Testing for concept shift online

Vladimir Vovk

практические выводы
теории вероятностей
могут быть обоснованы
в качестве следствий
гипотез о *предельной*
при данных ограничениях
сложности изучаемых явлений

On-line Compression Modelling Project (New Series)

Working Paper #31

First posted December 28, 2020. Last revised January 25, 2021.

Project web site:
http://alrw.net

Abstract

This paper continues study of exchangeability martingales, i.e., processes that are martingales under any exchangeable distribution for the observations. Such processes can be used for detecting violations of the IID assumption, which is commonly made in machine learning. Violations of the IID assumption are sometimes referred to as dataset shift, and dataset shift is sometimes subdivided into concept shift, covariate shift, etc. Our primary interest is in concept shift, but we will also discuss exchangeability martingales that decompose perfectly into two components one of which detects concept shift and the other detects what we call label shift. Our methods will be based on techniques of conformal prediction.

Contents

1 Introduction

The most standard way of testing statistical hypotheses is batch testing: we try to reject a given null hypothesis based on a batch of data. The alternative approach of online testing (see, e.g., [12] or [11]) consists in constructing a nonnegative process that is a martingale under the null hypothesis. The ratio of the current value of such a process to its initial value can be interpreted as the amount of evidence found against the null hypothesis.

The standard assumption in machine learning is the (general) IID assumption, sometimes referred to (especially in older literature) as the assumption of randomness: the observations are assumed to be independent and identically distributed, but nothing is assumed about the probability measure generating a single observation. Interestingly, there exist processes, *exchangeability martingales*, that are martingales under the IID assumption; they can be constructed (see, e.g., [16, Section 7.1] or [15]) using the method of conformal prediction [16, Chapter 2].

Deviations from the IID assumption have become a popular topic of research in machine learning under the name of dataset shift [8, 9]; in my terminology I will follow mostly [8]. Analysing general dataset shift is usually regarded as too challenging a problem, and researchers concentrate on restricted versions, with restrictions imposed on marginal or conditional probabilities associated with the probability measure generating a single observation. Different restrictions are appropriate for different kinds of learning problems.

In this paper we consider problems of classification, in which random observations (X, Y) consist of objects X and labels Y, the latter taking a finite number of possible values. We will be interested in $Y \to X$ *domains*, in the terminology of [3], in which the objects are causally dependent on the labels. Under the IID assumption, the consecutive pairs (X, Y) have the same probability distribution P. There is a *dataset shift* if P in fact changes between observations. Let us say that there is a *label shift* if the marginal distribution P_Y of Y under P changes. Finally, there is a *concept shift* if the conditional distribution $P_{X|Y}$ of X given Y changes. Later in this paper we will adopt a wider understanding of a label shift.

As an example, suppose we are interested in the differential diagnosis between cold, flu, and Covid-19 given a set of symptoms. Under a pure label shift, the properties of the three diseases do not change (there is no concept shift), and only their prevalence changes, perhaps due to epidemics and pandemics. Under a concept shift, one or more of the diseases change leading to different symptoms. Examples are new variants of Covid-19 and new strains of flu that appear every year.

In general, exchangeability martingales may detect both label shift and concept shift. In some cases we might not be interested in label shift and only be interested in concept shift (or, perhaps less commonly, vice versa). The goal of this paper is to develop and start investigating exchangeability martingales targeting only concept shift. It would be ideal to decompose the amount of evidence found by an exchangeability martingale for dataset shift into two com-

ponents, one reflecting the amount of evidence found for concept shift and the other reflecting the amount of evidence found for label shift. Such decomposable martingales are our secondary object of study.

New exchangeability martingales and their simple theoretical properties will be the topic of Section 2, and in Section 3 they will be applied to the well-known USPS dataset. The results reported in the latter section suggest that the exchangeability martingales constructed for this dataset in [16, Section 7.1] are dominated (and greatly improved) by an exchangeability martingale decomposable into a product of an exchangeability martingale for detecting concept shift and an exchangeability martingale for detecting label shift.

The most obvious application of exchangeability martingales is to help in deciding when to retrain predictors, as discussed in [15]. We should be particularly worried about the changes that invalidate ROC analysis, which is the case of concept shift in a $Y \to X$ domain [3, 17]. Our exchangeability martingales for concept shift are designed to detect such dangerous changes.

In the context of conformal prediction, concept shift in $Y \to X$ domains requires retraining label-conditional predictors [16, Section 4.5]. For connection between label-conditional predictors and ROC analysis, see [1, Section 2.7].

2 Theory

For a detailed review of conformal prediction see, e.g., [16], but in this section I will mainly follow [1, Chapters 1 and 2] (for the generation of conformal p-values) and [15] (for gambling against those p-values).

As mentioned earlier, we consider *observations* $z = (x, y)$ that consist of two components, the *object* x and the *label* y. Let $\mathbf{X}$ be the measurable space of all possible objects, and $\mathbf{Y}$ be the set of all possible labels. Set $\mathbf{Z} := \mathbf{X} \times \mathbf{Y}$; this is our *observation space*. We are interested in classification and so always assume $|\mathbf{Y}| < \infty$; $\mathbf{Y}$ is always equipped with the discrete σ-algebra.

A *conformity measure* A is a function that maps any finite sequence $(z_1, \dots, z_n) \in \mathbf{Z}^n$ of observations of any length $n \in \{1, 2, \dots\}$ to a sequence $(\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$ of real numbers of the same length that is *equivariant* in the following sense: for any $n \in \{1, 2, \dots\}$, any permutation $\pi : \{1, \dots, n\} \to \{1, \dots, n\}$, and any sequences $(z_1, \dots, z_n) \in \mathbf{Z}^n$ and $(\alpha, \dots, \alpha_n) \in \mathbb{R}^n$,

$$(\alpha_1, \dots, \alpha_n) = A(z_1, \dots, z_n) \Longrightarrow \left(\alpha_{\pi(1)}, \dots, \alpha_{\pi(n)}\right) = A\left(z_{\pi(1)}, \dots, z_{\pi(n)}\right). \tag{1}$$

In our experiments in Section 3 we will only use conformity measures, but in theory we are also interested in the following generalization. A *label-conditional conformity measure* A is a function that maps any finite sequence $(z_1, \dots, z_n) \in \mathbf{Z}^n$ of observations of any length $n \in \{1, 2, \dots\}$ to a sequence $(\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$ of real numbers of the same length that is *label-conditionally equivariant*: for any $n \in \{1, 2, \dots\}$, any permutation $\pi : \{1, \dots, n\} \to \{1, \dots, n\}$, and any sequences

$(z_1, \dots, z_n) = ((x_1, y_1), \dots, (x_n, y_n)) \in \mathbf{Z}^n$ and $(\alpha, \dots, \alpha_n) \in \mathbb{R}^n$,

$$\left.\begin{array}{c} y_1 = y_{\pi(1)}, \dots, y_n = y_{\pi(n)} \\ (\alpha_1, \dots, \alpha_n) = A(z_1, \dots, z_n) \end{array}\right\} \Longrightarrow \left(\alpha_{\pi(1)}, \dots, \alpha_{\pi(n)}\right) = A\left(z_{\pi(1)}, \dots, z_{\pi(n)}\right).$$

In other words, we only require (1) to hold for the permutations that leave the labels intact.

