Lecture 4: Conformal hypothesis testing

Vladimir Vovk

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

Mahalanobis Memorial Lectures 2020–21
15 March, 2021

Main points of this lecture

- An interesting application of conformal prediction: the existence of exchangeability martingales (conformal test martingales).
- Exchangeability martingales can be used for detecting a point at which the IID assumption becomes violated.
- Open problem: how efficient are conformal test martingales?

Plan

Setting the problem

- Conformal prediction also allows us to test the IID model.
- Why is this important? For example, to decide when we need to retrain a prediction algorithm.
- For this, our testing procedure has to be online: the observations keep arriving sequentially, and at each point in time we want to know the amount of evidence we have found against the IID hypothesis.
- Conformal prediction is the only known method for doing that.

Gambling against randomness

- The online way of testing the IID model: a test martingale.
- We start from a capital of \$1 and gamble against each observation in such a way that the game is fair and our capital is always nonnegative.
- Formally, our capital S_n at time n satisfies $\mathbb{E}(S_n \mid S_1, \dots, S_{n-1}) = S_{n-1}$ (with $S_0 = 1$) under the IID model (i.e., under any probability measure in the IID model). Such S is called a test martingale.
- How to get a test martingale (valid under the IID model, which is massive): gamble against the p-values output by conformal prediction (which is easy).
- Remember Ville's inequality.

Details of betting

- A betting martingale is a measurable function $F : [0,1]^* \to [0,\infty]$ such that $F(\square) = 1$ and, for each sequence $(u_1, \dots, u_{n-1}) \in [0,1]^{n-1}$ for any $n \in \{1, 2, \dots\}$,

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

- 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

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

Details of randomized p-values

- Let $(x_1, y_1), (x_2, y_2), \ldots$ be an input stream of data.
- For each $n = 1, 2, \ldots,$
  - compute the conformity scores

$$\alpha_i := A(\{z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_n)\}, z_i), \quad i = 1, \ldots, n,$$

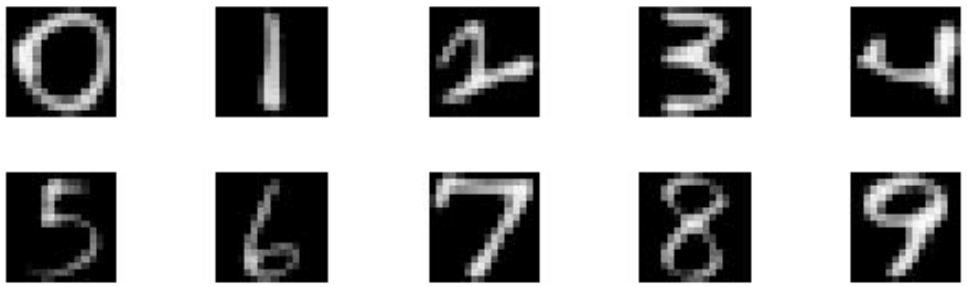
    with the dependence on n suppressed;
  - compute the (randomized) p-value

$$p_n := \frac{|\{i : \alpha_i < \alpha_n\}| + \tau_n |\{i : \alpha_i = \alpha_n\}|}{n},$$

