Journal of Machine Learning Research 23 (2022) 1-27 Submitted 9/21; Published 4/22

A universally consistent learning rule with a universally monotone error

Vladimir Pestov VLADIMIR.PESTOV@UOTTAWA.CA
Departamento de Matemática
Universidade Federal de Santa Catarina
Campus Universitário Trindade
CEP 88.040-900 Florianópolis-SC, Brasil
and
Departement of Mathematics and Statistics
University of Ottawa
STEM Complex, 150 Louis-Pasteur Pvt
Ottawa, Ontario K1N 6N5 Canada

Editor: Mehryar Mohri

Abstract

We present a universally consistent learning rule whose expected error is monotone non-increasing with the sample size under every data distribution. The question of existence of such rules was brought up in 1996 by Devroye, Györfi and Lugosi (who called them "smart"). Our rule is fully deterministic, a data-dependent partitioning rule constructed in an arbitrary domain (a standard Borel space) using a cyclic order. The central idea is to only partition at each step those cyclic intervals that exhibit a sufficient empirical diversity of labels, thus avoiding a region where the error function is convex.

Keywords: Learning rule, learning error, "smart" rule, partitioning rule, cyclic order

1. Introduction

Here we are interested in the learning rules for a binary classification problem. Given a labelled n-sample σ_n, such a rule outputs a binary classifier for the domain Ω, that is, predicts a label, 0 or 1, for every point x of the domain. We denote the rule by $g = (g_n)_{n=1}^{\infty}$, and the predicted label for x based on a sample σ_n, by $g_n(\sigma_n)(x)$. Now let $\tilde{\mu}$ be an unknown distribution of the labelled datapoints (X, Y), that is, a probability measure on $\Omega \times \{0, 1\}$. The learning (or generalization, or misclassification) error $L_{\tilde{\mu}}(g_n)$ is the random variable

$$P_{\tilde{\mu}}[g_n(D_n)(X) \neq Y \mid D_n],$$

where D_n is a random labelled n-sample. The rule g is *consistent* (under $\tilde{\mu}$), if the error converges to the smallest possible classification error (the Bayes error), $L^*(\tilde{\mu})$, in expectation (or probability):

$$\mathbb{E}_{\tilde{\mu}}[L_{\tilde{\mu}}(g_n)] \overset{n\to\infty}{\longrightarrow} L^*(\tilde{\mu}).$$

The rule is *universally consistent* if it is consistent under every data distribution $\tilde{\mu}$. Intuitively, this means that the more data we have, the better is the prediction of the learning

rule, and asymptotically as $n \to \infty$, it is as good as it can possibly get under the (unknown) data law.

It is therefore tempting to think that the learning error does not increase under the transition $n \mapsto n+1$, that is, the sequence

$$\mathbb{E}_{\tilde{\mu}}[L_{\tilde{\mu}}(g_n)], \quad n = 1, 2, 3, \ldots \tag{1}$$

is monotone nonincreasing. Perhaps surprisingly, it is not the case. See Devroye et al. (1996), Sect. 6.8 for a simple example of a data distribution on the interval, under which the nearest neighbour rule has a strictly smaller learning error for $n = 1$ than it has for $n = 2$. It is not difficult to construct similar counter-examples for other common universally consistent learning rules (cf. Problems 6.14 and 6.15, *loco citato*).

Devroye, Györfi and Lugosi called a rule g *smart* (Devroye et al. (1996), Sect. 6.8) if for all labelled data distributions μ on $\Omega \times \{0, 1\}$, the sequence in Eq. (1) is nonincreasing. Based on the above, they have conjectured that no universally consistent learning rule is "smart". (Cf. *loc. cit.*, bottom of p. 106 and Problem 6.16, p. 109.)

Our aim is to show that "smart" universally consistent rules do exist, even without requiring any amount of randomization.

We use a partitioning rule: the domain is divided in disjoint cells, and the label for each cell is determined by the majority vote among all datapoints contained in it. It is easy to show that for a fixed partition, the error does not increase with the sample size (Problem 6.13 in Devroye et al. (1996)). However, for a partitioning rule to be consistent, the cells have to be divided, and this is where the error jump may occur.

Here is the root of the problem. Let Y, Y_i, $i = 1, \ldots, n$ be i.i.d. random labels following a Bernoulli distribution with $p = P[Y = 1]$. Consider the predictor for the value of Y based on the majority vote among $Y_1, \ldots, Y_n$. For the odd values of n, the voting ties are avoided, and the misclassification error is a polynomial function in p:

$$L(p, n) = P[Y = 1]P\left[\frac{1}{n}\sum_{i=1}^{n} Y_i < \frac{1}{2}\right] + P[Y = 0]P\left[\frac{1}{n}\sum_{i=1}^{n} Y_i > \frac{1}{2}\right].$$

For $n > 1$, the error function is not concave: there is a straight line segment joining two points on the graph that is strictly above the underlying part of the graph (Fig. 1).

This is because, for $n > 1$, the derivative of the polynomial function $L(p, n)$ at $p = 0$ and 1 equals 1 and -1 respectively (see Problem 5.6(2) in Devroye et al. (1996), p. 84 for a Taylor polynomial). The optimal (Bayes) predictor for the problem gives the value 0 if $p < 1/2$ and 1 if $p > 1/2$, and the Bayes error is given by $L^*(p) = \min\{p, 1-p\}$. Since $L(p, n) > \min\{p, 1-p\}$ at all points except $0, 1/2, 1$, it follows that there are small neighbourhoods of 0 and 1 in which the polynomial function $L(p, n)$ is strictly convex.

This implies that no concave function strictly greater than $\min\{p, 1-p\}$ is contained under the graph of $L(p, n)$. In particular, given $N > n$, for some $p_0 > 0$ small enough

$$L(p_0, n) < \frac{1}{2}L(2p_0, N). \tag{2}$$

And here is how the error value can increase after we refine the partition, even if we increase the sample size. Suppose the domain Ω is subdivided into two cells of equal measure,

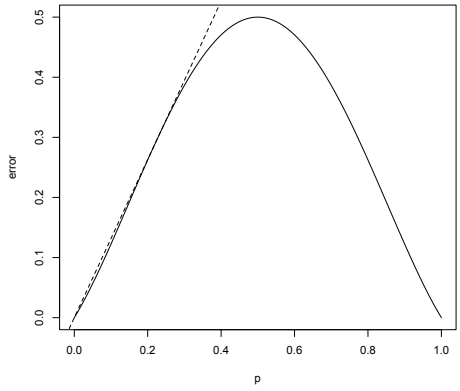

Figure 1: Non-concavity of the error function $L(p, 3)$.

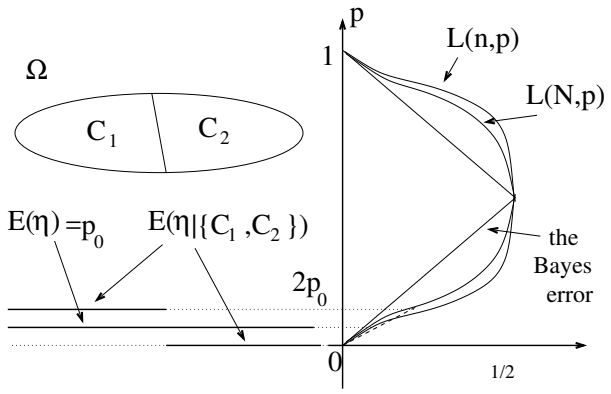

Figure 2: Splitting the domain into two cells in the area of convexity of error.

C_1 and C_2, with conditional probabilities of getting label 1 equal to $2p_0$ and 0 respectively. (Fig. 2.)

The learning error of the rule based on the trivial partition, $\{\Omega\}$, and a random n-sample σ equals $L(p_0, n)$. But when we proceed to the rule based on the finer partition $\{C_1, C_2\}$, the error, conditionally on each cell containing N sample points, strictly increases:

$$\begin{aligned} P[X \in C_1] \cdot L(2p_0, N) + P[X \in C_2] \cdot L(0, N) &= \frac{1}{2} L(2p_0, N) + \frac{1}{2} 0 \\ &> L(p_0, n). \end{aligned}$$

Using the monotonicity of $L(p, n)$ in n, it is easy to deduce that the expected error of the histogram rule based on the trivial partition, $p_{\{\Omega\}}$, over i.i.d. n-samples is less than the expected error of the rule based on the finer partition $\{C_1.C_2\}$, over the i.i.d. N-samples.