The *label-conditional conformal transducer* associated with a label-conditional conformity measure A is the function p defined by

$$p(z_1, \dots, z_n, \tau) := \frac{|\{i : y_i = y_n \wedge \alpha_i < \alpha_n\}| + \tau\,|\{i : y_i = y_n \wedge \alpha_i = \alpha_n\}|}{|\{i : y_i = y_n\}|}, \quad (2)$$

where i ranges over $1, \dots, n$, $z_i = (x_i, y_i)$ for all $i \in \{1, \dots, n\}$,

$$(\alpha_1, \dots, \alpha_n) := A(z_1, \dots, z_n), \quad (3)$$

and $\tau \in [0, 1]$. The values (2) will be referred to as *p-values*. If the label-conditional conformity measure A is in fact a conformity measure, we will say that the label-conditional conformal transducer p associated with it is *simple*.

Let $Z_1, Z_2, \dots$ be a sequence of random observations, i.e., random elements whose domain is a fixed probability space with probability measure $\mathbb{P}$ and which take values in the observation space $\mathbf{Z}$. Each random observation Z_n is a pair $Z_n = (X_n, Y_n)$, where X_n is a random object and Y_n is a random label.

Let us say that the random sequence of observations $Z_1, Z_2, \dots$ is *label-conditional exchangeable* if, for any $n \in \{1, 2, \dots\}$, any sequence $(y_1, \dots, y_n) \in \mathbf{Y}^n$, any sequence of measurable sets $E_1, \dots, E_n$ in $\mathbf{X}$, and any permutation $\pi : \{1, \dots, n\} \to \{1, \dots, n\}$,

$$\begin{aligned} &y_1 = y_{\pi(1)}, \dots, y_n = y_{\pi(n)} \\ &\quad \Longrightarrow \mathbb{P}(Y_1 = y_1, \dots, Y_n = y_n, X_1 \in E_1, \dots, X_n \in E_n) \\ &\quad\quad = \mathbb{P}\left(Y_1 = y_1, \dots, Y_n = y_n, X_{\pi(1)} \in E_1, \dots, X_{\pi(n)} \in E_n\right). \end{aligned}$$

This is an instance of de Finetti's [2] notion of partial exchangeability. The sequence $Z_1, Z_2, \dots$ is *exchangeable* if, for any $n \in \{1, 2, \dots\}$, any sequence of measurable sets $E_1, \dots, E_n$ in $\mathbf{Z}$, and any permutation $\pi : \{1, \dots, n\} \to \{1, \dots, n\}$,

$$\mathbb{P}(Z_1 \in E_1, \dots, Z_n \in E_n) = \mathbb{P}\left(Z_{\pi(1)} \in E_1, \dots, Z_{\pi(n)} \in E_n\right).$$

Of course, exchangeability is a stronger property than label-conditional exchangeability.

Proposition 1. *If the sequence of random observations $Z_1, Z_2, \dots$ is label-conditional exchangeable, $(\tau_1, \tau_2, \dots)$ is an independent sequence of independent random variables each distributed uniformly in $[0, 1]$, and p is a label-conditional conformal transducer, the sequence of random p-values*

$$P_n := p(Z_1, \dots, Z_n, \tau_n), \quad n = 1, 2, \dots, \quad (4)$$

is distributed uniformly in $[0, 1]^\infty$.

For a proof of Proposition 1, see [16, Section 8.7] (Proposition 1 is a special case of Theorem 8.1 in [16]).

If Z is a measurable space, Z^* stands for the set of all finite sequences of elements of Z (equipped with the natural σ-algebra). It includes the empty sequence $\Box$. A *betting martingale* is a measurable function $F : [0,1]^* \to [0,\infty]$ such that $F(\Box) = 1$ and, for each sequence $(u_1, \dots, u_{n-1}) \in [0,1]^{n-1}$ for any $n \in \{1, 2, \dots\}$,

$$\int_0^1 F(u_1, \dots, u_{n-1}, u) \, \mathrm{d}u = F(u_1, \dots, u_{n-1}). \tag{5}$$

(The three unusual features of this definition are that betting martingales are required to be nonnegative, start from 1, and are allowed to take value ∞.) The *test martingale* associated with the betting martingale F and a sequence $(P_1, P_2, \dots)$ uniformly distributed in $[0,1]^\infty$ (the *input p-values*) is the sequence of random variables

$$S_n = F(P_1, \dots, P_n), \quad n = 0, 1, \dots. \tag{6}$$

The sequence $(S_n)_{n=0,1,\dots}$ is a nonnegative martingale, in the usual sense of probability theory [14, Definition 7.1.1], in its own filtration $\mathcal{F}_n := \sigma(S_1, \dots, S_n)$ or the filtration $\mathcal{F}_n := \sigma(P_1, \dots, P_n)$ generated by the input p-values. Intuitively, this martingale describes the evolution of the capital of a player who gambles against the hypothesis that the input p-values are distributed uniformly and independently.

In this paper we will be interested in several classes of test martingales. The *label-conditional conformal martingales* are defined as the test martingales associated with any betting martingale F and a sequence $(P_1, P_2, \dots)$ defined by (4) (under the conditions of Proposition 1) as the input p-values.

Label-conditional conformal martingales are main topic of this paper. They detect concept shift. It was shown, once again, in [16, Section 7.1] that the USPS dataset is non-exchangeable, and in Section 3 we will explore sources of this lack of exchangeability.

Remark 1. It is important that our exchangeability martingales for detecting concept shift can be used in situations where the labels are so far from being IID that it would be unusual to talk about label shift. Discussion of label shift usually presuppose at least approximate independence of labels. Suppose a sequence of hand-written characters $x_1, x_2, \dots$ comes from a user writing a letter. The objects x_n are matrices of pixels and the corresponding labels y_n take values in the set $\{a, b, \dots\}$. Different instances of the same character, say "a", may well be exchangeable among themselves (even conditionally on knowing the full text of the letter), whereas the text itself will be far from IID; for example, "q" will be almost invariably followed by "u" if the letter is in English. For discussions of such partial exchangeability, see, e.g., [2], [10], and [16, Section 8.4].