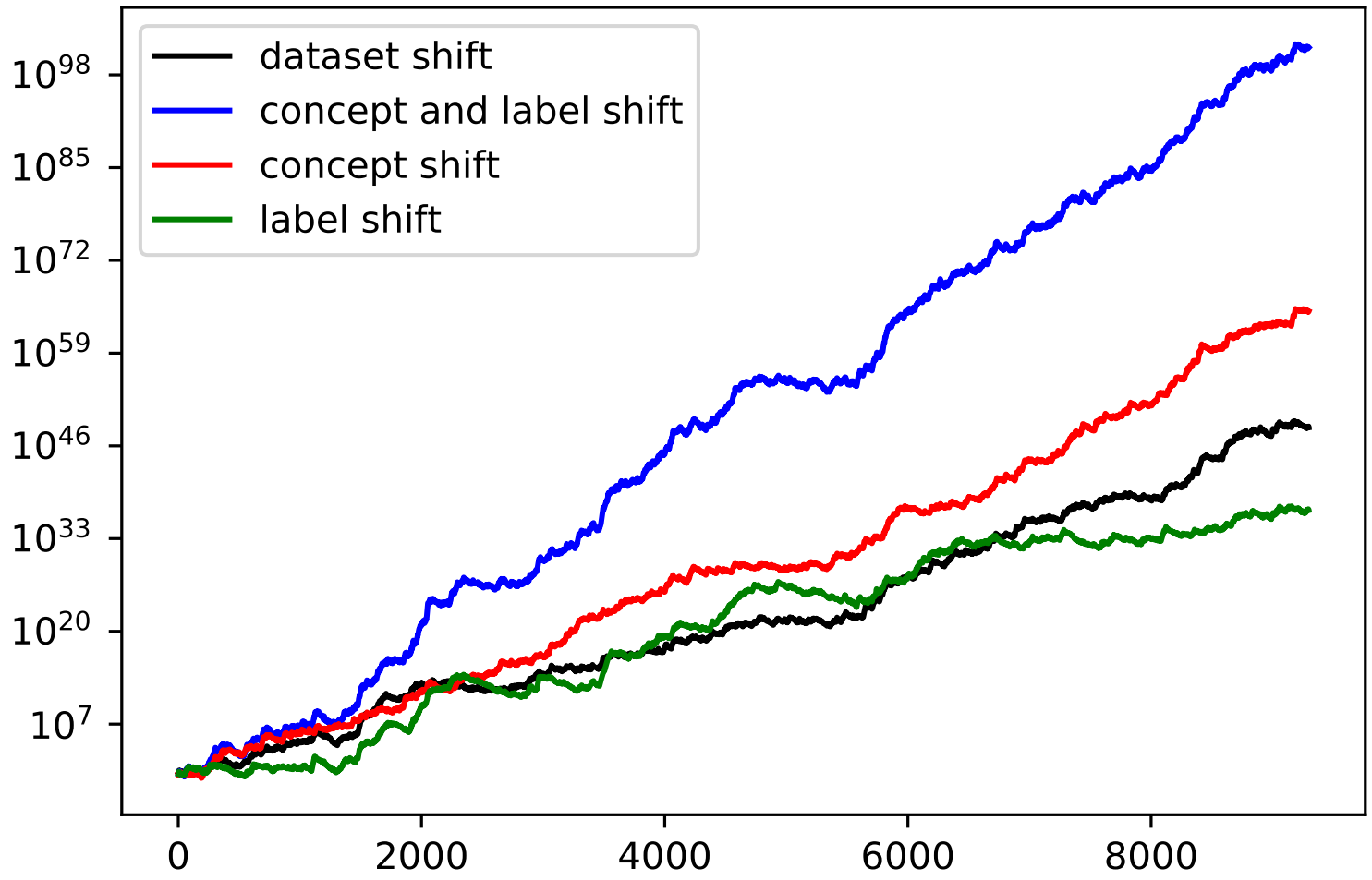
    i ranging over $1, \ldots, n$ and $\tau_n \in [0, 1]$ being IID uniformly distributed numbers.
- Theorem (intuitive backward argument): $p_n \in [0, 1]$ are IID and uniformly distributed.

USPS dataset

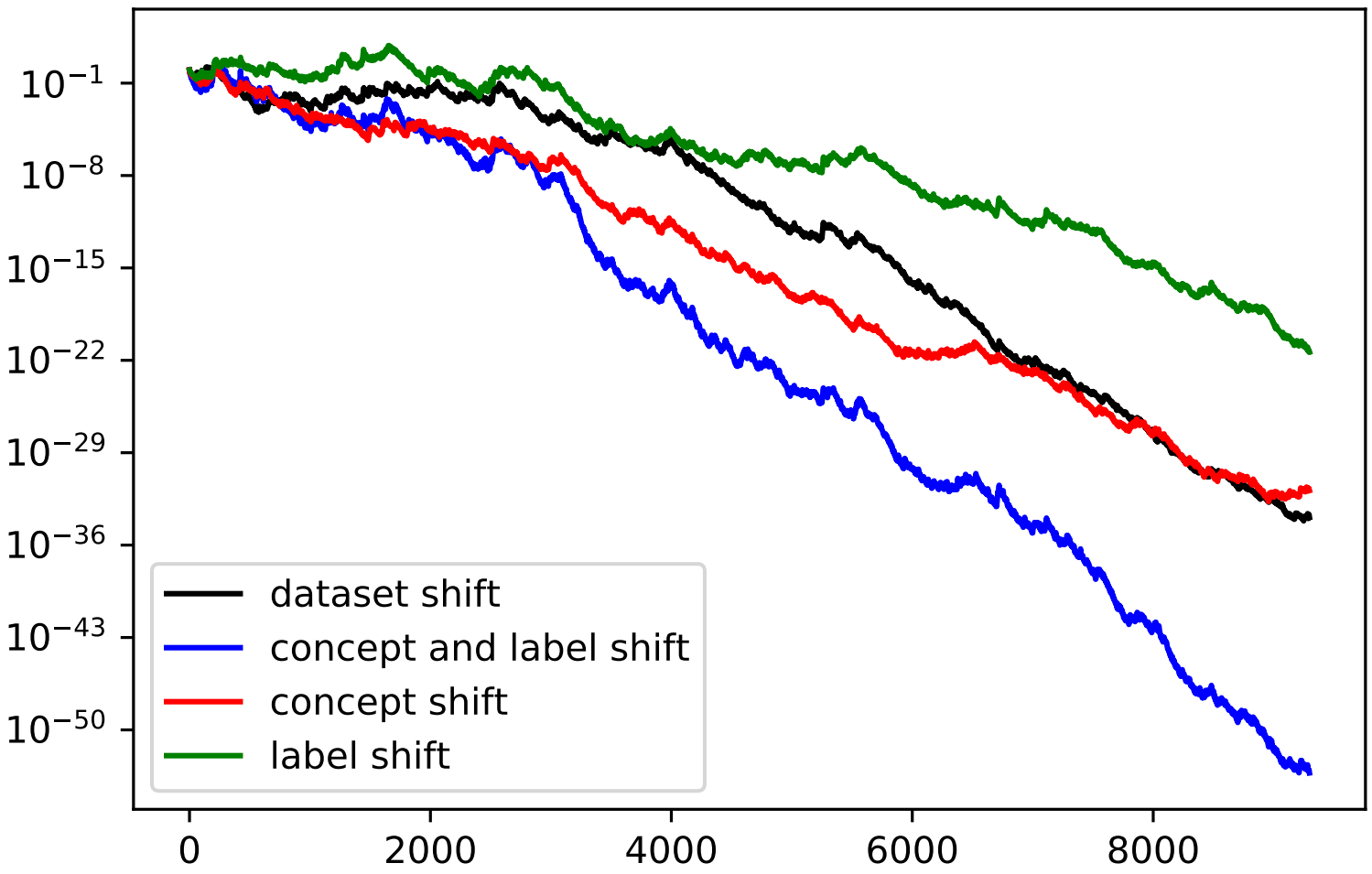
- A dataset of 9298 hand-written digits, popular in machine learning and known to be somewhat non-IID.
- Objects: 16×16 matrices of pixels (see below).
- Labels: 0 to 9.
- Can we detect non-exchangeability using our methods?



