Testing Exchangeability On-Line

Vladimir Vovk, Ilia Nouretdinov, Alex Gammerman

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

On-line Compression Modelling Project

Working Paper #5

February 21, 2003

Project web site:
http://vovk.net/kp

Abstract

The majority of theoretical work in machine learning is done under the assumption of exchangeability: essentially, it is assumed that the examples are generated from the same probability distribution independently. This paper is concerned with the problem of testing the exchangeability assumption in the on-line mode: examples are observed one by one and the goal is to monitor on-line the strength of evidence against the hypothesis of exchangeability. We introduce the notion of exchangeability martingales, which are on-line procedures for detecting deviations from exchangeability; in essence, they are betting schemes that never risk bankruptcy and are fair under the hypothesis of exchangeability. Some specific exchangeability martingales are constructed using Transductive Confidence Machine. We report experimental results showing their performance on the USPS benchmark data set of hand-written digits (known to be somewhat heterogeneous); one of them multiplies the initial capital by more than 10^{18}; this means that the hypothesis of exchangeability is rejected at the significance level 10^{-18}.

Contents

1 Introduction

The majority of theoretical results in machine learning (such as PAC theory and statistical learning theory) depend on the exchangeability assumption, so it is surprising that so little work has been done on testing this assumption. Of course, testing is a traditional concern for statisticians, but the usual statistical techniques do not work for high-dimensional data sets such as the USPS data set (257 variables; see §4). This paper approaches the problem of testing exchangeability building on the theory of Transductive Confidence Machine (TCM), first introduced in [13] as a practically useful way of attaching confidence measures to predictions. It was shown in [11] that TCM is automatically well-calibrated under any exchangeable distribution when used in the on-line mode. In this paper we strengthen that result showing (Theorem 1 on p. 6) that a modification of TCM, which we call "randomised confidence transducer", produces p-values that are independent and distributed according to the uniform distribution U in $[0, 1]$. It turns out that Theorem 1 is a convenient tool for testing the hypothesis of exchangeability.

We start, in §2, with the definition of exchangeability and stating formally the problem of testing exchangeability on-line. TCM, in the form of "confidence transducers", is introduced in §3. In §4 we briefly describe the USPS data set and a particular confidence transducer, the Nearest Neighbour transducer, which works reasonably well for predicting the digits in the USPS data set. In §5 we define a family of exchangeability martingales, which we call "power martingales", and report experimental results for a simple mixture of NN power martingales (i.e., power martingales constructed from the Nearest Neighbour transducer). We found that the simple mixture, which is a non-negative exchangeability martingale that starts from 1, ends with more than 10^{10} on the USPS data set. The probability of this event under the null hypothesis of exchangeability is less than 10^{-10}, which contrasts sharply with typical significance levels, such as 1% or 5%, used in statistics. In §6 we describe procedures for "tracking the best power martingale"; one particular, very simple, procedure performs considerably better than the best power martingale on the USPS data set, achieving a final value exceeding 10^{18}.

2 Problem

In this section we set up our basic framework, making some important distinctions that have been glossed over so far: exchangeability vs. randomness, martingales vs. supermartingales, etc.

In our learning protocol, Nature outputs elements $z_1, z_2, \ldots$, called *examples*, of a measurable space $\mathbf{Z}$. (It is often the case that each example consists of two parts: an object and its label; we will not, however, need this additional structure in the theoretical considerations of this paper.) The hypothesis of *randomness* is that each example z_n, $n = 1, 2, \ldots$, is output according to a probability distribution P in $\mathbf{Z}$ and the examples are independent of each other (so the sequence $z_1 z_2 \ldots$ is output by the power distribution P^∞). The almost identical hypothesis of *exchangeability* is that the examples $z_1 z_2 \ldots$ are output according to an *exchangeable* probability distributions Q in $\mathbf{Z}^\infty$, i.e., such that under Q the permuted examples $z_{\pi(1)}, \ldots, z_{\pi(n)}$ are distributed as the original examples $z_1, \ldots, z_n$, for any n and any permutation π of $\{1, \ldots, n\}$. It is clear *a priori* that the exchangeability hypothesis is as weak as or weaker than the randomness hypothesis, since all power distributions are exchangeable.

We are interested in testing the hypothesis of randomness/exchangeability *on-line*: after observing each new example z_n Learner is required to output a number M_n reflecting the strength of evidence found against the hypothesis. The most natural way to do this is to use non-negative supermartingales starting from 1 (cf. [5]). Suppose first that we want to test the simple hypothesis that $z_1, z_2, \ldots$ are generated from a probability distribution Q in $\mathbf{Z}^\infty$. We say that a sequence of random variables $M_0, M_1, \ldots$ is a *Q-supermartingale* if, for all $n = 0, 1, \ldots$, M_n is a measurable function of $z_1, \ldots, z_n$ (in particular, M_0 is a constant) and

$$M_n \geq \mathbb{E}^Q \left(M_{n+1} \,|\, M_1, \ldots, M_n \right). \tag{1}$$

If $M_0 = 1$ and $\inf_n M_n \geq 0$, M_n can be regarded as the capital process of a player who starts from 1, never risks bankruptcy, at the beginning of each trial n places a fair (cf. (1)) bet on the z_n to be chosen by Nature, and maybe sometimes throws money away (since (1) is an inequality). If such a supermartingale M ever takes a large value, our belief in Q is undermined; this intuition is formalized by Doob's inequality, which implies

$$Q \{\exists n : M_n \geq C\} \leq 1/C, \tag{2}$$

where C is an arbitrary positive constant.

When testing a *composite hypothesis* $\mathcal{P}$ (i.e., a family of probability distributions), we will use $\mathcal{P}$*-supermartingales*, i.e., sequences of random variables $M_0, M_1, \ldots$ which are Q-supermartingales for all $Q \in \mathcal{P}$ simultaneously. If $\mathcal{P}$ is the set of all power distributions P^∞, P ranging over the probability distributions in $\mathbf{Z}$, $\mathcal{P}$-supermartingales will be called *randomness supermartingales*. We will be even more interested in the wider family $\mathcal{P}$ consisting of all exchangeable probability distributions Q in $\mathbf{Z}^\infty$; in this case we will use the term *exchangeability supermartingales* for $\mathcal{P}$-supermartingales.