In the rest of this section we will look for possible explanations of the difference between the amount of evidence found against concept shift and against exchangeability. We will see that in some situation the amount of evidence found against exchangeability decomposes into two components:

- the amount of evidence found for concept shift;
- the amount of evidence found for label shift.

In these situations the second component can be said to explain the difference.

A *label conformity measure* A is a conformity measure that satisfies, additionally, the following property: for any finite sequence $(z_1, \dots, z_n) \in \mathbf{Z}^n$ of observations of any length $n \in \{1, 2, \dots\}$, any sequence $(\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$ of real numbers of the same length, and any $i, j \in \{1, \dots, n\}$,

$$\left.\begin{array}{c} y_i = y_j \\ (\alpha_1, \dots, \alpha_n) = A(z_1, \dots, z_n) \end{array}\right\} \Longrightarrow \alpha_i = \alpha_j, \tag{7}$$

where y_i and y_j are the labels in z_i and z_j, respectively. In other words, it assigns conformity scores only to the labels rather than to the full observations. (Notice that the requirement of equivariance only ensures (7) with "$z_i = z_j$" in place of "$y_i = y_j$".) The *conformal transducer* associated with a conformity measure A outputs the p-values

$$p'(z_1, \dots, z_n, \tau) := \frac{|\{i : \alpha_i < \alpha_n\}| + \tau |\{\alpha_i = \alpha_n\}|}{n}, \tag{8}$$

where $i \in \{1, \dots, n\}$, $\alpha_1, \dots, \alpha_n$ are defined by (3), and $\tau \in [0, 1]$. We will say that p' is a *label conformal transducer* if A is a label conformity measure.

Our method of decomposing exchangeability martingales will be based on the following result (version of Theorem 8.1 in [16]). Its proof is given in Appendix A.1.

Theorem 2. *If the sequence of random observations $Z_1, Z_2, \dots$ is exchangeable, $(\tau_1, \tau_2, \dots)$ and $(\tau'_1, \tau'_2, \dots)$ are independent (between themselves and of the observations) sequences distributed uniformly in $[0, 1]^\infty$, p is a simple label-conditional conformal transducer, and p' is a label conformal transducer, the interleaved sequence of random p-values $P_1, P'_1, P_2, P'_2, \dots$, where*

$$P_n := p(Z_1, \dots, Z_n, \tau_n), \quad P'_n := p'(Z_1, \dots, Z_n, \tau'_n),$$

is distributed uniformly in $[0, 1]^\infty$.

A *conformal martingale* is defined to be the test martingale associated (via (6), where F is a betting martingale) with a conformal transducer. If the underlying conformity measure is a label conformity measure, the conformal martingale will be called a *label conformal martingale*.

We will say that a label-conditional conformal martingale is *simple* if its underlying label-conditional conformal transducer is simple.

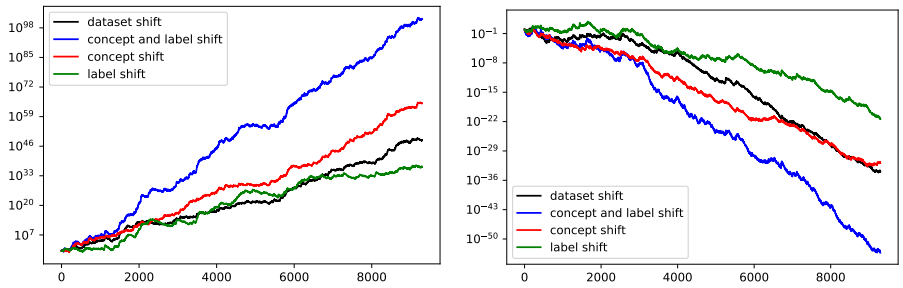

Figure 1: Four exchangeability martingales for the Simple Jumper based on the nearest-neighbour conformity measure applied to the USPS dataset (left panel) and its random permutation (right panel)

Having the stream of random p-values $P_1, P'_1, P_2, P'_2, \dots$ produced as in Theorem 2, we can define two derivative exchangeability martingales: a label-conditional conformal martingale associated with $P_1, P_2, \dots$ and a label conformal martingale associated with $P'_1, P'_2, \dots$. (There are no restrictions on the underlying betting martingales.)

Corollary 3. *The product of a simple label-conditional conformal martingale and a label conformal martingale with independent randomizations (i.e., their sequences of random numbers τ) is an exchangeability martingale.*

Such product exchangeability martingales decompose perfectly into components for detecting concept shift and label shift. For a short proof of this corollary, see Appendix A.2.

3 Experiments

The dataset used in our experiment is the well-known USPS dataset of handwritten digits [16, Appendix B.1], which is known to be non-exchangeable. The objects x_n are 16×16 matrices with entries in $[-1, 1]$ (representing pixel intensities), and the labels y_n are elements of $\{0, \dots, 9\}$; overall there are 9298 labelled images (obtained by merging the original training set of 7291 and test set of 2007). This dataset is clearly in the $Y \to X$ domain (the intended digit in the writer's mind causes the resulting matrix of pixels, not vice versa).

Online methods for testing the exchangeability of the USPS dataset are described in [16, Section 7.1]; those methods are modified and greatly improved here. Figure 1 shows trajectories of four specific exchangeability martingales, which will be defined in the next few paragraphs. It plots $n \in \{0, \dots, 9298\}$ vs the values S_n of the four exchangeability martingales with initial value 1 after processing the first n observations. The martingales are randomized, but their trajectories on the USPS dataset do not depend much on the seed used in the

random number generator. The final values on the USPS dataset (the left panel of Figure 1) are huge, exceeding 10^{93} for the blue one, and show that the USPS dataset is far from exchangeable.

The conformity measure used in Figure 1 is of the nearest-neighbour type: namely, the conformity score α_i of the ith observation (x_i, y_i) in a sequence $(x_1, y_1), \ldots, (x_n, y_n)$ is defined as

$$\alpha_i := \min_{j \in \{1,\ldots,n\}} \|x_i - x_j\|, \tag{9}$$

where $\|\ldots\|$ is Euclidean norm. (Using the tangent distance in place of the Euclidean distance $\|x - x'\|$ leads to similar results for all experiments reported in this paper, unlike for the batch experiments in [15, Section 2].) Notice that the conformity measure (9) completely ignores the labels, although the alternative $\alpha_i := \min_{j \neq i : y_j = y_i} \|x_i - x_j\|$ leads to similar (usually slightly better) results.