However, away from the endpoints of the interval $[0, 1]$ this phenomenon no longer occurs. Given $\epsilon > 0$, if N is sufficiently large, then on the interval $[\epsilon, 1 - \epsilon]$ the concave envelope of the function $L(p, N)$ (that is, the smallest concave function majorising it) is

smaller than the function $L(p, n)$. This means that a cell C can be safely partitioned (into any finite number of smaller cells) once the conditional probability $p = P[Y = 1 \mid X \in C]$ is bounded away from 0 and 1, that is, belongs to some interval $[\epsilon_n, 1 - \epsilon_n]$, where $\epsilon_n \downarrow 0$. Therefore, the solution is to only partition a cell C when it is empirically confirmed that $P[Y = 1 \mid X \in C]$ is in the interval $[\epsilon_n, 1 - \epsilon_n]$:

$$P_\sigma[Y = 1 \mid X \in C] = \frac{\sum_{i\colon X_i \in C} Y_i}{\sharp\{i\colon X_i \in C\}} \in [\epsilon_n, 1 - \epsilon_n].$$

Here P_σ stands for the empirical (conditional) probability based on a random labelled sample σ.

There is still a probability of empirical error, but, near 0 and 1 and for ϵ_n fixed, this error is a polynomial function in p (resp. $1-p$) of higher order $\epsilon_n N$, where N is the number of points of the testing sample contained in the cell. Thus, even if we may from time to time erroneously partition a cell when we should not, the expected compound error under the transition $n \mapsto N$ still can be kept below the curve of the error function $L(p, n)$, provided N is large enough. (Lemmas 8 and 11.)

Now, a description of the learning rule, $g = (g_n)$. The empirical path (x_n) is divided into points of three kinds. A subsequence (n_k) is chosen, starting with $n_1 = 1$. The points (x_{n_k+1}) (with labels stripped off) are used to form a partition of the domain into half-open subintervals, for which purpose we fix a circular order on the domain. Equivalently, we identify the domain with the unit circle $\mathbb{S}^1$. This way, almost surely the measure of every cell of the partition is strictly positive. (Fig. 3.)

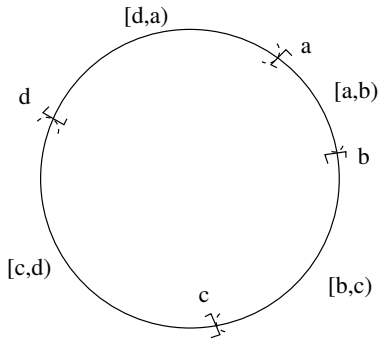

Figure 3: Half-open cyclic intervals.

The hypothesis is only updated at the steps of the form $n = n_k$, so for all the intermediate values $n_k < n < n_{k+1}$, the rule g_n just repeats the hypothesis output by the rule g_{n_k}:

$$g_n(\sigma) = g_{n_k}(\sigma[1, \ldots, n_k]).$$

The hypotheses g_{n_k} are generated recursively, that is, in order to output the hypothesis g_{n_k}, we need to know the hypotheses g_{n_i}, $i = 1, 2, \ldots, k-1$. In particular, the partitioning set $\mathcal{Q}_k \subseteq \{x_{n_1}, \ldots, x_{n_k}\}$ is selected recursively as well.

The interval of integers $[n_k + 2, n_{k+1}]$ is divided into two contiguous blocks, A_k and B_k, of length a_k and b_k respectively. Thus, $n_{k+1} = n_k + 1 + a_k + b_k$. We call A_k the testing block, and B_k, the labelling block. (See Fig. 4.)

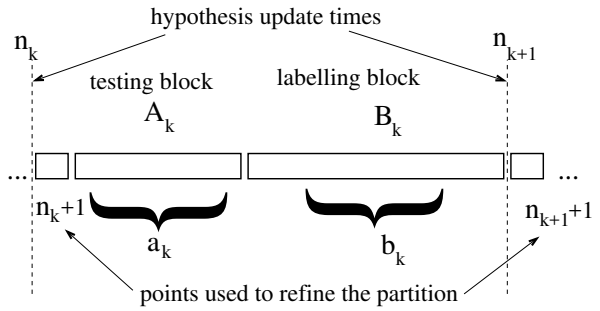

Figure 4: Natural numbers divided into blocks.

The testing block, A_k, or rather the corresponding subsample $\sigma[A_k]$, is used for empirical testing of the probability value $P[Y = 1 \mid X \in I]$ for every cell I (a half-open cyclic interval) of the partition, $\hat{\mathcal{Q}}_{k-1}$, generated by the current partitioning set $\mathcal{Q} = \mathcal{Q}_{k-1}$. And the labelling block, B_k, is used to generate the predicted labels based on the partitioning rule under the possibly updated partition.

We only perform the testing and then generate a new hypothesis if every cell of the partition $\hat{\mathcal{Q}}_{k-1}$, contains sufficiently many points of the testing sample, and every cell of the partition $\hat{\mathcal{P}}_k$ contains sufficiently many points of the labelling sample. If this is the case, then for every interval $I \in \hat{\mathcal{Q}}_{k-1}$ satisfying $P_{\sigma[A_k]}[Y = 1 \mid X \in I] \in (\epsilon_k, 1 - \epsilon_k)$ we update the partitioning set $\mathcal{Q}_{k-1}$ by adding to it the set $\mathcal{P}_k \cap I$. After the update is done, we generate the new hypothesis, given by the histogram rule on the updated partition $\hat{\mathcal{Q}}_k$ and the sample $\sigma[B_k]$. Otherwise, no testing and no partitioning set update are made, and the rule returns the previous hypothesis output at the moment $n = n_{k-1}$.

At the beginning, the set of partitioning points is initialized to be an empty set, $\mathcal{Q}_0 = \emptyset$, hence the corresponding partition $\hat{\mathcal{Q}}_0$ is trivial, $\hat{\mathcal{Q}}_0 = \{\mathbb{S}^1\}$, having the entire domain $\Omega = \mathbb{S}^1$ as the only cell. Thus, for $n = 1$, the rule g_1 on the labelled sample input (x_1, y_1) assigns the label y_1 to every point of the domain. In this way, (x_1, y_1) is the subsample $\sigma[B_1]$, the first labelling block is $B_1 = \{1\}$, and $b_1 = 1$. The first testing block is empty: $A_1 = \emptyset$, and $a_1 = 0$.

Further values a_k, b_k will be chosen recursively so as to grow fast enough and guarantee that almost surely along the sample path, the hypothesis is being updated infinitely often. A conditioning argument shows the misclassification error does not increase with each step. As to the universal consistency, notice that in the cases where the regression function is identically zero or one (that is, the same label is assigned to every point), the division of cells will almost surely never occur, but obviously the partitioning rule with only one cell is still consistent. However, a certain variation of existing results on the universal consistency of partitioning rules (which require the cell diameter to go to zero in probability) does the job.

Now an outline of the article structure. The sections 2–5 lay the technical groundwork for our learning rule, presented and studied in the sections 6–8. We start with a revision of the standard model of supervised statistical learning in Sect. 2. The concavity properties of the error function $L(p, n)$ are dealt with in Section 3. In Sect. 4 we discuss the partitioning rules, and here appears the key technical result, Lemma 11, showing how to partition cells without increasing the learning error. Section 5 is devoted to the cyclic orders and the probability measures on the circle. Our learning rule is formally described in Section 6. In Sect. 7 we show the rule has a universally monotone expected error, and in Sect. 8 we prove the universal consistency. A few concluding remarks in Sect. 9 were motivated by the referees' comments.