De Finetti's theorem and the fact that Borel spaces are closed under countable products (see, e.g., [4], Theorem 1.49 and Lemma B.41) imply that each exchangeable distribution Q in $\mathbf{Z}^\infty$ is a mixture of power distributions P^∞ provided $\mathbf{Z}$ is Borel. By Property A of conditional probability distributions (see Appendix) the notions of non-negative randomness and exchangeability supermartingales coincide in the Borel case. But even without the assumption that $\mathbf{Z}$ is Borel, all randomness supermartingales are exchangeability supermartingales.

In this paper we will also need *randomised exchangeability martingales*; these are sequences of measurable functions $M_n(z_1, \theta_1, \ldots, z_n, \theta_n)$ (each example z_n is extended by adding a random number $\theta_n \in [0, 1]$) such that, for any exchangeable probability distribution Q in $\mathbf{Z}^\infty$,

$$M_n = \mathbb{E}^{Q \times U^\infty} \left(M_{n+1} \,|\, M_1, \ldots, M_n \right), \tag{3}$$

U being the uniform distribution in $[0, 1]$. We refrain from giving the analogous definition of randomised randomness martingales; the discussion in the previous paragraphs about the relation between randomness and exchangeability is also applicable in the randomised case (remember that $\mathbf{Z} \times [0, 1]$ is Borel when $\mathbf{Z}$ is). Doob's inequality (2) is also true for non-negative randomised exchangeability martingales starting from 1.

Remark Our definitions of martingale (3) and supermartingale (1) are from [1]; a more modern approach (cf. [6, 5]) would be to replace the condition "$\,|\, M_1, \ldots, M_n$" in (3) and (1) by "$\,|\, \mathcal{F}_n$", where $\mathcal{F}_n$ is the σ-algebra generated by $z_1, \ldots, z_n$ in the case of (1) and $z_1, \theta_1, \ldots, z_n, \theta_n$ in the case of (3) (i.e., $\mathcal{F}_n$ represents all information available by the end of trial n). To see how restrictive conditions (3) and (1) are, notice that the notions of randomised exchangeability martingale and exchangeability supermartingale become trivial when this apparently small change is made: the latter will be

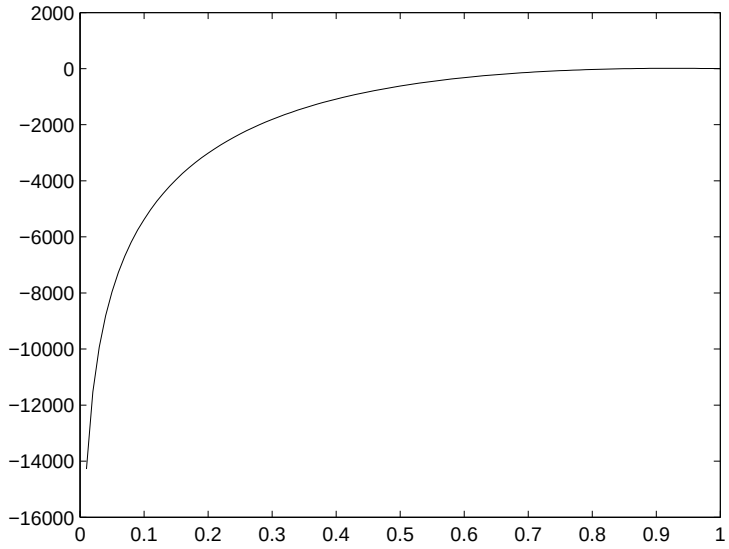

Figure 1: The final values, on the logarithmic (base 10) scale, attained by the randomised NN power martingales $M_n^{(\epsilon)}$ on the (full) USPS data set.

non-increasing processes ($M_0 \geq M_1 \geq \cdots$) and the former will only gamble on the random numbers $\theta_1, \theta_2, \ldots$.

Now we can state the goal of this paper. We will construct non-negative exchangeability supermartingales and randomised exchangeability martingales that, starting from 1, take large final values on data sets (concentrating on the USPS data set) deviating from exchangeability; as discussed earlier, this will also provide us with randomness supermartingales and randomised randomness martingales. Before this paper, it was not even clear that non-trivial supermartingales of this kind exist; we will see that they not only exist, but can attain huge final values starting from 1 and never risking bankruptcy.

3 Confidence Transducers

In this section we introduce the main tool for constructing exchangeability martingales. A family of measurable functions $\{A_n : n \in \mathbb{N}\}$, where $A_n : \mathbf{Z}^n \to \mathbb{R}^n$ for all n, $\mathbb{N}$ is the set of all positive integers and $\mathbb{R}$ is the set of all real numbers (equipped with the Borel σ-algebra), is called an *individual*

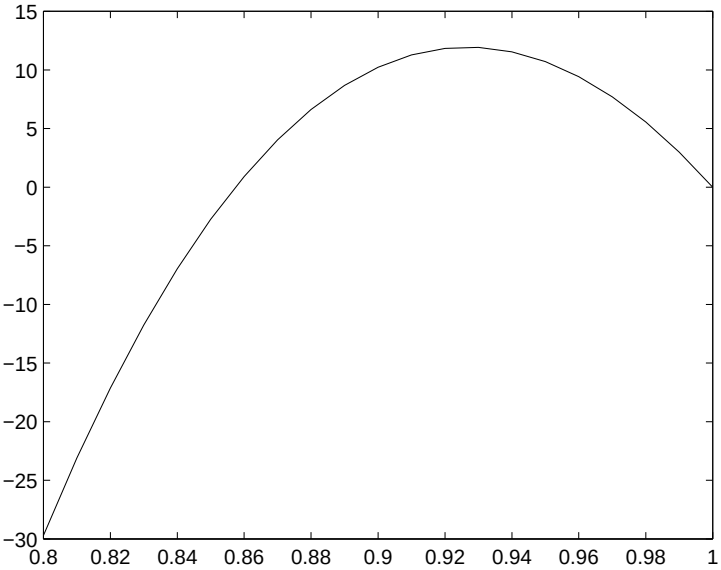

Figure 2: The final values for a narrower range of the parameter ϵ.