The betting martingale used in all our experiments is the *Simple Jumper* (*SJ*), a modification of the "Sleepy Jumper" as described in [16, Section 7.1]. The SJ involves one parameter, J, which is set to $J := 0.1$. (Inevitably, there is some element of data snooping here, since this value was chosen because of its reasonable performance on the USPS dataset, but it is limited by the use of round figures.)

The main components of the SJ are two *betting functions*,

$$f_\epsilon(p) := 1 + \epsilon(p - 0.5), \quad p \in [0, 1], \tag{10}$$

where $\epsilon \in \{-1, 1\}$. For any probability measure μ on $\{-1, 1\}^\infty$ the function

$$F(u_1, \ldots, u_n) := \int \prod_{i=1}^{n} f_{\epsilon_i}(u_i) \mu(\mathrm{d}(\epsilon_1, \epsilon_2, \ldots)) \tag{11}$$

is a betting martingale. The measure μ is defined as the probability distribution of the following Markov chain with state space $\{-1, 1\}$. The initial state is $\epsilon_1 := \pm 1$ with equal probabilities. The transition function prescribes maintaining the same state with probability $1 - J$ and, with probability $1 - J$, choosing a new state from the set $\{-1, 1\}$ with equal probabilities. Notice that the betting martingale (11) is a deterministic function, even though the Markov chain is stochastic.

Remark 2. The intuition behind the betting functions (10) is that $\epsilon = -1$ corresponds to betting on small p-values, and $\epsilon = 1$ corresponds to betting on large p-values. The main difference from the Sleepy Jumper [16, p. 176] is that along with betting on small p-values now we also allow betting on large p-values. The idea of gambling against the non-uniformity of p-values in the context of conformal prediction goes back to [4]. Intuitively, the SJ tracks the best value of the parameter ϵ used to "calibrate" the p-values produced by conformal prediction into a martingale. The idea of tracking the best value of ϵ goes back to [5] ("tracking the best expert").

Each of the four exchangeability martingales in Figure 1 apart from the product (the blue martingale) is determined by three components:

- the underlying conformity measure, which is either (9) or its modification ensuring the label invariance (7);
- the transducer, which is either the label-conditional conformal transducer (2) or the conformal transducer (8); feeding the conformity measure of the previous item into this transducer we obtain a sequence of p-values;
- the betting martingale F, which in this paper is always the SJ; we feed the p-values resulting from the previous item into F, as per (6).

The black martingale in the left panel of Figure 1 uses the conformity measure (9), the conformal transducer (8), and the SJ.

The black martingale may detect any deviations from exchangeability, but in this paper we are particularly interested in concept shift. In our current context, concept shift means that, for some reason, the same digit (such as "0") starts looking different; perhaps people start writing digits differently, or the digits are scanned with different equipment. To detect concept shift, we use the same conformity measure (9), but feed it into the label-conditional conformal transducer (2); the resulting sequence of p-values is fed into the SJ, as usual. The resulting test martingale is shown in red in the left panel of Figure 1. Its final value, of the order of magnitude 10^{55}, is even more impressive than the final value of the black martingale.

There is, of course, another reason why exchangeability may be violated: we may have label shift. To detect it, we use the label conformity measure that assigns the conformity score

$$\alpha'_i := \operatorname{med}\left(\{\alpha_j : j \in \{1,\dots,n\}, y_j = y_i\}\right) \tag{12}$$

to the ith observation (x_i, y_i) in a sequence $(x_1, y_1), \dots, (x_n, y_n)$, where med stands for the median (the convention for $\operatorname{med}(\emptyset)$ does not affect the resulting p-values). In other words, we average, in the sense of median, the conformity scores for each class to ensure the requirement of invariance (7).

The label conformal martingale obtained by applying the SJ to the p-values produced by the label conformal transducer (8) applied to the conformity scores (12) is shown as the green line in the left panel of Figure 1. It is interesting that, despite the invariance restriction, the final value of the green martingale, which is more volatile than the black and red ones, is not much worse than the final value of the black martingale. The relatively high volatility of the green line stems from large values of the term $|\{\alpha_i = \alpha_n\}|$ in (8) for label conformity measures, which assign the same conformity score to all images of the same class.

According to Corollary 3, the product of a label-conditional conformal martingale and a label conformal martingale is still an exchangeability martingale. The product is shown as the blue line in the left panel of Figure 1. By construction, the blue martingale is perfectly decomposable. Its final value greatly exceeds the previous record for the USPS dataset.

Remark 3. Corollary 3 has an important condition, "with independent randomizations". It is satisfied in our experiments (if we ignore the fact that NumPy can only generate pseudorandom numbers) since all our plots are produced by a single Python program that sets the seed for its random number generator only once, at the beginning (to 0).

The blue exchangeability martingale, on the one hand, dominates the black martingale over the USPS dataset and, on the other hand, decomposes into a product of exchangeability martingales for detecting concept shift and for detecting label shift. Therefore, the red and green pair in the left panel of Figure 1 appears to be a significant improvement over the black martingale.

Of course, when the USPS dataset is permuted, as in the right panel of Figure 1, these successful martingales start quickly losing capital (much quicker than the more cautious Sleepy Jumper used in [16]: see [16, Figure 7.8] and, especially, [16, Figure 7.9]).

4 Conclusion

We have seen that the existing methods of constructing exchangeability martingales can be adapted to detecting concept shift. Perfectly decomposable exchangeability martingales turned out to be surprisingly successful on the USPS dataset of handwritten digits.

This paper concentrated on concept shift in $Y \to X$ classification domains. It is clear, however, that the same methods are applicable, verbatim, when the observations z_i take values in any measurable space and y_i are no longer the labels but defined as $f(z_i)$ for a function f taking finitely many values. For example, y_i can be an important feature of the object in z_i that we do not wish to model, but we wish our analysis to be conditional on it (e.g., $y_i \in \{\text{male}, \text{female}\}$ can be a feature).

Acknowledgments

Many thanks to Alex Gammerman, Ivan Petej, and Ilia Nouretdinov for numerous useful discussions. This work has been supported by Amazon (project "Conformal martingales for change-point detection", including generous funding for computational experiments) and Stena Line.

References

[1] Vineeth N. Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk, editors. *Conformal Prediction for Reliable Machine Learning: Theory, Adaptations, and Applications*. Elsevier, Amsterdam, 2014.