Gambling against the USPS dataset (black line)



Gambling against a permuted USPS dataset



How we can achieve this (1)

- The conformity measure is of the nearest-neighbour type: the conformity score of (x, y) as compared with $\zeta = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ is

$$\alpha = A(\zeta, (x, y)) := \min_{i \in \{1,\ldots,n\}} \|x_i - x\|,$$

  where $\|\ldots\|$ is Euclidean norm.
- The betting martingale: Simple Jumper (SJ). The main components of the SJ are two calibrators,

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

  where $\epsilon \in \{-1, 1\}$.

How we can achieve this (2)

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

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

  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 J, choosing a new state from the set $\{-1,1\}$ with equal probabilities; $J := 0.1$.

How we can achieve this (3)

- Notice: our betting martingale is a deterministic function, even though the Markov chain is stochastic.
- The intuition behind the calibrators is that
  - $\epsilon = -1$ corresponds to betting on small p-values,
  - and $\epsilon = 1$ corresponds to betting on large p-values.
- Sometimes $\epsilon = 0$ is also useful (not betting), but not in this case.

Do non-trivial exchangeability martingales exist?

- It is interesting that exchangeability martingales, despite their impressive performance on many datasets, barely exist.
- Let $\mathcal{F}_n$ be the σ-algebra generated by the first n observations.
- The usual definition of a martingale requires $\mathbb{E}(S_n \mid \mathcal{F}_{n-1}) = S_{n-1}$.
- But since this holds under any distribution for the nth observation, we must have $S_n = S_{n-1}$.
- So S must be a constant!

Why they exist: two reasons

- One reason is that our filtration is not $\mathcal{F}_n$.
- It is the filtration generated by the martingale itself, or the larger filtration generated by the input p-values.
- By itself this is not enough (think of the binary case).
- Another reason: the input p-values are randomized.
- Even a tiny amount of randomization is often sufficient!
  - Other examples: defensive forecasting (we can replace the continuity used in lecture 1 by slight randomization); differential privacy.

Is all dataset shift interesting?

- In machine learning, the moment when the IID assumption becomes violated is known as dataset shift.
- Not all kinds of dataset shift are considered dangerous.
- Sometimes, an irrelevant marginal distribution changes.
- In other cases, we can have a “concept shift”, which is more important.

Two kinds of concept shift

- There are two kinds of datasets, $X \to Y$ and $Y \to X$: sometimes X causes Y ("regular case", e.g., weather prediction), and in other cases Y causes X (medicine, hand-written digits).
- For the former, concept shift means a change in the conditional distribution $Y \mid X$, whereas for the latter, it means a change in the conditional distribution $X \mid Y$.

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.

USPS dataset again

- The USPS dataset is obviously $Y \to X$.
- For it we can have a label shift (the distribution of Y changes) and a concept shift (the distribution of $X \mid Y$ changes).
- Interestingly, we can gamble against the label shift and concept shift separately obtaining two exchangeability martingales.
- Their product will also be an exchangeability martingale (gambling against dataset shift).
- But in this lecture I will discuss only concept shift.

Partial exchangeability

- It will be possible to use our exchangeability martingales for detecting concept shift for testing "partial exchangeability".
- 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.