2. Learning rules

Let $\Omega = (\Omega, \mathcal{A})$ be a measurable space, that is, a set equipped with a sigma-algebra of subsets $\mathcal{A}$. The main example is where $\mathcal{A}$ is the smallest sigma-algebra of subsets containing all open balls with regard to a metric, making Ω a complete separable metric space. Such a measurable space $(\Omega, \mathcal{A})$ is called a *standard Borel space*. The elements of $\mathcal{A}$ are known as *Borel sets.*

The Borel structure remembers very little of the generating metric, in the following sense. Two standard Borel spaces admit a Borel isomorphism (a bijection preserving the sigma-algebras) if and only if they have the same cardinality. Thus, there are only the following isomorphism types of standard Borel spaces: finite ones with n elements for each natural n, countably infinite (all of which are isomorphic to the natural numbers with the sigma-algebra of all subsets), and those of cardinality continuum. For example, the Borel spaces associated to the real line, to $\mathbb{R}^d$, to the Hilbert space ℓ^2, to the Cantor set, etc., are all pairwise isomorphic as standard Borel spaces. See Kechris (1995) as a general reference.

In statistical learning theory the learning domain Ω is usually assumed to be a standard Borel space of cardinality continuum, which will be our standing assumption as well.

The product $\Omega \times \{0, 1\}$ now becomes a standard Borel measurable space in a natural way. The elements $x \in \Omega$ are known as *unlabelled points*, and the pairs $(x, y) \in \Omega \times \{0, 1\}$ are *labelled points.* A finite sequence of labelled points, $\sigma = (x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n) \in \Omega^n \times \{0, 1\}^n$, is a *labelled sample.*

A *classifier* in Ω is a mapping

$$T \colon \Omega \to \{0, 1\},$$

assigning a label to every point. The mapping is usually assumed to be measurable in order for things like the misclassification error to be well defined, although some authors are allowing for non-measurable maps, working with the outer measure instead.

Let $\tilde{\mu}$ be a probability measure defined on the measurable space $\Omega \times \{0, 1\}$. Denote (X, Y) a random element of $\Omega \times \{0, 1\}$ following the law $\tilde{\mu}$. The misclassification error of a classifier T is the quantity

$$\begin{aligned} L_{\tilde{\mu}}(T) &= \tilde{\mu}\{(x, y) \in \Omega \times \{0, 1\} \colon T(x) \neq y\} \\ &= P[T(X) \neq Y]. \end{aligned}$$

The *Bayes error* is the infimum (in fact, the minimum) of the misclassification errors of all the classifiers T defined on Ω:

$$L^* = L^*(\tilde{\mu}) = \inf_T L_{\tilde{\mu}}(T).$$

A *learning rule* in $(\Omega, \mathcal{A})$ is a mapping

$$g\colon \bigcup_{n=1}^{\infty} \Omega^n \times \{0,1\}^n \times \Omega \ni (\sigma, x) \mapsto g(\sigma)(x) \in \{0,1\}.$$

Again, the map above is usually assumed to be measurable. We denote g_n the restriction of g to $\Omega^n \times \{0,1\}^n \times \Omega$. For a labelled sample σ, we denote $g_n(\sigma)$ the binary function $\Omega \ni x \mapsto g_n(\sigma)(x) \in \{0,1\}$. Thought of as a subset of Ω on which the function takes value 1, this $g_n(\sigma)$ is also known as a *hypothesis* output by the rule g on the labelled sample input σ.

The labelled datapoints are modelled by a sequence of independent, identically distributed random elements $(X_n, Y_n) \in \Omega \times \{0,1\}$ following the law $\tilde{\mu}$. For each n, the *misclassification error* of the rule g_n is the random variable

$$L_{\tilde{\mu}} g_n = P\left[g_n(D_n)(X) \neq Y \mid D_n\right].$$

In other words, it is the error of the random classifier $g_n(D_n)$, where D_n is a random labelled n-sample.

Consider the probability measure $\mu = \tilde{\mu} \circ \pi^{-1}$ on Ω, where π is the first coordinate projection of $\Omega \times \{0,1\}$. Now define a finite measure μ_1 on Ω by $\mu_1(A) = \tilde{\mu}(A \times \{1\})$. Clearly, μ_1 is absolutely continuous with regard to μ. Define the *regression function*, $\eta\colon \Omega \to [0,1]$, as the Radon–Nikodým derivative

$$\begin{aligned} \eta(x) &= \frac{d\mu_1}{d\mu} \\ &= P[Y = 1 \mid X = x], \end{aligned}$$

that is, the conditional probability for x to be labelled 1. The pair (μ, η) completely defines the measure $\tilde{\mu}$ and is often more convenient to use. Thus, a learning problem in a measurable space $(\Omega, \mathcal{A})$ can be alternatively given either by the measure $\tilde{\mu}$ on $\Omega \times \{0,1\}$ or by the pair (μ, η).

A rule g is *consistent* under $\tilde{\mu}$ if

$$L_{\tilde{\mu}} g_n \xrightarrow{p} L^*(\tilde{\mu}), \text{ that is, } \mathbb{E} L_{\tilde{\mu}} g_n \to L^*(\tilde{\mu}),$$

and *universally consistent* if g is consistent under every probability measure on $\Omega \times \{0,1\}$.

The following simple observation allows to delay the hypothesis update.

Lemma 1 *Let F be a random function from the natural numbers to itself with the property $F(n) \leq n$ for all n. Given a learning rule g, define a rule g_F by $(g_F)_n(\sigma) = g_{F(n)}(\sigma[(1, \ldots, F(n))])$. Assume that $F(n) - n \to 0$ in probability. If g is consistent, then g_F is consistent.*

Proof Given $\epsilon > 0$, find N so that for $n \geq N$, we have $P(L_n(g_n) < L^* + \epsilon) > 1 - \epsilon$ and $P[F(n) = n] > 1 - \epsilon$ whenever $n \geq N$. It follows that for such n

$$\begin{aligned} P(L(g_F)_n < L^* + \epsilon) &\geq P(F(n) = n \text{ and } L_n(g_n) < L^* + \epsilon) \\ &> 1 - 2\epsilon. \end{aligned}$$

■

3. Error in a trivial one-point domain

Consider the "natural" learning rule g in the one-point domain $\Omega = \{*\}$, which is the majority vote among n i.i.d. random labels $D_n = (Y_1, \ldots, Y_n)$ following the Bernoulli law. To avoid ties, we assume n odd (so if n is even, the n-th label is not considered).

So, let $n = 2k + 1$, $k \geq 0$ and $p \in [0, 1]$. We have:

$$P[g_n(D_n) = 0] = P\left[\frac{1}{n}\sum_{i=1}^{n} Y_i < 1/2\right] = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i},$$

$$P[g_n(D_n) = 1] = \sum_{i=0}^{k} \binom{n}{i} p^{n-i} (1-p)^i.$$

The expression for the expected learning error becomes

$$\begin{aligned} L\,(p, n) &= P[g_n(D_n) = 0] \cdot P[Y = 1] + P[g_n(D_n) = 1] \cdot P[Y = 0] \\ &= \sum_{i=0}^{k} \binom{n}{i} p^{i+1} (1-p)^{n-i} + \sum_{i=0}^{k} \binom{n}{i} p^{n-i} (1-p)^{i+1}. \end{aligned} \tag{3}$$

As $n \to \infty$, simple ball volume considerations in the Hamming cube $\{p, 1-p\}^n$ show that $L(p, n)$ converges to the Bayes error,

$$L^*(p) = \min\{p, 1-p\}.$$

For the first part of the following statement, see Sect. VII of Cover and Hart (1967), or Problem 6.12 in Devroye et al. (1996); as neither source contains a proof, we present it here.