[2] Bruno de Finetti. *Sur la condition d'équivalence partielle*, volume 739 of *Actualités Scientifiques et Industrielles*. Hermann, Paris, 1938. An English translation is included in [6] as Chapter 9.

[3] Tom Fawcett and Peter A. Flach. A response to Webb and Ting's *On the application of ROC analysis to predict classification performance under varying class distributions*. *Machine Learning*, 58:33–38, 2005.

[4] Valentina Fedorova, Alex Gammerman, Ilia Nouretdinov, and Vladimir Vovk. Plug-in martingales for testing exchangeability on-line, On-line Compression Modelling project (New Series), `http://alrw.net`, Working Paper 4, April 2012. Conference version: ICML 2012.

[5] Mark Herbster and Manfred K. Warmuth. Tracking the best expert. *Machine Learning*, 32:151–178, 1998.

[6] Richard C. Jeffrey, editor. *Studies in Inductive Logic and Probability*, volume 2. University of California Press, Berkeley, 1980.

[7] Fredrik Johansson et al. `mpmath`: *a Python library for arbitrary-precision floating-point arithmetic (version 1.1.0)*, December 2018. `http://mpmath.org/`.

[8] Jose G. Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V. Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45:521–530, 2012.

[9] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence, editors. *Dataset Shift in Machine Learning*. MIT Press, Cambridge, MA, 2009.

[10] Daniil Ryabko. Pattern recognition for conditionally independent data. *Journal of Machine Learning Research*, 7:645–664, 2006.

[11] Glenn Shafer. The language of betting as a strategy for statistical and scientific communication. Technical Report arXiv:1903.06991 [math.ST], arXiv.org e-Print archive, March 2019. To appear as discussion paper in the *Journal of the Royal Statistical Society A*; read in September 2020.

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

[13] Albert N. Shiryaev. *Probability-1*. Springer, New York, third edition, 2016.

[14] Albert N. Shiryaev. *Probability-2*. Springer, New York, third edition, 2019.

[15] Vladimir Vovk. Testing randomness online, On-line Compression Modelling project (New Series), `http://alrw.net`, Working Paper 24, June 2019.

[16] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

[17] Geoffrey I. Webb and Kai Ming Ting. On the application of ROC analysis to predict classification performance under varying class distributions. *Machine Learning*, 58:25–32, 2005.

A Some proofs

A.1 Proof of Theorem 2

It suffices to prove, for a fixed horizon $N \in \{1, 2, \dots\}$, that the random p-values $P'_1, P_1, \dots, P'_N, P_N$ are distributed independently and uniformly in $[0, 1]$ (see, e.g., [16, Section 8.2]). Let us fix such an N.

The rest of this appendix is a modification of [16, Section 8.7]. First an informal argument. Imagine that the data sequence $Z_1, \dots, Z_n$ is generated in two steps: first a random multiset $\{\!|Z_1, \dots, Z_n|\!\}$ and then its random ordering. Already the second step ensures that $(P_1, P'_1, \dots, P_N, P'_N)$ are distributed uniformly in $[0, 1]^{2N}$ (even conditionally on $\{\!|Z_1, \dots, Z_n|\!\}$). This can be demonstrated using the following backward argument. Ignoring borderline effects, P'_N is uniformly distributed in $[0, 1]$ (at least approximately). When Y_N is disclosed, P'_N will be settled. Given what we already know, the distribution of P_N will be uniform. When X_N is disclosed, P_N will be settled. Now the distribution of P'_{N-1} given what we already know is uniform, etc.

For the formal proof, we will need the following σ-algebras. Let $\mathcal{G}_n$, $n = 0, \dots, N$, be the σ-algebra

$$\mathcal{G}_n := \sigma\left(\{\!|Z_1, \dots, Z_n|\!\}, Z_{n+1}, \tau_{n+1}, \tau'_{n+1}, \dots, Z_N, \tau_N, \tau'_N\right)$$

generated by the multiset $\{\!|Z_1, \dots, Z_n|\!\}$ and the other random elements listed in the parentheses. Let $\mathcal{G}'_n$, $n = 1, \dots, N$, be the σ-algebra $\sigma(\mathcal{G}_n, Y_n, \tau'_n)$ generated by $\mathcal{G}_n$, the label Y_n of the nth observation, and the random number τ'_n.

The following two lemmas (analogues of [16, Lemma 8.8]) say that

$$\begin{array}{ccccccccccccc} & & P'_N & & P_N & & P'_{N-1} & \dots & P_2 & & P'_1 & & P_1 \\ \mathcal{G}_N & \subseteq & \mathcal{G}'_N & \subseteq & \mathcal{G}_{N-1} & \subseteq & \mathcal{G}'_{N-1} & \subseteq \cdots \subseteq & \mathcal{G}_1 & \subseteq & \mathcal{G}'_1 & \subseteq & \mathcal{G}_0 \end{array}$$

is a stochastic sequence essentially in the usual sense of probability theory [14, Section 7.1.2]: in the second row we have a finite filtration, and the random variables in the first row are measurable w.r. to the σ-algebras directly below them.

Lemma 4. *For any trial $n = 1, \dots, N$, P'_n is $\mathcal{G}'_n$-measurable.*

Proof. The random multiset of conformity scores of $Z_1, \dots, Z_n$ is $\mathcal{G}_n$-measurable, and so, according to the definition (8) and the invariance requirement (7), P'_n is $\mathcal{G}'_n$-measurable. □

Lemma 5. *For any trial $n = 1, \dots, N$, P_n is $\mathcal{G}_{n-1}$-measurable.*

Proof. This follows from the definition (2) and our requirement that the label-conditional conformal transducer p should be simple. □

We will also need the following analogues of [16, Lemma 8.7]. As in [16], $\mathbb{E}_{\mathcal{F}}$ stands for the conditional expectation w.r. to a σ-algebra $\mathcal{F}$.