strangeness measure if, for any $n \in \mathbb{N}$, any permutation π of $\{1, \ldots, n\}$, any $(z_1, \ldots, z_n) \in \mathbf{Z}^n$, and any $(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n$,

$$\begin{aligned} (\alpha_1, \ldots, \alpha_n) = A_n(z_1, \ldots, z_n) \implies \\ (\alpha_{\pi(1)}, \ldots, \alpha_{\pi(n)}) = A_n(z_{\pi(1)}, \ldots, z_{\pi(n)}). \end{aligned} \tag{4}$$

In other words,

$$A_n : (z_1, \ldots, z_n) \mapsto (\alpha_1, \ldots, \alpha_n) \tag{5}$$

is an individual strangeness measure if every α_i is determined by the multiset $\{\!|z_1, \ldots, z_n|\!\}$ and z_i. (Sometime multisets are called "bags", whence our notation.) Individual strangeness measures will be our starting point when constructing exchangeability martingales.

A *randomised transducer* is a function f of the type $(\mathbf{Z} \times [0,1])^* \to [0,1]$. It is called "transducer" because it can be regarded as mapping each input sequence $(z_1, \theta_1, z_2, \theta_2, \ldots)$ in $(\mathbf{Z} \times [0,1])^\infty$ into the output sequence $(p_1, p_2, \ldots)$ of "p-values" defined by $p_n = f(z_1, \theta_1, \ldots, z_n, \theta_n)$, $n = 1, 2, \ldots$. We say that f is a *randomised E/U-transducer* if the output p-values $p_1 p_2 \ldots$ are always distributed according to the uniform distribution in $[0,1]^\infty$, provided

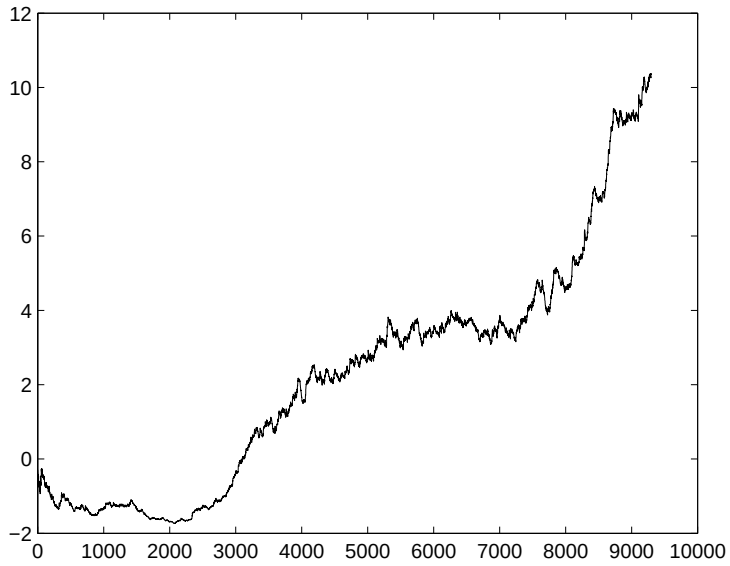

Figure 3: On-line performance of the randomised NN SM on the USPS data set. The growth is shown on the logarithmic (base 10) scale: $\log M_n$ is plotted against n. The final value attained is 2.18×10^{10}.

the input examples $z_1 z_2 \ldots$ are generated by an exchangeable probability distribution in $\mathbf{Z}^\infty$.

We will construct randomised exchangeability martingales from individual strangeness measures in two steps, first extracting randomised E/U transducers from the latter: given an individual strangeness measure A, for each sequence $(z_1, \theta_1, \ldots, z_n, \theta_n) \in (\mathbf{Z} \times [0,1])^*$ define

$$f(z_1, \theta_1, \ldots, z_n, \theta_n) := \frac{\#\{i : \alpha_i > \alpha_n\} + \theta_n \#\{i : \alpha_i = \alpha_n\}}{n}, \tag{6}$$

where α_i, $i = 1, 2, \ldots$, are computed from z_i using A as per (5). Each randomised transducer f that can be obtained in this way will be called a *randomised confidence transducer*.

Theorem 1 *Each randomised confidence transducer is a randomised E/U transducer.*

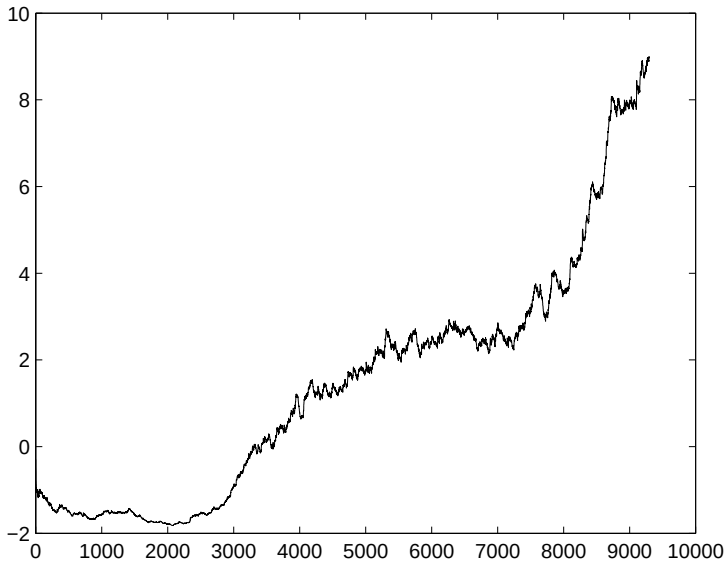

Figure 4: On-line performance of the deterministic NN SM on the USPS data set The final value is 9.13×10^8.