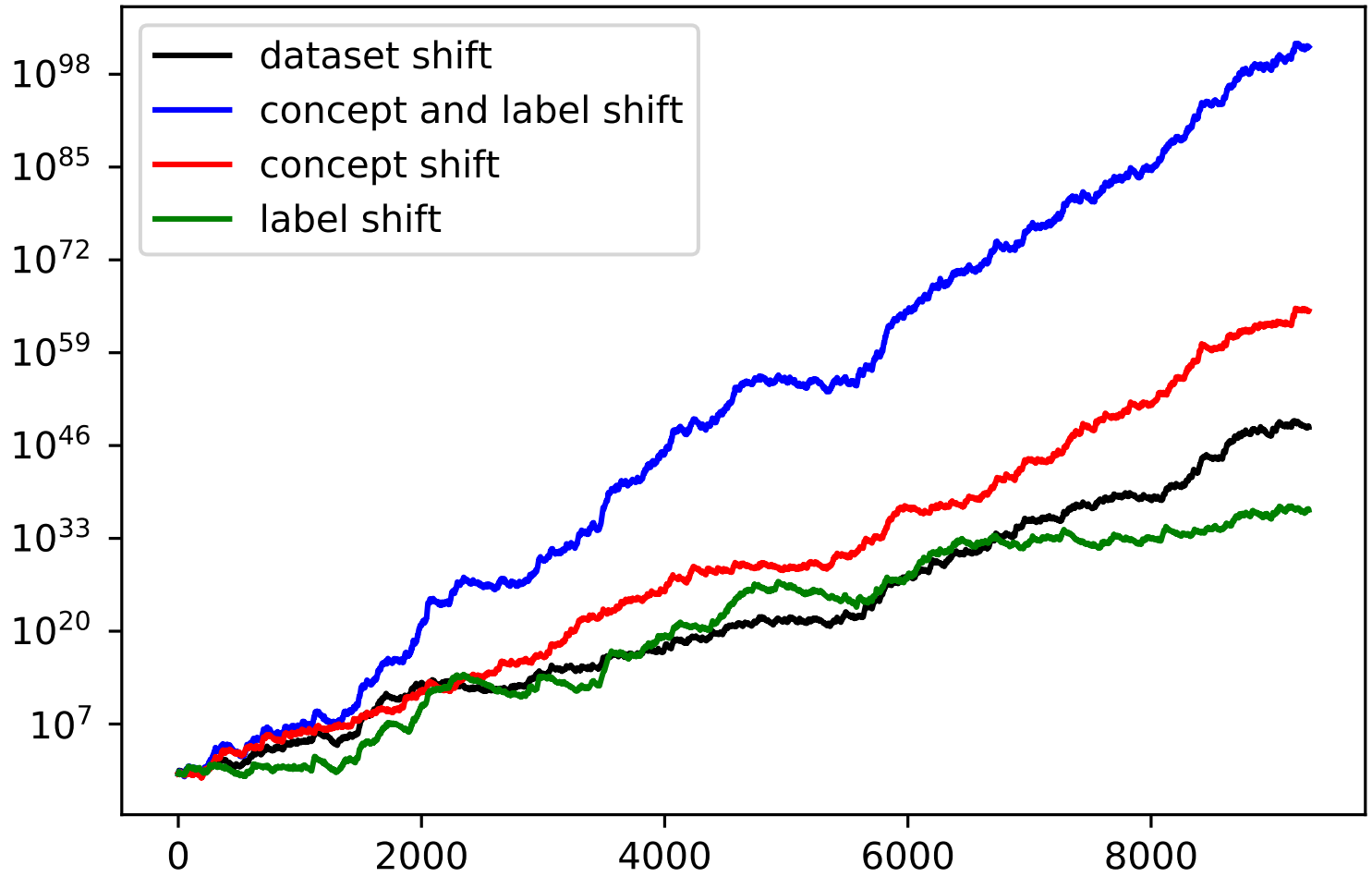
Label-conditional p-values

- For the input stream of data $z_i = (x_i, y_i)$, $i = 1, 2, \ldots$, compute α_i as before, but the p-values are label-conditional:

$$p_n := \frac{|\{i : y_i = y_n \wedge \alpha_i < \alpha_n\}| + \tau_n |\{i : y_i = y_n \wedge \alpha_i = \alpha_n\}|}{|\{i : y_i = y_n\}|}.$$

- Let us now assume that the observations are partially exchangeable (conditionally on $y_1, y_2, \ldots$, the objects of the same class are exchangeable); formal definition omitted.
- Theorem: $p_n \in [0, 1]$ are IID and uniformly distributed.

Same conformity measure and betting martingale for the USPS dataset (red line)



Plan

Experimental setting

- As an example, let's consider the Wine Quality dataset.
- It consists of two parts, 4898 white wines and 1599 red wines.
- We randomly choose a subset of 1599 white wines and refer to it as test set 0, and the remaining white wines (randomly permuted) will be our training set.
- All 1599 red wines form our test set 1; therefore we have two test sets of equal sizes.

Two scenarios (1)

- To detect a possible change point in the test set, the training set of 3299 white wines is randomly split into three folds of nearly equal sizes, about 1100.
- We use each fold in turn as the calibration set and the remaining folds as the training set proper.
- For each fold $k \in \{1, 2, 3\}$ we train a prediction algorithm on the training set proper and run an exchangeability martingale (based on the resulting model) on the calibration set followed by a test set (one of the two).