Lemma 2 *The convergence $L\,(p, n) \to L^*(p)$ is monotone along the odd values of n. More exactly, for every $n = 2k + 1$,*

$$L\,(p, n) - L\,(p, n+2) = \binom{n}{k} (2p-1)^2 p^{k+1} (1-p)^{k+1} \geq 0. \tag{4}$$

Proof Let $\{p, 1-p\}^n$ denote the Hamming cube $\{0,1\}^n$ with the product measure $\{p, 1-p\}^{\otimes n}$. Denote $w(\tau) = \sum_{i=1}^n \tau_i$ the weight of the string τ. The learning error $L(p, 2k+1)$ is the expected value of the random variable $L(g_n(D_n))$, represented by the function

$$\{p,1-p\}^n \ni \tau \mapsto L(g_n(\tau)) = \begin{cases} p, & \text{if } w(\tau) \le k, \\ 1-p, & \text{if } w(\tau) \ge k+1. \end{cases}$$

We will study the behaviour of this expected value under the transition $k \mapsto k+1$, that is, $n \mapsto n+2$. In the latter case,

$$\tau \mapsto L(g_{n+2}(\tau)) = \begin{cases} p, & \text{if } w(\tau) \le k+1, \\ 1-p, & \text{if } w(\tau) \ge k+2. \end{cases}$$

For every $i = 0, 1, \ldots, n$, denote C_i the cylindrical set of all strings $\tau \in \{0,1\}^{n+2}$ satisfying $w(\tau_n) = i$, where τ_n is the n-prefix of τ. We have

$$\begin{aligned} L(p,n) - L(p,n+2) &= \mathbb{E}\, L(g_n(\tau_n)) - \mathbb{E} L(g_{n+2}(\tau)) \\ &= \sum_{i=0}^{n} \int_{C_i} \left(L(g_n(\tau_n)) - L(g_{n+2}(\tau))\right)\, d\{p,1-p\}^{\otimes n}. \end{aligned}$$

If the prefix τ_n of τ satisfies $w(\tau_n) \le k-1$, then $w(\tau) \le k+1$, and so on C_i the two error variables are identical. The same applies if $w(\tau_n) \ge k+2$. We have

$$L(p,n+2) - L(p,n) = \sum_{i=k,k+1} \int_{C_i} \left(L(g_n(\tau_n)) - L(g_{n+2}(\tau))\right)\, d\{p,1-p\}^{\otimes n}.$$

Case $i = k$. The value of the error variable changes from p to $1-p$ for the strings of the form $\tau = \tau_n 11$, whose total number is $\binom{n}{k}$. For every such string, the singleton $\{\tau_n 11\}$ has measure $p^{k+2}(1-p)^{k+1}$. Other strings in C_k keep the same error value, p. We conclude:

$$\int_{C_k} \left(L(g_n(\tau_n)) - L(g_{n+2}(\tau))\right)\, d\{p,1-p\}^{\otimes n} = (2p-1)\binom{n}{k} p^{k+2}(1-p)^{k+1}.$$

Case $i = k+1$. The error value changes from $1-p$ to p for the strings of the form $\tau_n 00$, whose total number is $\binom{n}{k+1} = \binom{n}{k}$:

$$\int_{C_{k+1}} \left(L(g_n(\tau_n)) - L(g_{n+2}(\tau))\right)\, d\{p,1-p\}^{\otimes n} = (1-2p)\binom{n}{k} p^{k+1}(1-p)^{k+2}.$$

Thus,

$$\begin{aligned} L(p,n) - L(p,n+2) &= (2p-1)\binom{n}{k} p^{k+2}(1-p)^{k+1} + (1-2p)\binom{n}{k} p^{k+1}(1-p)^{k+2} \\ &= (2p-1)\binom{n}{k} p^{k+1}(1-p)^{k+1}\left[p - (1-p)\right] \\ &= (2p-1)^2 \binom{n}{k} p^{k+1}(1-p)^{k+1} \ge 0. \end{aligned}$$

■

Lemma 3 *The convergence $L(p,n) \to L^*(p)$ is uniform in p along the odd values of n.*

Proof The sequence of non-negative continuous functions $L(p,n) - L^*(p)$ with n odd converges to zero pointwise and monotonically on a compact set $[0,1]$, which implies the uniform convergence (Dini's theorem). ■

A real function f is *concave* if for all x, y in the domain of f and each $t \in [0,1]$, $f((1-t)x + ty) \geq (1-t)f(x) + tf(y)$. Equivalently, for any collection of x_i and t_i with $\sum_i t_i = 1$, we have $f\left(\sum_i t_i x_i\right) \geq \sum_i t_i f(x_i)$.

Lemma 4 *For every n odd, the function $L(p,n)$ is concave in a sufficiently small neighbourhood of $p = 1/2$.*

Proof For $n = 1$, $L(p,1) = 2p(1-p) = 2(p - p^2)$ is globally concave. Write

$$p(1-p) = \frac{1}{4} - \left(p - \frac{1}{2}\right)^2.$$

Then, by Lemma 2,

$$\begin{aligned} L(p, n+2) &= L(p,n) - \binom{n}{k}(2p-1)^2 p^{k+1}(1-p)^{k+1} \\ &= L(p,n) - 4\binom{n}{k}\left(p - \frac{1}{2}\right)^2 \left[\frac{1}{4} - \left(p - \frac{1}{2}\right)^2\right]^{k+1}. \end{aligned}$$

The lowest term in the Taylor expansion of the second polynomial around $p = 1/2$ is of second degree, $(p - 1/2)^2$ with negative coefficient $-4\binom{n}{k}(1/4)^{k+1}$, meaning the function is concave in a sufficiently small neighbourhood of $p = 1/2$. As the sum of concave functions is concave, we conclude by induction. ■

If f is a bounded real-valued function defined on some set X, it is easy to see that there exists the smallest concave function of the same domain of definition majorizing f. This function is called the *concave envelope* of f, and we will denote it $\widehat{f}$.

Lemma 5 *Given n and $\epsilon > 0$, there is N so that for all values $\epsilon \leq p \leq 1 - \epsilon$*

$$\widehat{L}(p, N) \leq L(p, n).$$

Proof Choose $\delta > 0$ so that in the δ-neighbourhood of $p = 1/2$, the function $L(p,n)$ is concave (Lemma 4). Since the only values where $L(p,n) = p$ are $p = 0, 1/2, 1$, there is $\gamma > 0$ so small that for all $p \in [\epsilon, 1/2 - \delta]$ we have $L(p,n) > p + \gamma$. By the intermediate value theorem, there is $q \in [1/2 - \delta, 1/2)$ with $L(q,n) = q + \gamma$. Reducing γ further if needed, we may assume that such a q is unique. By the uniform convergence of error functions

(Lemma 3), there is N with $L(p,N) \leq p+\gamma$ for all p. The monotonicity of the error function (Lemma 2) implies $L(p,N) \leq L(p,n)$ for all p.

The function

$$\psi(p) = \begin{cases} p+\gamma, & \text{if } p \leq q, \\ L(p,n), & \text{if } q \leq p \leq 1/2, \end{cases}$$

extended by symmetry over $[1/2,1]$, is concave over $[0,1]$. (Indeed, for every $p \in [q,1/2]$, the gradient of the chord joining $(p, L(p,n))$ with $(q, q+\epsilon) = (q, L(q,n))$ is less than 1.) By the construction, we have $L(p,N) \leq \psi(p)$. Therefore, for all p,

$$\widehat{L}(p,N) \leq \psi(p).$$

Since on the interval $[\epsilon, 1-\epsilon]$ we have $\psi(p) \leq L(p,n)$, we conclude. ■

The proof of the following is left out as an exercise in Devroye et al. (1996), problem 5.6(2).

Lemma 6 *Let $n = 2k+1$, where $k \geq 1$. Up to higher degree terms, $L(p,n)$ at zero has the form*

$$L(p,n) = p + \binom{2k+1}{k} p^{k+1} + o(p^{k+1}).$$

Proof Consider the expression for the learning error (Eq. 3):

$$L(p,n) = \sum_{i=0}^{k} \binom{n}{i} p^{i+1}(1-p)^{n-i} + \sum_{i=0}^{k} \binom{n}{i} p^{n-i}(1-p)^{i+1}.$$

The monomial of the lowest order in the right hand sum comes from the term corresponding to $i=k$ and equals exactly $\binom{2k+1}{k}p^{k+1}$. The monomial of the lowest order in the left hand sum corresponds to $i=0$ and equals p. Thus, it is enough to show that in the polynomial

$$\sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i} = \sum_{i=0}^{k} \binom{n}{i} p^i \sum_{j=0}^{n-i} \binom{n-i}{j} (-1)^j p^j$$

(the l.h.s. after we took p out) all the powers of p between $m=1$ and $m=k$ inclusive vanish. Let $1 \leq m \leq k$. Using the classical binomial formula, we calculate the coefficient of p^m:

$$\begin{aligned} \sum_{i+j=m} \binom{n}{i}\binom{n-i}{j}(-1)^j &= \frac{n!}{m!(n-m)!} \sum_{j=0}^{m} \frac{m!(-1)^j}{j!(m-j)!} \\ &= \frac{n!}{m!(n-m)!}(1-1)^m \\ &= 0. \end{aligned}$$

■

Remark 7 *Note that*

$$\binom{2k+1}{k} = \binom{2k+1}{k+1} = \binom{2k}{k} + \binom{2k}{k+1},$$

which is how the expression for the coefficient appears in Devroye et al. (1996). Also, Lemma 6 is false for $k = 0$ (that is, $n = 1$), in which case $L(p, 1) = 2p - 2p^2$. There are two reasons why the proof fails: first, $p^{k+1} = p$, and second, we cannot conclude that for $m = k = 0$ the power $(1-1)^m$ vanishes.