A special case (labelled there as Theorem 2) of this theorem was used in [11] as a tool for region prediction.

In a similar way we can define (deterministic) *confidence transducers* f: given an individual strangeness measure A, for each sequence $(z_1, \dots, z_n) \in \mathbf{Z}^*$ set

$$f(z_1, \dots, z_n) := \frac{\#\{i : \alpha_i \geq \alpha_n\}}{n},$$

where α_i are computed as before. In general, a (deterministic) *transducer* is a function f of the type $\mathbf{Z}^* \to [0, 1]$; as before, we associate with f a mapping from $z_1 z_2 \dots$ to the p-values $p_1 p_2 \dots$ ($p_n = f(z_1, \dots, z_n)$). We say that f is an *E/U-supertransducer* if $p_1 \leq \overline{p}_1, p_2 \leq \overline{p}_2, \dots$ for some random variables $\overline{p}_1, \overline{p}_2, \dots$ distributed independently according to U, whatever the exchangeable distribution generating $z_1, z_2, \dots$ is. The following implication of Theorem 1 is obvious:

Corollary 1 *Each deterministic confidence transducer is an E/U supertransducer.*

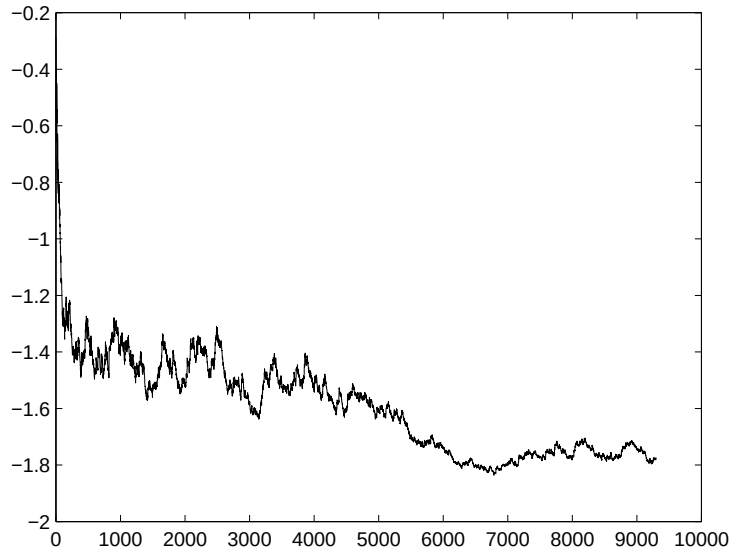

Figure 5: On-line performance of the randomised NN SM on a randomly permuted USPS data set. The final value is 0.0117.

4 USPS Data Set and Nearest Neighbour Transducers

The USPS data set (described in, e.g., [7]) consists of 7291 training examples and 2007 test examples; we merged the training set and the test set, in this order, obtaining what we call the *full USPS data set* (or just USPS data set). Each example consists of an image (16×16 matrix of pixels) and its label (0 to 9). In region prediction (e.g., [11]) it is usually beneficial to pre-process the images; no pre-processing is done in this paper.

It is well-known that the USPS data set is heterogeneous; in particular, the training and test sets seem to have different distributions. (See, e.g., [2].) In the next two sections we will see the huge scale of this heterogeneity.

It was shown in [12] that a Nearest Neighbours TCM provides a universally optimal, in an asymptotic sense, on-line algorithm for predicting classifications under the assumption of exchangeability. On the empirical side, Figures 1 and 2 in [11] show that a Nearest Neighbour TCM performs reasonably well on the USPS data set. Therefore, it is natural to expect that the Nearest Neighbour(s) idea will also perform well in the problem of

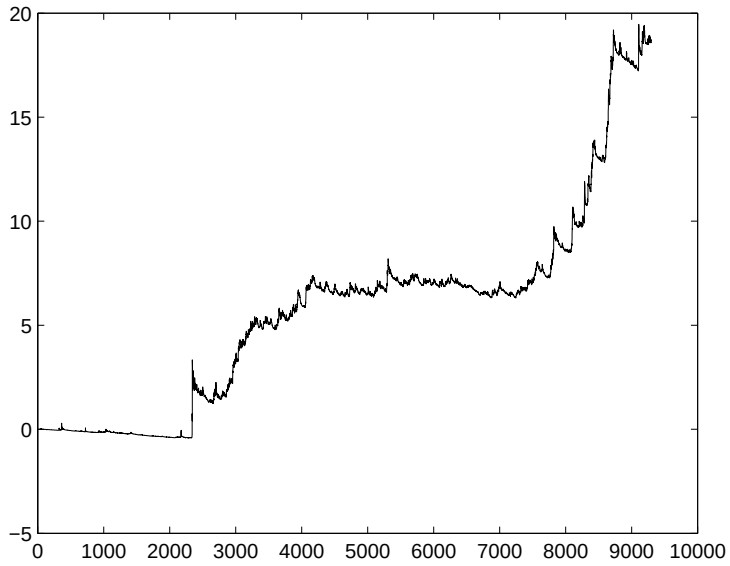

Figure 6: On-line performance of the randomised NN SJ with parameters $(R, J) = (1\%, 1‰)$ on the USPS data set. The final value is 4.71×10^{18}.

testing exchangeability.

We define the *Nearest Neighbour* (NN) individual strangeness measure as mapping (5) where

$$\alpha_i := \frac{\min_{j \neq i : y_j = y_i} d(x_i, x_j)}{\min_{j \neq i : y_j \neq y_i} d(x_i, x_j)}; \tag{7}$$

in this formula, $x_i \in \mathbb{R}^{256}$ is the image in a USPS example z_i, $y_i \in \{0, 1, \ldots, 9\}$ is the corresponding label (so that $z_i = (x_i, y_i)$), and d is the Euclidean distance. Intuitively, an image is considered strange if it is in the middle of images labelled in a different way and is far from the images labelled in the same way. The corresponding confidence transducer (randomised or deterministic) will be called the *NN transducer*.