Two scenarios (2)

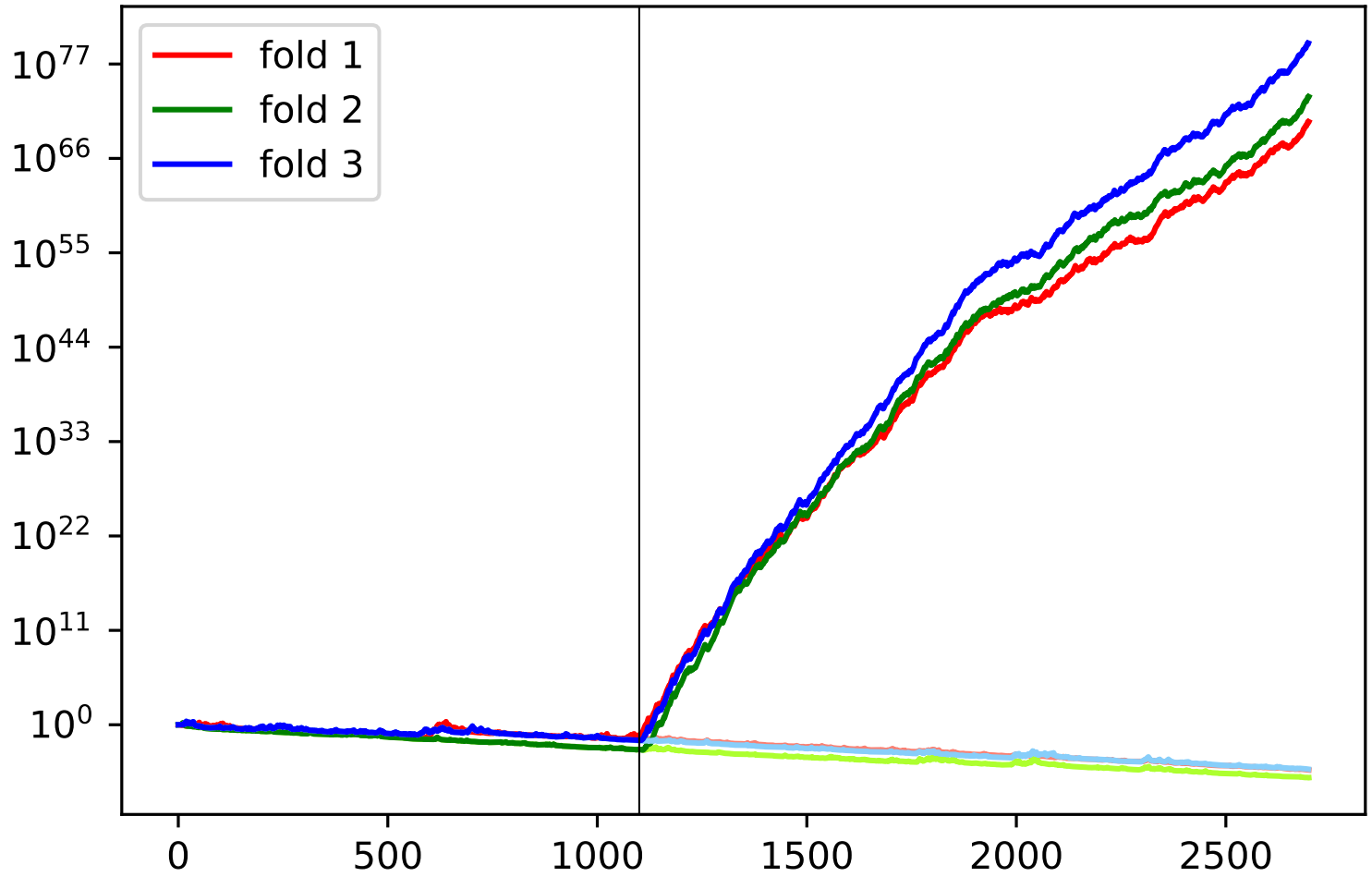
- This way we obtain three paths, plots of the values of the exchangeability martingales vs time.
- We have two scenarios: scenario 0 uses test set 0, and scenario 1 uses test set 1; thus we have 6 paths overall.
- Let us use the conformity measure

$$\alpha := y - \hat{y},$$

where $\hat{y}$ is the prediction for the label y of the object x produced by the Random Forest (in `scikit-learn`) found from the training set proper.

- The betting martingale is the Simple Jumper (but now we have $\epsilon \in \{-1, 0, 1\}$ with equal probabilities).

The martingale paths



Procedures for change detection

- We can use the Ville procedure: retrain when a conformal test martingale on the previous slide exceeds, say, 100.
- A disadvantage of the Ville procedure for deciding when to retrain: when the IID model is valid but we keep gambling against it (in vain), our capital goes down (typically exponentially), and it becomes more difficult to recover it before exceeding the threshold when a change does happen.
- We can apply the standard CUSUM and Shiryaev–Roberts procedures for change detection on top of conformal test martingales to obtain procedures for raising an alarm when the IID model ceases to be true.