Lemma 6. *For any trial $n = 1, \dots, N$ and any $\epsilon \in [0,1]$,*

$$\mathbb{P}_{\mathcal{G}'_n}\{P_n \le \epsilon\} = \epsilon.$$

Proof. Follow the proof of [16, Lemma 8.7]. □

Lemma 7. *For any trial $n = 1, \dots, N$ and any $\epsilon \in [0,1]$,*

$$\mathbb{P}_{\mathcal{G}_n}\{P'_n \le \epsilon\} = \epsilon.$$

Proof. Follow the proof of [16, Lemma 8.7]. □

Let us now prove the following double sequence of equalities:

$$\mathbb{P}_{\mathcal{G}'_n}\left\{P_n \le \epsilon_n, P'_{n-1} \le \epsilon'_{n-1}, P_{n-1} \le \epsilon_{n-1}, \dots, P'_1 \le \epsilon'_1, P_1 \le \epsilon_1\right\} = \epsilon_n \epsilon'_{n-1} \epsilon_{n-1} \dots \epsilon'_1 \epsilon_1 \quad (13)$$

and

$$\mathbb{P}_{\mathcal{G}_n}\{P'_n \le \epsilon'_n, P_n \le \epsilon_n, \dots, P'_1 \le \epsilon'_1, P_1 \le \epsilon_1\} = \epsilon'_n \epsilon_n \dots \epsilon'_1 \epsilon_1. \quad (14)$$

We will use induction arranging these equalities into a single sequence: the equality for $\mathbb{P}_{\mathcal{G}'_1}$, the equality for $\mathbb{P}_{\mathcal{G}_1}$, the equality for $\mathbb{P}_{\mathcal{G}'_2}$, the equality for $\mathbb{P}_{\mathcal{G}_2}$, etc. The first of these equalities is a special case of Lemma 6. When proving any other of these equalities, we will assume that all the previous equalities are true.

The equality for $\mathbb{P}_{\mathcal{G}_n}$, $n \in \{1, \dots, N\}$, follows from

$$\begin{aligned}
&\mathbb{P}_{\mathcal{G}_n}\{P'_n \le \epsilon'_n, P_n \le \epsilon_n, \dots, P'_1 \le \epsilon'_1, P_1 \le \epsilon_1\} \\
&\quad = \mathbb{E}_{\mathcal{G}_n}\left(\mathbb{E}_{\mathcal{G}'_n}\left(1_{P'_n \le \epsilon'_n} 1_{P_n \le \epsilon_n} \dots 1_{P'_1 \le \epsilon'_1} 1_{P_1 \le \epsilon_1}\right)\right) \\
&\quad = \mathbb{E}_{\mathcal{G}_n}\left(1_{P'_n \le \epsilon'_n}\, \mathbb{E}_{\mathcal{G}'_n}\left(1_{P_n \le \epsilon_n} \dots 1_{P'_1 \le \epsilon'_1} 1_{P_1 \le \epsilon_1}\right)\right) \\
&\quad = \mathbb{E}_{\mathcal{G}_n}\left(1_{P'_n \le \epsilon'_n} \epsilon_n \dots \epsilon'_1 \epsilon_1\right) = \epsilon'_n \epsilon_n \dots \epsilon'_1 \epsilon_1.
\end{aligned}$$

The first equality is just the tower property of conditional expectations. The second equality follows from Lemma 4. The third equality follows from the inductive assumption, namely (13). The last equality follows from Lemma 7.

The equality for $\mathbb{P}_{\mathcal{G}'_n}$, $n \in \{2, \dots, N\}$, follows from

$$\begin{aligned}
&\mathbb{P}_{\mathcal{G}'_n}\left\{P_n \le \epsilon_n, P'_{n-1} \le \epsilon'_{n-1}, P_{n-1} \le \epsilon_{n-1}, \dots, P'_1 \le \epsilon'_1, P_1 \le \epsilon_1\right\} \\
&\quad = \mathbb{E}_{\mathcal{G}'_n}\left(\mathbb{E}_{\mathcal{G}_{n-1}}\left(1_{P_n \le \epsilon_n} 1_{P'_{n-1} \le \epsilon'_{n-1}} 1_{P_{n-1} \le \epsilon_{n-1}} \dots 1_{P'_1 \le \epsilon'_1} 1_{P_1 \le \epsilon_1}\right)\right) \\
&\quad = \mathbb{E}_{\mathcal{G}'_n}\left(1_{P_n \le \epsilon_n}\, \mathbb{E}_{\mathcal{G}_{n-1}}\left(1_{P'_{n-1} \le \epsilon'_{n-1}} 1_{P_{n-1} \le \epsilon_{n-1}} \dots 1_{P'_1 \le \epsilon'_1} 1_{P_1 \le \epsilon_1}\right)\right) \\
&\quad = \mathbb{E}_{\mathcal{G}'_n}\left(1_{P_n \le \epsilon_n} \epsilon'_{n-1} \epsilon_{n-1} \dots \epsilon'_1 \epsilon_1\right) = \epsilon_n \epsilon'_{n-1} \epsilon_{n-1} \dots \epsilon'_1 \epsilon_1.
\end{aligned}$$

Now the second equality follows from Lemma 5. The third equality follows from the inductive assumption, namely (14) with $n-1$ in place of n. The last equality follows from Lemma 6.

Plugging $n := N$ into (14), we obtain

$$\mathbb{P}\{P_1 \le \epsilon_1, P'_1 \le \epsilon'_1, \dots, P_N \le \epsilon_N, P'_N \le \epsilon'_N\} = \epsilon_1\epsilon'_1 \dots \epsilon_N\epsilon'_N.$$

This implies the uniform distribution of $(P_1, P'_1, \dots, P_N, P'_N)$ in $[0,1]^{2N}$ (see, e.g., [13, Lemma 2.2.3]).

A.2 Proof of Corollary 3

Let the simple label-conditional conformal martingale be

$$S_n = F(P_1, \dots, P_n), \quad n = 0, 1, \dots,$$

and the label conformal martingale be

$$S'_n = F'(P'_1, \dots, P'_n), \quad n = 0, 1, \dots,$$

where F and F' are betting martingales and $P_1, P'_1, P_2, P'_2, \dots$ is a stream of p-values as in Theorem 2. Let us check that $S_nS'_n$, $n = 0, 1, \dots$, is a martingale w.r. to the filtration generated by the p-values: for any $n \in \{1, 2, \dots\}$,

$$\begin{aligned}
&\mathbb{E}_{P_1,P'_1,\dots,P_{n-1},P'_{n-1}}(S_nS'_n) \\
&\quad = \mathbb{E}_{P_1,P'_1,\dots,P_{n-1},P'_{n-1}}\left(\mathbb{E}_{P_1,P'_1,\dots,P_{n-1},P'_{n-1},P_n}(S_nS'_n)\right) \\
&\quad = \mathbb{E}_{P_1,P'_1,\dots,P_{n-1},P'_{n-1}}\left(S_n\,\mathbb{E}_{P_1,P'_1,\dots,P_{n-1},P'_{n-1},P_n}(S'_n)\right) \\
&\quad = \mathbb{E}_{P_1,P'_1,\dots,P_{n-1},P'_{n-1}}\left(S_nS'_{n-1}\right) \\
&\quad = \mathbb{E}_{P_1,P'_1,\dots,P_{n-1},P'_{n-1}}\left(S_n\right)S'_{n-1} = S_{n-1}S'_{n-1},
\end{aligned}$$

where each lower index for $\mathbb{E}$ signifies the conditioning σ-algebra (namely, the conditioning σ-algebra is generated by the listed random variables). The third and last equalities follow from (5).