5 Power Martingales

In this and next sections we discuss the second step in transforming individual strangeness measures into (randomised) exchangeability (super)martingales:

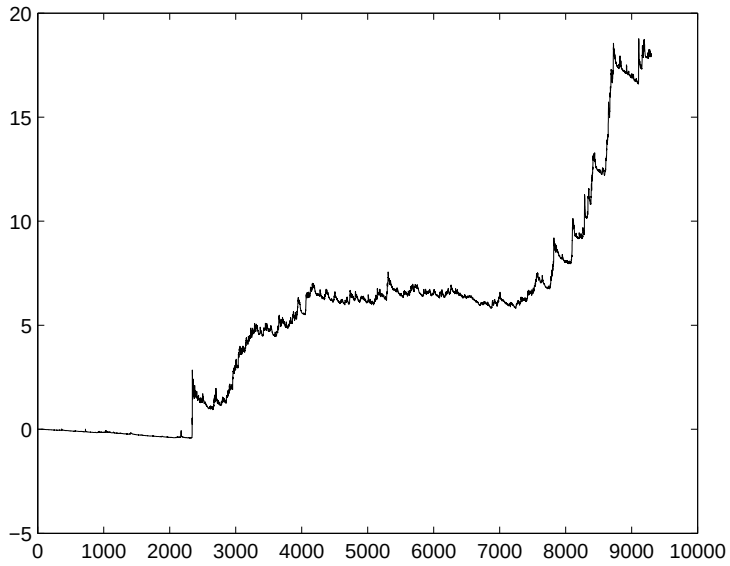

Figure 7: On-line performance of the deterministic NN SJ with parameters $(1\%, 1\permil)$ on the USPS data set. The final value is 1.01×10^{18}.

constructing the latter from (randomised) E/U (super)transducers. To this end we use the procedure suggested in [9].

Since $\int_0^1 \epsilon p^{\epsilon-1} dp = 1$, the random variables

$$M_n^{(\epsilon)} := \prod_{i=1}^{n} \left(\epsilon p_i^{\epsilon-1} \right), \tag{8}$$

where p_n are the p-values output by a randomised confidence transducer, will be a non-negative randomised exchangeability martingale with initial value 1; this family of martingales, indexed by $\epsilon \in [0, 1]$, will be called the *randomised power martingales*.

When applied to the NN transducer, the family of randomised power martingales (*randomised NN power martingales*) might at first not look very promising (Figure 1), but if we concentrate on a narrower range of ϵ (Figure 2), it becomes clear that the final values for some ϵ are very large.

To eliminate the dependence on ϵ, we may use the randomised exchangeability martingale

$$M_n := \int_0^1 M_n^{(\epsilon)} d\epsilon, \tag{9}$$

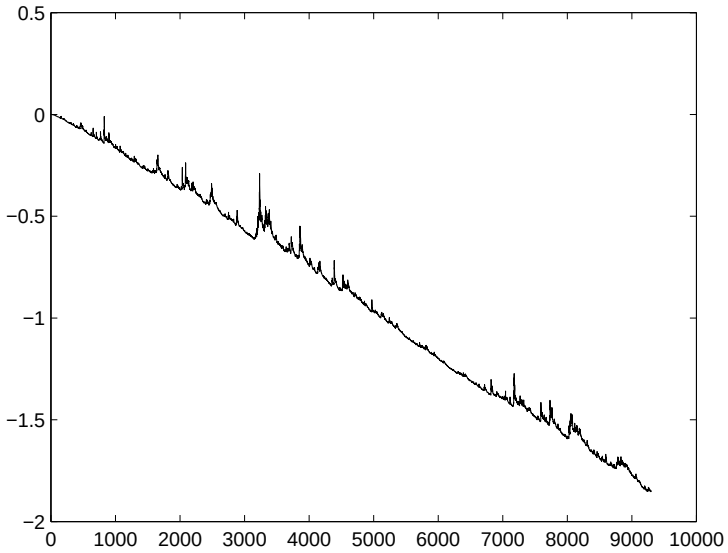

Figure 8: On-line performance of the randomised NN SJ with parameters $(1\%, 1‰)$ on the randomly permuted USPS data set. The final value is 0.0142.

which is called the *simple mixture* of $M_n^{(\epsilon)}$. The simple mixture of randomised NN power martingales (which will also be referred to as the *randomised NN SM*) usually ends up with more than 10^{10}; a typical trajectory is shown in Figure 3. This figure and Figures 5, 6, 8 below are affected by statistical variation (since the outcome depends on the random numbers θ_i actually generated), but the dependence is not too heavy. For example, in the case of the randomised NN SM the final values are: 2.18×10^{10} (MATLAB pseudo-random numbers generator started from state 0), 1.32×10^{10} (state 1), 1.60×10^{10} (state 2),...; in what follows we only give results for initial state 0.

As clear from Figure 3, the difference between the training and test sets is not the only anomaly in the USPS data set: the rapid growth of the randomised NN SM starts already on the training set.

If p_n are output by the deterministic NN transducer, we call (8) the *NN power supermartingales* and we refer to (9) as the *deterministic NN SM*. As Figure 4 shows, the growth rate of the latter is slightly less than that of its randomised counterpart.

The result for a randomly permuted USPS data set is shown in Figure 5.