CUSUM procedure

- Let S be an exchangeability martingale that never takes value 0 (typical case).
- The CUSUM procedure (Page, 1954) raises an alarm at the time

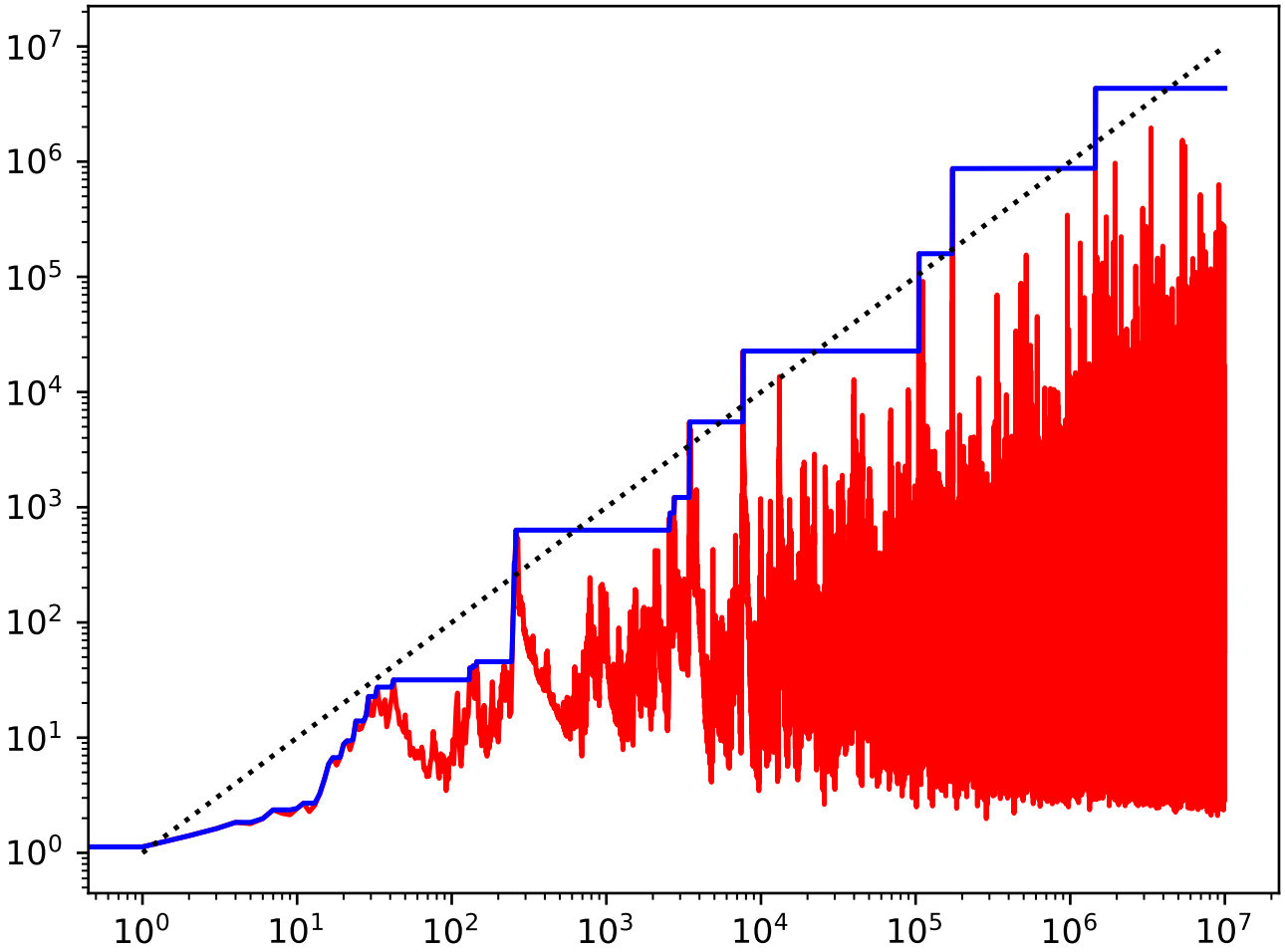
$$\min\{n \mid \gamma_n \geq c\}, \text{ where } \gamma_n := \max_{i=0,\dots,n-1} \frac{S_n}{S_i}$$

and $c > 1$ is the parameter.

CUSUM procedure for deciding when to retrain

- Suppose we have an idea of the target lifespan of our predictor, say we would like it to process 10^6 observations.
- Let us retrain when γ_n hits a suitable barrier.
- The next figure shows in red the maximum of 100 simulated CUSUM paths.
- It suggests that a reasonable barrier is a straight line with slope 1 in the loglog representation.
- In the original (x, y)-axes the barrier has the equation $y = cx$.
- The blue line: the records.