The following key technical result together with its corollary underpins our learning rule by saying that a certain amount of empirical error when testing a cell for partitioning is admissible. An application to random partitions appears in Lemma 11.

Lemma 8 *Given n odd and $t \in (0, 1]$, for all $N = N(n, t)$ (odd) large enough,*

$$P[\text{binomial}(p, N) > tN] \cdot \widehat{L}(p, N) + P[\text{binomial}(p, N) \leq tN] \cdot L(p, N) \leq L(p, n)$$

over all $p \in [0, 1/2]$.

Proof Let $n = 2k + 1$. By force of Lemma 6,

$$\lim_{p \to 0} \frac{L(p, n+2) - p}{L(p, n) - p} = 0,$$

and for some $\delta > 0$ small enough,

$$L(p, N) - p \leq L(p, n+2) - p < \frac{1}{2}(L(p, n) - p)$$

when $p \in [0, \delta]$ and $N > n$ is odd. Rewrite the inequality as

$$L(p, N) < \frac{1}{2}(L(p, n) + p).$$

Now note a very rough estimate

$$\begin{aligned} P[\text{binomial}(p, N) \geq tN] &= \sum_{i=\lceil tN \rceil}^{N} \binom{N}{i} p^i (1-p)^{N-i} \\ &\leq p^{\lceil tN \rceil} \sum_{i=0}^{N} \binom{N}{i} \\ &\leq p^{tN} 2^N \\ &= (2^{t^{-1}} p)^{tN}. \end{aligned}$$

When $N > t^{-1}(k+1)$, thanks to Lemma 6, the ratio of the polynomials $P[\text{binomial}(p, N) > tN]$ and $(1/2)L(p, n) - p/2$ converges to zero as $p \to 0$, and so for some $\delta' > 0$, we have

$$P[\text{binomial}(p, N) > tN] < \frac{1}{2}(L(p, n) - p)$$

as long as $p \in [0, \delta']$.

Use Lemma 5 to further increase N_0 so that for all $N \geq N_0$ and $p \in [\min\{\delta, \delta'\}, 1/2]$,

$$\widehat{L}(p, N) \leq L(p, n).$$

For p in the interval $[\min\{\delta, \delta'\}, 1/2]$ and N sufficiently large, we have

$$\begin{aligned}
&P[\text{binomial}(p, N) > tN] \cdot \widehat{L}(p, N) + P[\text{binomial}(p, N) \leq tN] \cdot L(p, N) \\
&\leq P[\text{binomial}(p, N) > tN] \cdot L(p, n) + (1 - P[\text{binomial}(p, N) > tN]) \cdot L(p, n) \\
&= L(p, n),
\end{aligned}$$

and if $p \leq \min\{\delta, \delta'\}$,

$$\begin{aligned}
&P[\text{binomial}(p, N) > tN] \cdot \widehat{L}(p, N) + P[\text{binomial}(p, N) \leq tN] \cdot L(p, N) \\
&\leq P[\text{binomial}(p, N) > tN] + L(p, N) \\
&\leq \frac{1}{2}(L(p, n) - p) + \frac{1}{2}(L(p, n) + p) \\
&= L(p, n).
\end{aligned}$$

■

Lemma 9 *Given n odd and $t \in (0, 1/2)$, for all $N = N(n, t)$ (odd) large enough,*

$$\begin{aligned}
P[\text{binomial}(p, N) \in (tN, (1-t)N)] \cdot \widehat{L}(p, N) + \\
P[\text{binomial}(p, N) \notin (tN, (1-t)N)] \cdot L(p, N) \\
\leq L(p, n),
\end{aligned}$$

over all $p \in [0, 1]$.

Proof Let $N = N(n, t)$ be chosen as in Lemma 8. Write the expression on the left hand side above as

$$\begin{aligned}
&P[\text{binomial}(p, N) \in (tN, (1-t)N)] \cdot \widehat{L}(p, N) + \\
&P[\text{binomial}(p, N) \leq tN] \cdot L(p, N) + P[\text{binomial}(p, N) \geq (1-t)N] \cdot L(p, N).
\end{aligned}$$

For $p \in [0, 1/2]$, bounding the third term by $P[\text{binomial}(p, N) \geq (1-t)N] \cdot \widehat{L}(p, N)$, we get the expression in Lemma 8. For $p \in [1/2, 1]$, we apply the same bound to the second term, and use the symmetry of the binomial distribution and the functions $L(p, n)$ and $\widehat{L}(p, n)$:

$$\begin{aligned}
&\leq P[\text{binomial}(p, N) < (1-t)N] \cdot \widehat{L}(p, N) + P[\text{binomial}(p, N) \geq (1-t)N] \cdot L(p, N) \\
&= P[\text{binomial}(1-p, N) > tN] \cdot \widehat{L}(1-p, N) + P[\text{binomial}(1-p, N) \leq tN] \cdot L(1-p, N),
\end{aligned}$$

again applying Lemma 8. ■

4. Partitioning rules

A partition, $\mathscr{P}$, of the domain (a standard Borel space) Ω is a finite family of disjoint measurable subsets, called cells, covering Ω. To a partition $\mathscr{P}$ and a labelled sample σ associate a classifier, $h_{\mathscr{P}}$, as follows. The predicted label of a point x is determined by the majority vote among the elements of a labelled sample contained in the same cell as x. To avoid voting ties, we will remove if necessary the datapoint having the largest index, leaving an odd number of labels for the vote. The labels of those cells entirely missed by σ are not relevant, and for instance can be chosen at random, or always be equal to 1. (In our future rule, this will almost surely never happen.)

Lemma 10 *Let $\mathscr{P}$ be a partition of the domain. Denote $p = P[Y = 1]$. Then, conditionally on each cell of the partition containing at least n sample points, the expected error of the histogram classifier satisfies*

$$\mathbb{E}L(h_{\mathscr{P}}) \leq \widehat{L}(p, n).$$

Proof Denote $p_C = P[Y = 1 \mid X \in C]$. Then $p = \sum \mu(C)p_C$. Using the monotonicity of the function $L(p, n)$ in n (Lemma 2),

$$\begin{aligned} P[h_{\mathscr{P}}(X) \neq Y \colon \sharp\sigma \restriction C \geq n,\ C \in P] &= \sum_{C \in P} \mu(C) P[h_{\mathscr{P}}(X) \neq Y \mid X \in C, \sharp\sigma \restriction C \geq n] \\ &\leq \sum_{C \in P} \mu(C) P[h_{\mathscr{P}}(X) \neq Y \mid X \in C, \sharp\sigma \restriction C = n] \\ &= \sum_{C \in P} \mu(C) L(p_C, n) \\ &\leq \widehat{L}(p, n). \end{aligned}$$

■

A partitioning rule $h = (h_{\mathscr{P}_n})$ is based on a sequence of partitions of the domain, $(\mathscr{P}_n)$. Those partitions can be either deterministic and fixed in advance (as the histogram rule), or random, for instance determined by the (unlabelled) elements of a subsample. To talk about random partitions, one needs of course a standard Borel structure on the family of partitions that may emerge. This happens naturally, for example, in our case, where the partitions are into cyclic intervals of the circle: the family of all such partitions is naturally identifiable with a standard Borel space.

There are various known sufficient conditions for a partitioning rule to be consistent. For example (Devroye et al. (1996), Th. 6.1) this is the case if Ω is a Euclidean domain, and the cell $C(X)$ containing a random element $X \in \Omega$ has two properties: the diameter of $C(X)$ converges to zero in probability, and the number of points of a sample contained in $C(X)$ converges to infinity in probability.