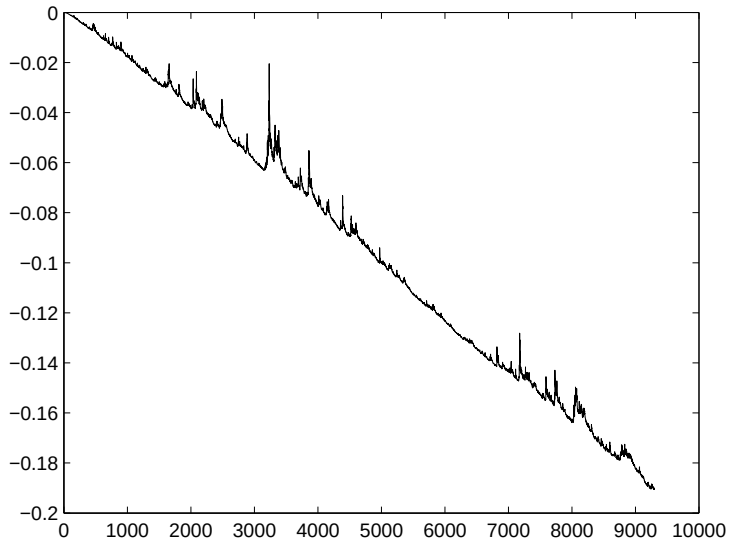

Figure 9: On-line performance of the deterministic NN SJ with parameters $(1\%, 1\%_{00})$ on the randomly permuted USPS data set. The final value is 0.646.

A low final value (about %) results from NN SM's futile attempts to gamble against a random sequence; to make possible spectacular gains against highly untypical sequences such as the original USPS data set, it has to underperform against random sequences.

6 Tracking the Best Power Martingale

The simple mixture of the previous section has modest goals; the best it can do is to approximate the performance of the best power martingale. In this section we will see that it is possible to "track" the best power martingale, so that the resulting performance considerably exceeds that of the best "static" martingale (8).

We first generalise (8) as follows: for each $\epsilon = \epsilon_1 \epsilon_2 \ldots \in [0,1]^\infty$, we set

$$M_n^{(\epsilon)} := \prod_{i=1}^{n} \left(\epsilon_i p_i^{\epsilon_i - 1} \right). \tag{10}$$

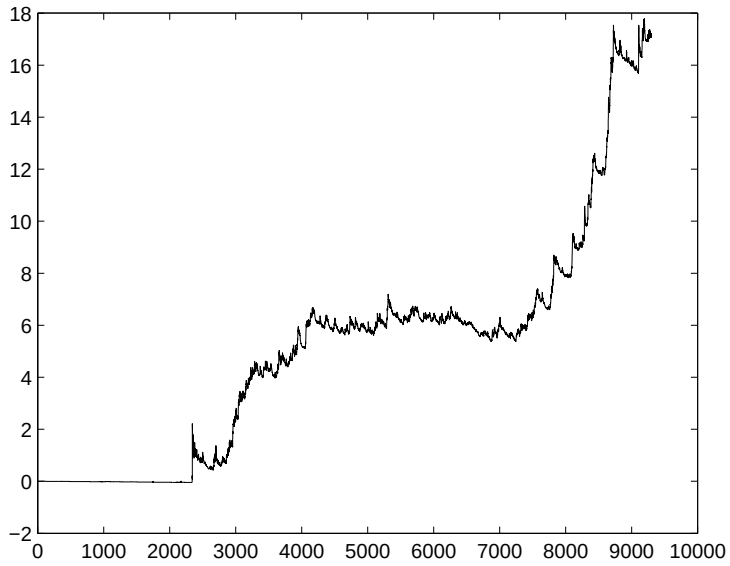

Figure 10: On-line performance of the deterministic NN SJ with parameters $(1\%, 1\%_{00})$ on the USPS data set. The final value is 1.48×10^{17}.

For any probability distribution μ in $[0,1]^\infty$, define

$$M_n := \int_{[0,1]^\infty} M_n^{(\epsilon)} \mu(d\epsilon). \tag{11}$$

It is convenient to specify μ in terms of the distribution of the coordinate random variables ϵ_n (but of course, since we integrate over μ, this does not involve any extra randomisation; in particular, the mixture (11) is deterministic if p_n are generated by a deterministic confidence transducer). One possible μ is generated by the following *Sleepy Jumper* automaton. The states of Sleepy Jumper are elements of the Cartesian product $\{\text{awake}, \text{asleep}\} \times [0,1]$. Sleepy Jumper starts from the state $(\text{asleep}, 1)$; when he is in a state (s, ϵ), his transition function prescribes that:

- if $s = \text{asleep}$, he moves to the state (awake, ϵ) ("wakes up") with probability R ($R \in [0,1]$ is one of two parameters of the automaton) and stays in the state $(\text{asleep}, \epsilon)$ with probability $1 - R$;
- if $s = \text{awake}$, he moves to the state $(\overline{s}, \overline{\epsilon})$, where $\overline{\epsilon}$ and $\overline{s}$ are generated independently as follows: $\overline{\epsilon} = \epsilon$ with probability $1 - J$ ($J \in [0,1]$,

the "probability of jumping", is the other parameter) and $\overline{\epsilon}$ is chosen randomly from U with probability J; $\overline{s}$ = awake with probability $1-R$ and $\overline{s}$ = asleep with probability R.

The output of the Sleepy Jumper automaton starting from $(s_1, \tilde{\epsilon}_1) = (\text{passive}, 1)$ and further moving through the states $(s_2, \tilde{\epsilon}_2), (s_3, \tilde{\epsilon}_3), \ldots$ is the sequence $\epsilon_1, \epsilon_2, \ldots$, where

$$\epsilon_n := \begin{cases} \tilde{\epsilon}_n & \text{if } s_n = \text{awake} \\ 1 & \text{otherwise.} \end{cases}$$

The probability distribution μ of $\epsilon_1, \epsilon_2, \ldots$ generated in this way defines, by (11), a randomised exchangeability martingale (or exchangeability supermartingale), which we call the *randomised Sleepy Jumper martingale* (resp. *Sleepy Jumper supermartingale*); if p_n are produced by the NN transducer (randomised or deterministic, as appropriate), we refer to these processes as the randomised/deterministic *NN SJ*.