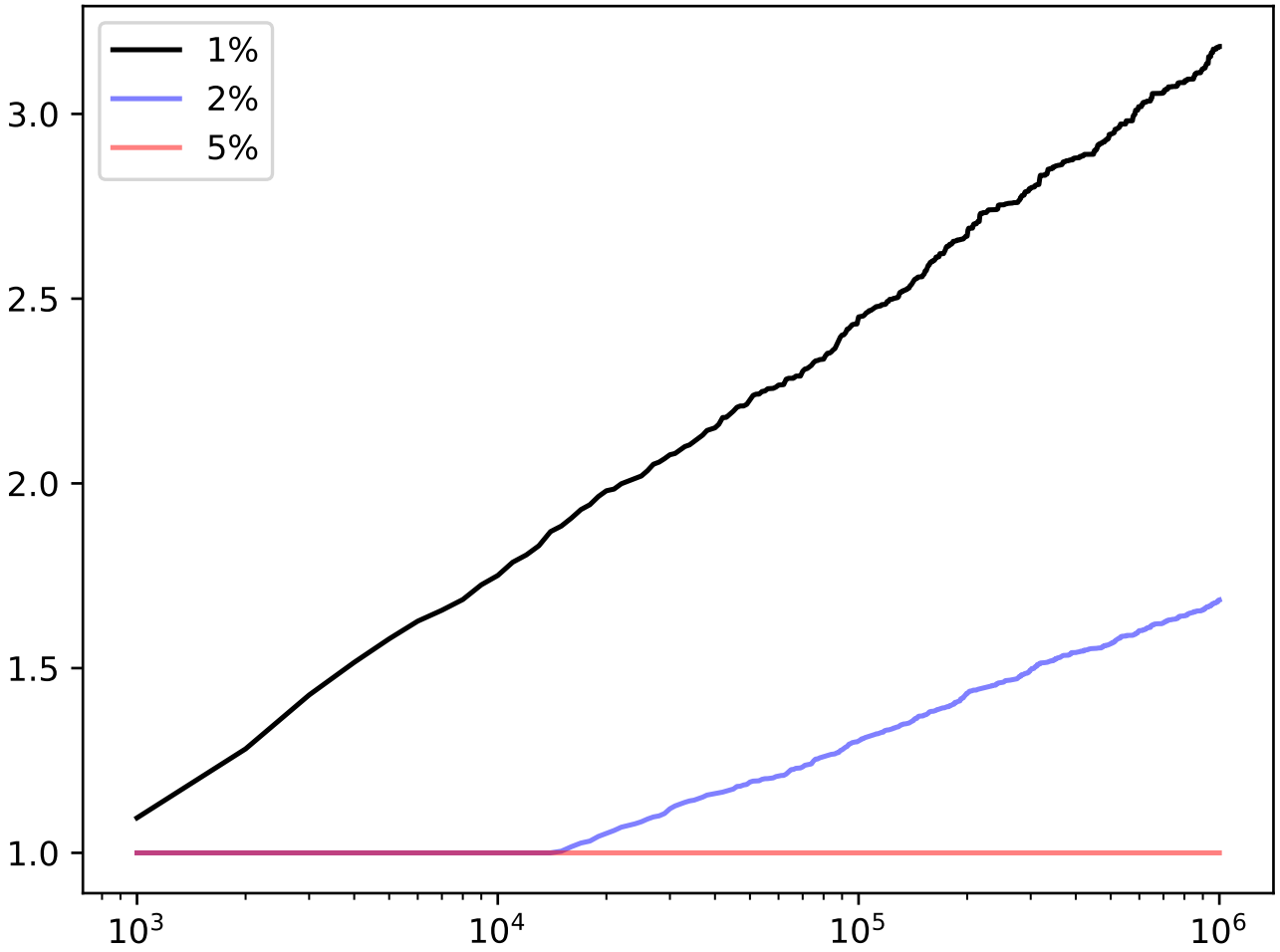
Max of 100 CUSUM paths



Choosing the slope of the barrier

- A path of the CUSUM statistic γ_n over $n = 1, \dots, N$ will trigger an alarm for a barrier $y = cx$ if $\gamma_n \geq cn$ for some $n \leq N$.
- Let us generate a large number K ($K = 10^5$ in the next figure and table) of paths of the CUSUM statistic in the ideal setting (uniform p-values).
- For each N, let c_N be the number such that 1% of the K paths trigger an alarm (a false one) for the barrier $y = c_N x$.
- The black line in the next figure: c_N vs N for $N \in \{1000, 2000, \dots, 10^6\}$.
- It looks like a straight line (like the similar blue and red lines, except that they cannot go under $y = 1$).

Slopes in the ideal setting



Confidence intervals for the probability of false alarm

c	alarms	99.9% confidence interval
3.2	988	[0.89%, 1.10%]
3.3	958	[0.86%, 1.06%]
3.4	930	[0.83%, 1.03%]
3.5	901	[0.81%, 1.00%]
4	793	[0.70%, 0.89%]
5	622	[0.54%, 0.71%]

Choosing the slope of the barrier

- For example, if the target lifespan is $N = 10^6$, we can see from the figure that $c_N \approx 3.2$ for 1%.
- To find a suitable barrier, we need a confidence interval for the probability of a false alarm at a suitable confidence level that is completely inside $[0, 1\%]$.
- According to the table, we can take the barrier $y = 4x$ (for the confidence level 99.9%).

Plan

Setting the problem

- How good are conformal test martingales at detecting violations of the IID assumption?
- Are they a universal way of testing that assumption?
- We can claim that they are, if, for any way W of testing the IID assumption, we can construct a conformal test martingale that rejects IID whenever W rejects IID.
- What is "any way of testing the assumption"?