For a labelled sample $\sigma = (x_1, \ldots, x_n, y_1, \ldots, y_n)$, we denote P_σ the corresponding empirical probability. In particular,

$$P_\sigma[Y = 1] = \frac{1}{n} \sharp\{i \colon y_i = 1\}.$$

The following lemma is our entire learning rule in a nutshell. It demonstrates the protocol for partitioning cells without increasing the error of the partitioning rule.

Lemma 11 *Let the domain Ω be equipped with a learning problem (μ, η). Let $\mathscr{P}$ be a random finite partition of Ω, and σ, ς, τ three jointly independent i.i.d. random labelled samples. Suppose also that $\mathscr{P}$ and ς are independent. Denote n the size of σ and N the size of ς. Let $0 < \epsilon < 1/2$, and let $N(n, \epsilon)$ be chosen as in Lemma 9. Suppose $N \geq N(n, \epsilon)$. Define a random partition $\mathscr{Q}$ as follows: if $P_\varsigma[Y = 1] \in (\epsilon, 1 - \epsilon)$, then $\mathscr{Q} = \mathscr{P}$, otherwise $\mathscr{Q} = \{\Omega\}$. Conditionally on the event that every cell of $\mathscr{P}$ contains at least N points of τ,*

$$\mathbb{E}L(h_{\mathscr{Q}}(\tau)) \leq \mathbb{E}L(h_{\{\Omega\}}(\sigma)).$$

Proof Denote for short the events

$$A = [P_\varsigma[Y = 1] \in (\epsilon, 1 - \epsilon)] \text{ and } B = [\text{ for all cells } C \in \mathscr{P}, \sharp\tau \restriction C \geq N].$$

Denoting $p = P[Y = 1] = \mathbb{E}\eta$, we have

$$P(A) = P[\text{binomial}(p, N) \in (\epsilon, 1 - \epsilon)],$$

and since the events A and B are independent,

$$\begin{aligned}
\mathbb{E}(L(h_{\mathscr{Q}}(\tau)) \mid B) &= P(A)\mathbb{E}(L(h_{\mathscr{P}}(\tau)) \mid B) + (1 - P(A))\mathbb{E}(L(h_{\{\Omega\}}(\tau)) \mid B) \\
\text{(Lemma 10)} &\leq P(A)\widehat{L}(p, N) + (1 - P(A))L(p, N) \\
\text{(Lemma 9)} &\leq L(p, n) \\
&= L(h_{\{\Omega\}}(\sigma)).
\end{aligned}$$

■

For $x \in \Omega$, let $C(x)$ denote the cell of the partition $\mathscr{P}_n$ containing x, and $N(x)$ the number of elements of σ belonging to the cell $C(x)$. The following is a variation on Theorem 6.1 in Devroye et al. (1996).

Theorem 12 *Let (μ, η) be a learning problem on a standard Borel space Ω. Let $(\mathscr{P}_k)$ be a sequence of random partitions of Ω, and let (D_k) be a sequence of finite i.i.d. labelled samples. Suppose that $\mathbb{E}(\eta \mid \mathscr{P}_k) \to \eta$ in probability, and the number $N(X)$ of elements of D_k in a random cell $C(X) \in \mathscr{P}_k$ goes to infinity in probability as $k \to \infty$. Then the expected error $\mathbb{E}h_{\mathscr{P}_k}(D_k)$ converges to $L^* = L^*(\mu, \eta)$ as $k \to \infty$.*

Proof Denote

$$\hat{\eta}_k(x) = \frac{1}{N(x)} \sum_{i:\ X_i \in C(x)} Y_i$$

the empirical regression function. According to Corollary 6.1 in Devroye et al. (1996), it is enough to show that $\mathbb{E}\,|\hat{\eta}_k(X) - \eta(X)| \to 0$. By the triangle inequality,

$$\mathbb{E}\,|\hat{\eta}_k(X) - \eta(X)| \leq \mathbb{E}\,|\hat{\eta}_k(X) - \mathbb{E}(\eta \mid \mathscr{P}_k)(X)| + \mathbb{E}\,|\mathbb{E}(\eta \mid \mathscr{P}_k)(X) - \eta(X)|\,.$$

The first term converges to zero through conditioning on $N(X)$ and using the fact that $N(x)\hat{\eta}_k(X)$ is distributed as binomial$(N(x), \mathbb{E}(\eta \mid \mathscr{P}_k)(x))$, it is exactly the first part of the proof of Theorem 6.1 in Devroye et al. (1996). The convergence to zero of the second term is our assumption. ■

5. Cyclic orders

Recall again a basic theorem in descriptive set theory: every standard Borel space of uncountable cardinality is isomorphic to the unit interval with its usual Borel structure (see Th. 15.6 in Kechris (1995)). In particular, every such space is Borel isomorphic to the unit circle:

$$\mathbb{S}^1 = \{e^{2\pi \mathbf{i} t} \colon t \in [0,1)\} \subseteq \mathbb{C}.$$

Thus, given an arbitrary domain Ω (a standard Borel space), we can fix a Borel isomorphism with the circle $\mathbb{S}^1$ and work directly with the circle from now on.

This is the same thing as choosing on Ω a cyclic order with certain properties, and we will give a minimum of necessary definitions. A *cyclic order* on a set X is a ternary relation, denoted $[x, y, z]$, satisfying the following properties:

1. Either $[x, y, z]$ or $[z, y, x]$, but not both.
2. $[x, y, z]$ implies $[y, z, x]$.
3. $[x, y, z]$ and $[y, u, z]$ implies $[x, u, z]$.

A linearly ordered set $(X, \leq)$ supports a cyclic order given by

$$[x, y, z] \text{ if and only if } x < y < z \text{ or } y < z < x \text{ or } z < x < y.$$

The circle has a natural cyclic order, where $x < y < z$ whenever y is between x and z when we traverse the arc from x to z in the counter-clockwise direction (although clockwise would do just as well). Here is a definition not requiring geometric notions: for any $t, s, w \in [0, 1)$, $\left[e^{2\pi \mathbf{i} t}, e^{2\pi \mathbf{i} s}, e^{2\pi \mathbf{i} w}\right]$ if and only if $[t, s, w]$, where the cyclic order on the interval is defined as above. (See Świerczkowski (1959), remark to Lemma 1.)

Any two points x, y of a cyclically ordered set define an open interval, (x, y), consisting of all points z with $[x, z, y]$. Similarly one defines other types of intervals. We will be interested in half-open intervals of the form $[x, y) = (x, y) \cup \{x\}$. A cyclic order on a standard Borel space Ω is Borel if the corresponding ternary relation is a Borel subset of Ω^3, which in particular implies that every interval is a Borel set.

It is easy to verify that the Vapnik–Chervonenkis dimension of the family of all intervals (open, closed, and half-open) of a cyclically ordered set with at least 3 points is exactly 3. Indeed, every three-point set is shattered, while the axioms imply that a set of four points cannot be shattered.

Fixing any point ξ of a cyclically ordered set X, we obtain a linear order $<_\xi$ on X, with ξ as the smallest element, and for all other elements, $y <_\xi z$ if and only if $[\xi, y, z]$. Now the original cyclic order is exactly the cyclic order defined by the linear order $<_\xi$.

A cyclic order is *dense* if for every x, y, $x \neq y$, there is z with $[x, y, z]$. A cyclic order is *order-separable* if there is a countable subset meeting each non-empty open interval. Say that a cyclic order is *Dedekind complete* if every non-empty proper subset C has the greatest lower bound with regard to the linear order $<_\xi$ for every $\xi \notin C$. It can be shown that a standard Borel space equipped with a Dedekind complete dense order-separable Borel order admits a Borel isomorphism with the circle $\mathbb{S}^1$ preserving the cyclic order. Thus, technically, we construct our learning rule by fixing a cyclic order on a domain having the above listed properties, but it is more convenient to work by directly identifying the domain with the circle $\mathbb{S}^1$ and its standard cyclic order.