B Testing the validity of exchangeability martingales

In this appendix I will discuss ways of testing the validity of exchangeability martingales. This may be useful, e.g., for debugging their computer implementations. Notice that already the right panel of Figure 1 can be interpreted as a weak validity argument.

Our argument for conformal martingales being exchangeability martingales was based on two statements:

- The uniform distribution in $[0,1]^\infty$ of the sequence of conformal p-values $(p_1, p_2, \dots)$.
- The function F used in (6) is a betting martingale.

We can test these two statements separately and jointly. In this version of the paper we will concentrate on testing the second statement for the Simple Jumper.

Let us see how we can test that a function F of the type used in (6) is a betting martingale. Checking $F(\Box) = 1$ is straightforward, and we will assume that this is satisfied. Fix $N \in \{1, 2, \dots\}$ (already $N = 1$ is sufficient for testing the martingale property). To check that $\mathbb{E}(F(P_1, \dots, P_N)) \le 1$, we can generate K independent sequences $(P_1, \dots, P_N)$ of uniformly distributed independent p-values. For each of them compute $F(P_1, \dots, P_N)$, and let F_k be the kth value, $k = 1, \dots, K$. By the (informal) law of large numbers, we expect

$$\frac{1}{K} \sum_{k=1}^{K} F_k \approx \mathbb{E}(F(P_1, \dots, P_N)) \le 1 \tag{15}$$

for a large K, and we can test whether $\mathbb{E}(F(P_1, \dots, P_N)) \le 1$ by comparing the left-hand side of (15) with 1. If the left-hand side of (15) is much greater than 1, we can reject the hypothesis of validity. We will use a large-deviation inequality (Theorem 8 below) to judge whether the difference between the left-hand side of (15) and 1 is large enough. We can apply the same method for checking the analogous inequality, $\mathbb{E}(F(P_1, \dots, P_N) \mid P_1 = p_1, \dots, P_n = p_n) \le 1$, for conditional probabilities, where $n \in \{1, \dots, N-1\}$ and $p_1, \dots, p_n \in [0, 1]$; the only difference is that now we generate K independent replicas for $(P_{n+1}, \dots, P_N)$.

The problem with applying standard large-deviation inequalities, such as Hoeffding's, Bernstein's, and Bennet's (see, e.g., [12, Chapter 3]), is that they require knowing some characteristics of the random variables F_k that we are not given, such as upper bounds on them or their variances. However, a suitable inequality can be derived using Doléans's supermartingales [12, Section 3.2]); it is natural to call this inequality Doléans's inequality.

Theorem 8 (Doléans's inequality)**.** *Let $F_1, \dots, F_K$ be independent nonnegative random variables with expected value 1, and let $\kappa_1, \dots, \kappa_M$ be constants in $(0, 1/2]$. Then, for any $\epsilon > 0$,*

$$\mathbb{P}\left(\frac{1}{K}\sum_{k=1}^{K} F_k < 1 + \min_{m=1,\dots,M}\left(\frac{\kappa_m}{K}\sum_{k=1}^{K}(F_k - 1)^2 + \frac{1}{\kappa_m K}\ln\frac{M}{\epsilon}\right)\right) \ge 1 - \epsilon. \tag{16}$$

Our proof will show that it suffices to assume that $F_1 - 1, \dots, F_K - 1$ is a martingale difference, but we will never need this generalization. Moreover, we are only interested in the case where $F_1, \dots, F_K$ are IID.

The base form of (16) is where $M = 1$, and so we have only one $\kappa = \kappa_1$:

$$\mathbb{P}\left(\frac{1}{K}\sum_{k=1}^{K} F_k < 1 + \frac{\kappa}{K}\sum_{k=1}^{K}(F_k - 1)^2 + \frac{1}{\kappa K}\ln\frac{1}{\epsilon}\right) \ge 1 - \epsilon.$$

In hindsight, after seeing the data, the optimal value of κ is the point where

$$\frac{\kappa}{K}\sum_{k=1}^{K}(F_k-1)^2+\frac{1}{\kappa K}\ln\frac{1}{\epsilon}\to\min,$$

which is

$$\kappa=\sqrt{\frac{\ln\frac{1}{\epsilon}}{\sum_{k=1}^{K}(F_k-1)^2}}.$$

If F_k are IID and have a finite second moment, the optimal κ will have the order of magnitude $K^{-1/2}$. However, the assumption of a finite second moment seems unrealistic in our context, and in our applications, we will take $\kappa_m := K^{-m/(2M)}\wedge 1/2$, $m=M,M+1,\dots,2M-1$. When $K\ge 4$ (which is the case in all our experiments), we can ignore the "$\wedge 1/2$" bit.

Proof of Theorem 8. Since, for $\kappa\in(0,1/2]$, the random process

$$\exp\left(\kappa\sum_{k=1}^{n}(F_k-1)-\kappa^2\sum_{k=1}^{n}(F_k-1)^2\right),\quad n=0,1,\dots, \tag{17}$$

is a positive supermartingale (namely, Doléans's supermartingale [12, Proposition 3.4]) with initial value 1, the expected value of the positive random variable

$$\frac{1}{M}\sum_{m=1}^{M}\exp\left(\kappa_m\sum_{k=1}^{K}(F_k-1)-\kappa_m^2\sum_{k=1}^{K}(F_k-1)^2\right)$$

is at most 1. Replacing the first $\sum$ by max and using Markov's inequality shows that

$$\frac{1}{M}\max_{m=1,\dots,M}\exp\left(\kappa_m\sum_{k=1}^{n}(F_k-1)-\kappa_m^2\sum_{k=1}^{n}(F_k-1)^2\right)<\frac{1}{\epsilon}$$

with probability at least $1-\epsilon$. The last inequality is equivalent to the inequality $\le$ in (16). □

Remark 4. Theorem 8 gives an upper bound on the average of F_k that holds with a high probability (assuming $\epsilon\ll 1$). It is clear that there are no non-trivial lower bounds: we can make the probability

$$\mathbb{P}\left(\frac{1}{K}\sum_{k=1}^{K}F_k=0\right)$$

arbitrarily small even when F_k with $\mathbb{E}(F_k)=1$ are IID.