Figures 6 and 7 show the performance of the randomised and deterministic NN SJ for parameters $R = 0.01$ and $J = 0.001$. When applied to the randomly permuted USPS data set, the randomised NN SJ's performance is as shown in Figure 8. One way to improve the performance against a random data set is to decrease the jumping rate: if $J = 0.0001$, we obtain a much better performance (Figure 9), even for a deterministic NN SJ. It is easy to see the cause of the improvement: when $J = 0.0001$, the μ-measure of supermartingales (10) that make no jumps on the USPS data set will be at least $0.9999^{9298} > e^{-1}$. The performance on the original data set deteriorates (Figure 10) but not drastically.

Remark The approach of this section is reminiscent of "tracking the best expert" in the theory of prediction with expert advice. A general "Aggregating Algorithm" (AA) for merging experts was introduced in [8]; in the context of this section, the experts are the power martingales and the mixing operation (9) plays the role of (and is a special case of) the AA. Herbster and Warmuth [3] showed how to extend the AA to "track the best expert", to try and outperform even the best static expert. Vovk [10] noticed that Herbster and Warmuth's algorithm is in fact a special case of the AA, when it is applied not to the original experts (in our case, (8)) but to "superexperts" (in our case, (10)).

Of course, there are other ideas that can used when combining (10); e.g., it would be natural to allow ϵ not only occasionally to jump randomly but also to allow it to drift slowly.

References

[1] J. L. Doob. *Stochastic Processes*. Wiley, New York, 1953.

[2] Yoav Freund and Robert E. Schapire. Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the Thirteenth International Conference*, pages 148–156, 1996.

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

[4] Mark J. Schervish. *Theory of Statistics*. Springer, New York, 1995.

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

[6] Albert N. Shiryaev. *Probability*. Springer, New York, second edition, 1996.

[7] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

[8] Vladimir Vovk. Aggregating strategies. In M. Fulk and J. Case, editors, *Proceedings of the Third Annual Workshop on Computational Learning Theory*, pages 371–383, San Mateo, CA, 1990. Morgan Kaufmann.

[9] Vladimir Vovk. A logic of probability, with application to the foundations of statistics (with discussion). *Journal of the Royal Statistical Society* B, 55:317–351, 1993.

[10] Vladimir Vovk. Derandomizing stochastic prediction strategies. *Machine Learning*, 35(3):247–282, 1999.

[11] Vladimir Vovk. On-line Confidence Machines are well-calibrated, Online Compression Modelling project, `http://vovk.net/kp`, Working Paper #1. In *Proceedings of the Forty Third Annual Symposium on Foundations of Computer Science*, pages 187–196. IEEE Computer Society, 2002.

[12] Vladimir Vovk. Universal well-calibrated algorithm for on-line classification, On-line Compression Modelling project, `http://vovk.net/kp`, Working Paper #3, November 2002.

[13] Vladimir Vovk, Alex Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 444–453, San Francisco, CA, 1999. Morgan Kaufmann.

Appendix

We will use the notation $\mathbb{E}_{\mathcal{F}}$ for the conditional expectation w.r. to a σ-algebra $\mathcal{F}$; if necessary, the underlying probability distribution will be given as an upper index. Similarly, $\mathbb{P}_{\mathcal{F}}$ will stand for the conditional probability w.r. to $\mathcal{F}$. In this paper we use the following properties of conditional expectation:

A. If $\mathcal{F}$ is a σ-algebra, ξ is a non-negative random variable, and a probability distribution P is a mixture of probability distributions P_θ, $P = \int P_\theta \mu(d\theta)$, then $\mathbb{E}^P_{\mathcal{F}}(\xi) = \int \mathbb{E}^{P_\theta}_{\mathcal{F}}(\xi)\mu(d\theta)$ a.s.

B. If $\mathcal{G}$ and $\mathcal{F}$ are σ-algebras, $\mathcal{G} \subseteq \mathcal{F}$, ξ and η are bounded $\mathcal{F}$-measurable random variables, and η is $\mathcal{G}$-measurable, $\mathbb{E}_{\mathcal{G}}(\xi\eta) = \eta\,\mathbb{E}_{\mathcal{G}}(\xi)$ a.s.

C. If $\mathcal{G}$ and $\mathcal{F}$ are σ-algebras, $\mathcal{G} \subseteq \mathcal{F}$, and ξ is a random variable, $\mathbb{E}_{\mathcal{G}}(\mathbb{E}_{\mathcal{F}}(\xi)) = \mathbb{E}_{\mathcal{G}}(\xi)$ a.s.; in particular, $\mathbb{E}(\mathbb{E}_{\mathcal{F}}(\xi)) = \mathbb{E}(\xi)$.

The first property is obvious, and the other two are well-known (see, e.g., [6], §II.7.4).

Proof of Theorem 1

This proof is a generalization of the proof of Theorem 1 in [11], with the same basic idea: to show that $(p_1, \dots, p_N)$ is distributed as U^N (it is easy to get

rid of the assumption of a fixed horizon N), we use the old idea of reversing the time. Imagine that the sample $(z_1, \dots, z_N)$ is generated in two steps: first, the multiset $\{z_1, \dots, z_N\}$ is generated from some probability distribution (namely, the image of the exchangeable distribution Q generating $z_1, z_2, \dots$ under the mapping $(z_1, z_2, \dots) \mapsto \{z_1, \dots, z_N\}$), and then the actual sample $(z_1, \dots, z_N)$ is chosen randomly from the set of all orderings of $\{z_1, \dots, z_N\}$. Already the second step ensures that, conditionally on knowing $\{z_1, \dots, z_N\}$ (and, therefore, unconditionally), the sequence $(p_N, \dots, p_1)$ is distributed as U^N. Indeed, roughly speaking (i.e., ignoring borderline effects), p_N will be N^{-1} times the rank of α_N in the set $\{\alpha_i : i = 1, \dots, N\}$, and so distributed, at least approximately, as U, since all permutations are equiprobable; when z_N is disclosed, the value p_N will be settled; conditionally on knowing $\{z_1, \dots, z_N\}$ and z_N (and, therefore, knowing $\{z_1, \dots, z_{N-1}\}$), p_{N-1} will also be distributed as U, and so on.

