
GERMANY

Prof. Dr. Vladimir Vovk
Department of Computer Science
Royal Holloway, University of London
TW20 0EX Egham
UNITED KINGDOM

Prof. Dr. Ruodu Wang
Department of Statistics
University of Waterloo
Waterloo ON N2L 3G1
CANADA

Ian Waudby-Smith
Carnegie Mellon University
Pittsburgh, PA 15213-3890
UNITED STATES

Prof. Dr. Johanna Ziegel
ETH Zürich
Seminar für Statistik
Rämistrasse 101
8092 Zürich
SWITZERLAND