A mapping $f\colon X \to Y$ between two cyclically ordered sets is *monotone* if for all $x, y, z \in X$, whenever $f(x), f(y), f(z)$ are all pairwise distinct, we have $[x, y, z]$ if and only if $[f(x), f(y), f(z)]$. This is equivalent to saying that for some (or any) $\xi \in X$, the mapping f is monotone non-decreasing with regard to the linear orders $<_\xi$ on X and $<_{f(\xi)}$ on Y. A monotone map between two linearly ordered sets is monotone in this sense (but the converse does not hold). One can also talk of monotone maps between a cyclically ordered set and linearly ordered set. The composition of two monotone maps is monotone.

Perhaps it would be helpful to mention that the exponential map $\mathbb{R} \to \mathbb{S}^1$ is monotone on any interval of unit length, but not on the entire real line: for instance, $0 < 0.5 < 1.25$, therefore $[0, 0.5, 1.25]$ with regard to the cyclic order on $\mathbb{R}$, but the corresponding images $e^0 = 1$, $e^{\pi \mathbf{i}} = -1$ and $e^{\pi \mathbf{i}/2} = \mathbf{i}$ satisfy $[1, \mathbf{i}, -1]$, that is, $[1, -1, \mathbf{i}]$ does not hold. Similarly, the two-fold cover of $\mathbb{S}^1 \to \mathbb{S}^1$, $x \mapsto x^2$, is not cyclically monotone. On the contrary, every orientation-preserving self-homeomorphism of $\mathbb{S}^1$ is. It is further easily seen that every monotone map from the circle $\mathbb{S}^1$ to itself is Borel.

Say that y is a *successor* of x in a finite cyclically ordered set $\mathcal{P}$, if for all $z \in \mathcal{P} \setminus \{x, y\}$ one has $[x, y, z]$, that is, $x \neq y$ and $[x, z, y]$ does not happen. Clearly, the successor of a given element always exists, provided $|\mathcal{P}| \geq 2$, and is unique. Let now $\mathcal{P}$ be a finite subset of a cyclically ordered set X. Then $\mathcal{P}$ defines a partition of X into half-open intervals $[x, y)$, for all pairs $x, y \in \mathcal{P}$ where y is the successor of x in $\mathcal{P}$. We will denote this partition $\hat{\mathcal{P}}$. If $|\mathcal{P}| \leq 1$, then by definition the corresponding partition is trivial, $\hat{\mathcal{P}} = \{\Omega\}$. (If there is a single point, x, in $\mathcal{P}$, then one may say the only half-open interval contained in $\hat{\mathcal{P}}$ is $[x, x) = \Omega$.)

Lemma 13 *Let $f\colon X \to Y$ be a surjective monotone map between two cyclically ordered sets, and let $\mathcal{P} \subseteq X$ be a finite subset. Then every half-open interval in the partition $\widehat{f(\mathcal{P})}$ of Y defined by $f(\mathcal{P})$ is the image of some interval of the partition $\hat{\mathcal{P}}$ of X defined by $\mathcal{P}$.*

Proof Let $x, y \in \mathcal{P}$, where $b = f(y)$ is the successor of $a = f(x)$ in $f(\mathcal{P})$. Denote x' the maximal element in the finite set $f^{-1}(a)$ with regard to the linear order $<_y$. The interval $[x', y)$ contains no other elements of $f^{-1}(a)$. Now let y' be the minimal element in the finite set $f^{-1}(b)$ with regard to the linear order $<_x$. The interval $[x', y') \subseteq [x', y)$ still contains no elements of $f^{-1}(a)$ other than x', and no elements of $f^{-1}(b)$ other than y'. Then y' is the successor of x' in $\mathcal{P}$: any element w of $\mathcal{P}$ strictly between those two would have either satisied $[a, f(w), b]$ or coinside with a or b, both of which are impossible.

We claim that in this case, $f[x', y') = [a, b)$. Let $w \in (a, b)$, that is, $[a, w, b]$. Since f is surjective, there is $z \in X$ with $f(z) = w$. Because of monotonicity of f, we must have $[x', z, y']$, that is, $z \in (x', y')$. We conclude.

The trivial case $f(P) = \emptyset = \mathcal{P}$ is obvious. Finally, suppose $f(\mathcal{P})$ only contains one element, a, that is, $f^{-1}(a) = \mathcal{P}$. If Y only contains one element other than a, just select any interval of $\hat{\mathcal{P}}$ containing a preimage of this element. Else, we claim that all of $X \setminus \mathcal{P}$ is contained in only one interval of $\hat{\mathcal{P}}$. Indeed, let $x, y \in X \setminus \mathcal{P}$ be such that $f(x) \neq f(y)$. If x and y belong to different intervals of $\hat{\mathcal{P}}$, there exist $z, w \in \mathcal{P}$ with $[x, z, y]$ and $[x, y, w]$. This implies the incompatible properties $[f(x), a, f(y)]$ and $[f(x), f(y), a]$. From here the statement easily follows. ■

If $f\colon X \to Y$ is a measurable map between two standard Borel spaces and ν is a Borel probability measure on X, then the pushforward measure $\nu \circ f^{-1}$ on Y (which is also a Borel probability measure) is defined by letting $\nu \circ f^{-1}(A) = \nu(f^{-1}(A))$ for every Borel subset $A \subseteq Y$.

Lemma 14 *Given a Borel probability measure on the circle $\mathbb{S}^1$, there is a monotone (hence Borel) map $f\colon \mathbb{S}^1 \to \mathbb{S}^1$ with $\nu \circ i^{-1} = \mu$, where ν is the Haar measure on the circle.*

Proof The map $j\colon \mathbb{S}^1 \ni e^{2\pi \mathbf{i} t} \mapsto t \in [0, 1)$ is a Borel isomorphism. The push-forward measure $\nu \circ j^{-1}$ is the Lebesgue measure on the unit interval, λ, so j is an isomorphism between the Lebesgue probability spaces $(\mathbb{S}^1, \nu)$ and $([0, 1), \lambda)$. Denote $\mu' = \mu \circ j^{-1}$ the push-forward measure, and let F be the corresponding distribution function, $F(t) = \mu'(-\infty, t]$ $(= \mu'[0, t]$ for $t \in [0, 1])$. Let $i'\colon [0, 1) \ni \theta \mapsto \inf\{t \in [0, 1]\colon F(t) \geq \theta\} \in [0, 1)$. This is a monotone map with $\lambda \circ i'^{-1} = \mu'$. Finally, define $i = j^{-1} \circ i' \circ j$. This is the desired monotone map from $\mathbb{S}^1$ to itself that pushes forward ν to μ. ■

Lemma 15 *Given k, N, and $\delta > 0$, there exists $M = M(k, n, \delta)$ so large that for every Borel probability measure μ on the circle $\mathbb{S}^1$, if $k + M$ i.i.d. points following the law μ are chosen, then with confidence $1 - \delta$ every interval of the circle partition $\hat{\mathcal{P}}$ generated by the random finite set $\mathcal{P} = \{X_1, X_2, \ldots, X_k\}$ contains at least N points from among $X_{k+1}, X_{k+2}, \ldots, X_{k+M}$.*

Proof First, we prove the lemma for $\mathbb{S}^1$ with the Haar measure. Fix a sufficiently small $\epsilon > 0$. The probability of all the intervals of the circular partition made by $\mathcal{P} = \{x_1, x_2, \ldots, x_k\}$ to have arc length $\geq \epsilon$ is

$$\begin{aligned}(1 - 2\epsilon)(1 - 4\epsilon) \cdot \ldots \cdot (1 - (k-1)\epsilon) &> 1 - 2\epsilon - 4\epsilon - \ldots - (k-1)\epsilon \\ &= 1 - k(k-1)\epsilon.\end{aligned}$$

Thus, if we set $\epsilon = \delta/2k(k-1)$, then with confidence $1 - \delta/2$ every interval will have length $\geq \epsilon$.