We start the formal proof by defining the σ-algebra $\mathcal{G}_n$, $n = 0, 1, 2, \dots$, as the collection of all measurable sets $E \subseteq (\mathbf{Z} \times [0,1])^\infty$ which satisfy

$$(z_1, \theta_1, z_2, \theta_2, \dots) \in E \implies$$
$$(z_{\pi(1)}, \tilde{\theta}_1, \dots, z_{\pi(n)}, \tilde{\theta}_n, z_{n+1}, \theta_{n+1}, z_{n+2}, \theta_{n+2}, \dots) \in E$$

for any permutation π of $\{1, \dots, n\}$ and any sequences $z_1 z_2 \dots \in \mathbf{Z}^\infty$, $\theta_1 \theta_2 \dots \in [0,1]^\infty$, $\tilde{\theta}_1 \dots \tilde{\theta}_n \in [0,1]^n$. In particular, $\mathcal{G}_0$ (the most informative σ-algebra) coincides with the original σ-algebra on $(\mathbf{Z} \times [0,1])^\infty$; $\mathcal{G}_0 \supseteq \mathcal{G}_1 \supseteq \cdots$.

Fix a randomised confidence transducer f; it will usually be left implicit in our notation. Let p_n be the random variable $f(z_1, \theta_1, \dots, z_n, \theta_n)$ for each $n = 1, 2, \dots$; $\mathbb{P}$ will refer to the probability distribution $Q \times U^\infty$ (over examples z_n and random numbers θ_n) and $\mathbb{E}$ to the expectation w.r. to $\mathbb{P}$. The proof will be based on the following lemma.

Lemma 1 *For any trial n and any $\delta \in [0,1]$,*

$$\mathbb{P}_{\mathcal{G}_n} \{p_n \le \delta\} = \delta. \tag{12}$$

Proof This coincides with Lemma 1 in [11], since $\mathrm{err}_n = \mathbb{I}_{\{p_n \le \delta\}}$ (assuming the random numbers τ_n used by rTCM in [11] are $1 - \theta_n$), where $\mathbb{I}_E$ means the indicator of a set E. ∎

The other basic result that we will need is the following lemma (whose simple proof is omitted).

Lemma 2 *For any trial $n = 1, 2, \ldots$, p_n is $\mathcal{G}_{n-1}$-measurable.*

Fix temporarily positive integer N. First we prove that, for any $n = 1, \ldots, N$ and any $\delta_1, \ldots, \delta_n \in [0, 1]$,

$$\mathbb{P}_{\mathcal{G}_n}\{p_n \le \delta_n, \ldots, p_1 \le \delta_1) = \delta_n \cdots \delta_1. \tag{13}$$

The proof is by induction in n. For $n = 1$, (13) immediately follows from Lemma 1. For $n > 1$ we obtain, making use of Lemmas 1 and 2, properties B and C of conditional expectations, and the inductive assumption:

$$\begin{aligned}
&\mathbb{P}_{\mathcal{G}_n}\{p_n \le \delta_n, \ldots, p_1 \le \delta_1) \\
&= \mathbb{E}_{\mathcal{G}_n}\left(\mathbb{E}_{\mathcal{G}_{n-1}}\left(\mathbb{I}_{\{p_n \le \delta_n\}}\mathbb{I}_{\{p_{n-1} \le \delta_{n-1}, \ldots, p_1 \le \delta_1\}}\right)\right) \\
&= \mathbb{E}_{\mathcal{G}_n}\left(\mathbb{I}_{\{p_n \le \delta_n\}}\,\mathbb{E}_{\mathcal{G}_{n-1}}\left(\mathbb{I}_{\{p_{n-1} \le \delta_{n-1}, \ldots, p_1 \le \delta_1\}}\right)\right) \\
&= \mathbb{E}_{\mathcal{G}_n}\left(\mathbb{I}_{\{p_n \le \delta_n\}}\delta_{n-1} \cdots \delta_1\right) = \delta_n\delta_{n-1} \cdots \delta_1
\end{aligned}$$

almost surely.

By property C, (13) immediately implies

$$\mathbb{P}\{p_N \le \delta_N, \ldots, p_1 \le \delta_1\} = \delta_N \cdots \delta_1.$$

Therefore, we have proved that the distribution of the random sequence $p_1p_2\ldots \in [0, 1]^\infty$ coincides with U^∞ on the σ-algebra $\mathcal{F}_N$ generated by the first N coordinate random variables $p_1, \ldots, p_N$. It is well known (see, e.g., [6], Theorem II.3.3) that this implies that the distribution of $p_1p_2\ldots$ coincides with U^∞ on all measurable sets in $[0, 1]^\infty$.

On-line Compression Modelling Project
Working Papers

1. *On-line confidence machines are well-calibrated*, by Vladimir Vovk, April 2002.

2. *Asymptotic optimality of Transductive Confidence Machine*, by Vladimir Vovk, May 2002.

3. *Universal well-calibrated algorithm for on-line classification*, by Vladimir Vovk, November 2002.

4. *Mondrian Confidence Machine*, by Vladimir Vovk, David Lindsay, Ilia Nouretdinov and Alex Gammerman, March 2003.

5. *Testing exchangeability on-line*, by Vladimir Vovk, Ilia Nouretdinov and Alex Gammerman, February 2003.

6. *Criterion of calibration for Transductive Confidence Machine with limited feedback*, by Ilia Nouretdinov and Vladimir Vovk, April 2003.

7. *Online region prediction with real teachers*, by Daniil Ryabko, Vladimir Vovk and Alex Gammerman, March 2003.

8. *Well-calibrated predictions from on-line compression models*, by Vladimir Vovk, April 2003.