Now let us see how these results apply to the SJ martingale. Taking $N:=100$ and $K:=10^8$ in (15), we obtain the results given in Table 1. The mean is very close to 1 and well below its upper bound given in Theorem 8 (after the $<$ sign) for $M:=5$.

N	K	mean	bound	median	quartiles
100	10^8	1.00039	1.00167	0.43347	[0.22271,0.93031]
1000	10^7	0.77123	1.31286	0.00046	[0.00005,0.00498]

Table 1: The mean $\frac{1}{k}\sum_k F_k$, its bound in (16), and the median and interquartile range of $F_1, \dots, F_K$ for two pairs (N, K).

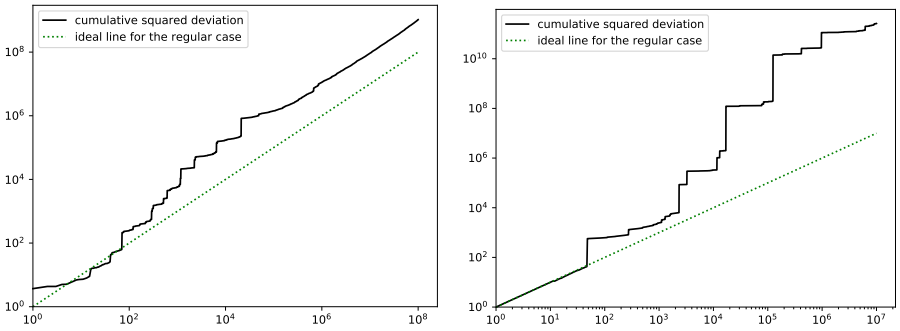

Figure 2: The *cumulative square deviation* $\sum_{k=1}^K (F_k - 1)^2$ vs K on the log scale for $N = 100$ (left panel) and $N = 1000$ (right panel)

The upper bound for $N = 100$ in Table 1 is computed as

$$\min(1.00167, 1.00409, 1.02477, 1.15611, 1.98495),$$

where the five numbers correspond to $\kappa = K^{-0.5}, K^{-0.6}, K^{-0.7}, K^{-0.8}, K^{-0.9}$. The optimal value among these κ is $K^{-0.5}$, which corresponds to the "regular" case $\sum_{k=1}^K (F_k - 1)^2 \sim K$. Figure 2 (left panel) plots $\sum_{k=1}^K (F_k - 1)^2$ vs K on the log scale for $N = 100$. We can see that the plot is above but fairly close to the diagonal.

Table 1 shows that the mean is well above the median, and even above the upper quartile. In general, our experiments demonstrate the difference between the average and typical behaviour of our exchangeability martingales in the situation when the null hypothesis of exchangeability is satisfied. Typical trajectories quickly go down, since gambling against the true null is futile. On the other, since they are martingales, the average trajectory is close to being horizontal for large K. This is illustrated in Figure 3. The trajectories shown in green are those corresponding to the seeds 0–9 of the random number generator; they can be considered to be typical. The trajectories shown in red are the 10 with the highest final values among K trajectories, where K is given in Table 1. Finally, the trajectories shown in blue are the 10 with the lowest final values. The average trajectory is nearly horizontal (as suggested by Table 1) even though typical trajectories go down.

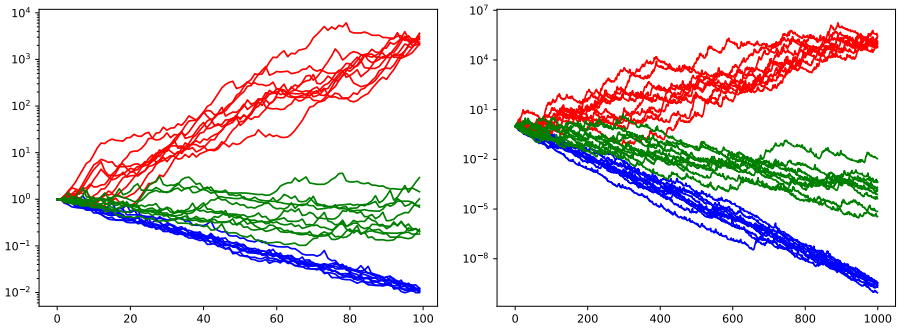

Figure 3: High, typical, and low paths of exchangeability martingales, as described in text, for $N = 100$ (left panel) and $N = 1000$ (right panel)

Table 1 and Figures 2–3 also have information for $N = 1000$ and $K = 10^7$. The mean in Table 1 is much less than 1, let alone the upper bound; the mean is still far from its limit 1. The upper bound is computed as

$$\min(9.24172, 2.65390, 1.37740, 1.31286, 2.25304),$$

the five numbers corresponding to $\kappa = K^{-0.5}, K^{-0.6}, K^{-0.7}, K^{-0.8}, K^{-0.9}$. Now the optimal value among these κ is $K^{-0.8}$, and indeed the right panel of Figure 2 shows a much more significant deviation from the diagonal than the left panel. The difference between the typical and average becomes much more pronounced for $N = 1000$.

One disadvantage of the upper bound on $\frac{1}{K}\sum_k F_k$ given in (16) is its dependence on M, the size of the grid of κs. If the grid is too crude, the first addend in

$$\frac{\kappa_m}{K}\sum_{k=1}^{K}(F_k - 1)^2 + \frac{1}{\kappa_m K}\ln\frac{M}{\epsilon}$$

is likely to be large (the grid may miss good values of κ), and making the grid too fine will inflate the second addend. A safer approach is to use $1 + X$ as an upper bound on $\frac{1}{k}\sum_k F_k$, where $X = X(F_1, \dots, F_K) \geq 0$ is the solution to the equation

$$2\int_{1/2}^{1} \exp\left(K^{1-u}X - K^{-2u}\sum_{k=1}^{K}(F_k - 1)^2\right) = \frac{1}{\epsilon}$$

(the left-hand side in increasing in X, and it is clear that there is unique solution). Since (17) is a supermartingale, we indeed have

$$\mathbb{P}\left(\frac{1}{K}\sum_{k=1}^{K} F_k < 1 + X(F_1, \dots, F_K)\right) \geq 1 - \epsilon;$$

this is a continuous version of Doléans's inequality (16).

When the discrete version (16) of Doléans's inequality is replaced by the continuous version, the bound 1.00167 in Table 1 for $N = 100$ becomes 0.00175, slightly worse. The bound 1.31286 for $N = 1000$, on the other hand, becomes slightly better, 1.27007. (To avoid an overflow error for $N = 100$ I had to use the arbitrary precision library `mpmath` [7].)