Glenn Shafer (2007).
From Cournot's principle to market efficiency.
Augustin Cournot: Modelling Economics (edited by Jean-Philippe Touffut), pages 55–95.

Using Cournot's principle

- The canonical way of testing a null hypothesis: choose an event E of a small probability under the hypothesis.
- If the event then happens, we are entitled to reject the null hypothesis.
- This is the basis of statistical hypothesis testing (both for Fisher and for Neyman–Pearson).
- In the history of probability, it is known as Cournot's principle.

Idea

- Suppose we have an event E of a small probability under the null hypothesis.
- Can we say that there exists a conformal test martingale that takes a large value when E happens?
- If "yes", conformal test martingales are a universal means of testing.
- But what should we take as our null hypothesis, IID or exchangeability?

IID and exchangeability probability (1)

- The relation between IID and exchangeability was of great interest to Kolmogorov in his research programme announced in his 1963 paper in *Sankhyā* (lecture 1).
- For simplicity, and as a starting point, he only considered finite binary sequences. Let's follow him.
- Let $\Omega := \{0,1\}^N$.

IID and exchangeability probability (2)

- The upper IID probability of $E \subseteq \Omega$ is

$$\mathbb{P}^{\text{iid}}(E) := \sup_{p \in [0,1]} B_p^N(E),$$

  B_p being the Bernoulli measures on $\{0, 1\}$.
- The upper exchangeability probability of $E \subseteq \Omega$ is

$$\mathbb{P}^{\text{exch}}(E) := \sup_P P(E),$$

  P being the exchangeable probability measures on Ω.
- The corresponding lower probabilities are $1 - \mathbb{P}^{\text{iid}}(\Omega \setminus E)$ and $1 - \mathbb{P}^{\text{exch}}(\Omega \setminus E)$, but we will never need them.

What happens for infinite sequences

- We can give the analogous definitions for infinite binary sequences.
- For infinite sequences, the relation between $\mathbb{P}^{\text{iid}}$ and $\mathbb{P}^{\text{exch}}$ is trivial: $\mathbb{P}^{\text{iid}} = \mathbb{P}^{\text{exch}}$.
- This follows from de Finetti's theorem: every exchangeable probability measure is a mixture of IID.
- In the finite case, this is no longer true, but the difference is minor in a crude sense (Kolmogorov).

And for finite sequences (1)

Proposition

For any $E \subseteq \Omega$,

$$\mathbb{P}^{\text{iid}}(E) \le \mathbb{P}^{\text{exch}}(E) \le 1.5\sqrt{N}\,\mathbb{P}^{\text{iid}}(E).$$

- This is straightforward (follows from Stirling's formula).
- Kolmogorov's interpretation (implicitly): the two probabilities differ by a factor that is polynomial in N, which can be regarded as small.
- On the log scale, which he used,

$$-\log \mathbb{P}^{\text{iid}}(E) = -\log \mathbb{P}^{\text{exch}}(E) + O(\log N).$$

And for finite sequences (2)

Intuitively, the IID and exchangeability assumptions are different:

- a long sequence with exactly one half of 1s may be exchangeable (if shuffled well),
- but it does not look IID.

Upper conformal probability

- The upper conformal probability of $E \subseteq \Omega$ is

$$\mathbb{P}^{\text{conf}}(E) := \inf\{\epsilon : \forall (z_1, \dots, z_N) \in E : S_N(z_1, \tau_1, z_2, \tau_2, \dots) \geq 1/\epsilon \ \tau\text{-a.s.}\},$$

  where S ranges over the conformal test martingales.
- Intuitively, for a small $\mathbb{P}^{\text{conf}}(E)$ the null hypothesis (IID or exchangeability) can be rejected by conformal test martingales.

Universality of conformal test martingales

Proposition

For any $E \subseteq \Omega$,

$$\mathbb{P}^{\text{exch}}(E) \le \mathbb{P}^{\text{conf}}(E) \le N\,\mathbb{P}^{\text{exch}}(E).$$

- In this crude sense, conformal test martingales are universal.
- They detect deviation from both exchangeability (factor of N) and IID (factor of $N^{3/2}$).
- Can we extend this to more interesting cases? (Open problem.)

References (1)


Vladimir Vovk, Alex Gammerman, and Glenn Shafer (2005).
Algorithmic learning in a random world.
New York: Springer.
Section 7.1: testing exchangeability.


Vladimir Vovk (2021).
Testing randomness online.
arXiv:1906.09256 [math.PR] (latest version: `http://alrw.net`). To appear in *Statistical Science*.
Testing the IID model and conformal change detection.

References (2)


Vladimir Vovk (2020).
Testing for concept shift online.
arXiv:2012.14246 [cs.LG] (latest version: http://alrw.net).


Vladimir Vovk, Ivan Petej, et al. (2021).
Retrain or not retrain: Conformal test martingales for change-point detection.
arXiv:2102.10439 [cs.LG] (latest version: http://alrw.net).

Thank you for your attention!