Since the VC dimension of the family of all half-open intervals of the circle is $d = 3$, the sample size that suffices to empirically estimate the measure of all the intervals with confidence $1 - \delta/2$ to within the precision $\epsilon/2$ does not exceed

$$M' = \max\left\{\frac{48}{\epsilon}\log\frac{16e}{\epsilon}, \frac{8}{\epsilon}\log\frac{4}{\delta}\right\}.$$

(Here we use the bounds from Vidyasagar (2003), p. 269, Th. 7.8.) Set

$$M = \max\left\{M', \frac{2N}{\epsilon}\right\}.$$

For $n \geq M$, if σ is an n-sample, then, denoting ν_n the empirical measure, we have with confidence $1-\delta$ that for each interval I of the partition:

$$\nu_n(I) \geq \nu(I) - \frac{\epsilon}{2} \geq \frac{\epsilon}{2},$$

that is, I contains at least $n\epsilon/2 \geq N$ points of the sample.

Now let μ be an arbitrary measure on $\mathbb{S}^1$. Select a monotone map $i\colon \mathbb{S}^1 \to \mathbb{S}^1$ pushing forward the Haar measure ν to μ (Lemma 14). The random elements $X_1, \ldots, X_{n+k} \sim \mu$ can be written as $i(X'_1), \ldots, i(X'_{n+k})$, where X'_i are i.i.d. random elements following the law ν. According to Lemma 13, for every interval of the partition generated by $X_1, \ldots, X_k$ its intersection with $i(\mathbb{S}^1)$ is the image of some interval of the partition generated by $X'_1, \ldots, X'_k$, and so, according to the first part of our proof, with confidence $1-\delta$, all those intervals contain at least N sample points each. ■

Say that a finite subset $\mathcal{P}$ of the circle $\mathbb{S}^1$ is ϵ-dense with regard to a probability measure μ, if $\mathcal{P}$ meets every half-open interval of measure $\geq \epsilon$.

Lemma 16 *Let μ be a Borel probability measure on the circle $\mathbb{S}^1$, and let $X_1, X_2, \ldots$ be a sequence of i.i.d. random elements of $\mathbb{S}^1$ following the law μ. Let $\epsilon > 0$. Almost surely, starting with some k large enough, the random finite set $\{X_1, \ldots, X_k\}$ is ϵ-dense.*

Proof Fix a cyclically monotone parametrization $i\colon \mathbb{S}^1 \to \mathbb{S}^1$ pushing forward the Haar measure ν to μ (Lemma 14). Let $\mathcal{Q}$ be a cover of the circle with $n_0 \geq 2\epsilon^{-1}$ intervals of Haar measure between $\epsilon/3$ and $\epsilon/2$ each. Let $Y_1, \ldots, Y_k$ be i.i.d. random elements of $\mathbb{S}^1$ following the law ν. The probability for all of them to miss at least one of the intervals from $\mathcal{Q}$ is bounded by $n_0(1-\epsilon/3)^k$, and this is a summable sequence in k. By the Borel-Cantelli lemma, almost surely, starting with some k high enough, in every interval $I \in \mathcal{Q}$ there is contained at least one random element from among Y_i, $i = 1, 2, \ldots, k$. Let J be a cyclic interval with $\mu(J) \geq \epsilon$. The inverse image $i^{-1}(J)$ is again a cyclic interval by the definition of a monotone map, and $\nu(i^{-1}(J)) = \mu(I)$. The interval $i^{-1}(J)$ must wholly contain at least one interval $I \in \mathcal{Q}$. We conclude: almost surely, some $X_i = i(Y_i)$ belongs to J. ■

6. The learning rule

Select a sequence (ϵ_n) of positive numbers converging to zero, with $\epsilon_1 < 1/2$. Select a summable sequence of positive numbers (δ_n), that is, $\sum_{n=1}^{\infty} \delta_n < \infty$, satisfying $\delta_1 < 1$.

Put $a_1 = 0$, $b_1 = 1$, and further select N_k, a_k, b_k, $k > 1$, recursively as follows.

1. Let $N_k = N(b_{k-1}, \epsilon_k)$ be chosen as in Lemma 9, with $n = b_{k-1}$ and $t = \epsilon_k$.

In other words, for all $N \geq N_k$, N odd, and all $p \in [0,1]$,

$$P[\text{binomial}(p,N) \in (\epsilon_k N, (1-\epsilon_k)N)] \cdot \widehat{L}(p,N) + \\ P[\text{binomial}(p,N) \notin (\epsilon_k N, (1-\epsilon_k)N)] \cdot L(p,N) \leq L(p, b_{k-1}).$$

2. Choose $a_k = M(k, N_k, \delta_k)$ as in Lemma 15.

   That is, a_k is so large that for every Borel probability measure μ on the circle $\mathbb{S}^1$, if $k + a_k$ i.i.d. points $\sim \mu$ are chosen, then with confidence $1 - \delta_k$ every interval of the circle partition generated by the random finite set $\mathcal{P} = \{X_1, X_2, \ldots, X_k\}$ contains at least N_k elements from among $X_{k+1}, X_{k+2}, \ldots, X_{k+a_k}$.

3. Now choose b_k, again using Lemma 15, as $b_k = M(k, a_k, \delta_k)$.

   In full, for every Borel probability measure μ on $\mathbb{S}^1$, if $k + b_k$ i.i.d. points $\sim \mu$ are chosen, then with confidence $1 - \delta_k$ every interval of the partition generated by $\{X_1, X_2, \ldots, X_k\}$ contains at least a_k elements from among $X_{k+1}, \ldots, X_{k+b_k}$.

Set $n_1 = 1$ and further, recursively,

$$n_k = n_{k-1} + a_k + b_k + 1.$$

Denote $A_1 = \emptyset$, $B_1 = \{1\}$, and for $k > 1$,

$$A_k = (n_{k-1} + 2, \ldots, n_{k-1} + a_k + 1), \quad B_k = (n_{k-1} + a_k + 2, \ldots, n_{k-1} + a_k + b_k + 1).$$

Denote $\mathcal{P}_1 = \emptyset$ and for every $i \geq 2$ set

$$\mathcal{P}_i = \{x_{n_j+1} \colon j = 1, \ldots, i-1\}.$$

For a finite subset I of the positive integers and a labelled sample σ, we will denote $\sigma[I]$ a labelled subsample of σ consisting of all pairs labelled with $i \in I$, in the same order.

Recall further that for a finite set $\mathcal{Q}$, we denote $\hat{\mathcal{Q}}$ the partition of the circle $\mathbb{S}^1$ into half-open cyclic intervals determined by the finite set $\mathcal{Q}$. Also, given a partition $\mathscr{P}$, the corresponding histogram classifier is denoted $h_{\mathscr{P}}$.

Finally, $P_{\sigma[A_i]}$ is the (conditional) empirical probability supported on the subsample $\sigma[A_i]$, in particular,

$$P_{\sigma[A_i]}[Y = 1 | X \in I] = \frac{\sharp\{j \in A_i \colon x_j \in I,\ y_j = 1\}}{\sharp\{j \in A_i \colon x_j \in I\}}.$$

Here is the algorithm description.

on input σ_n **do**
  $k \leftarrow \max\{i\colon n_i \le n\}$
  $\mathcal{Q} \leftarrow \emptyset$
  $\mathcal{R} \leftarrow \emptyset$
  for $i = 1 : k$ **do**
    if every interval $I \in \hat{\mathcal{P}}_i$ contains $\ge a_i$ points of $\sigma[B_i]$ **and**
      ($i = 1$ **or** every interval $I \in \hat{\mathcal{Q}}$ contains $\ge N_i$ points of $\sigma[A_i]$) **do**
        if $k > 1$ **do**
          for every $I \in \hat{\mathcal{Q}}$ **do**
            if $P_{\sigma[A_i]}[Y = 1 | X \in I] \in (\epsilon_i, 1 - \epsilon_i)$, **do**
              $\mathcal{R} \leftarrow \mathcal{R} \cup (\mathcal{P}_i \cap I)$
              end do
            end if
            end do
          end for
        end if
        $\mathcal{Q} \leftarrow \mathcal{R}$
        $H \leftarrow h_{\hat{\mathcal{Q}}}(\sigma[B_i])$
      end do
    end if
  end for
end do
return H

7. Monotonicity of the expected error

The hypothesis can only be updated at the moments $n = n_k$, so it is enough to compare the expected error of $g_{n_{k-1}}$ and g_{n_k}. Denote i the largest integer $< k$ such that the hypothesis was updated at the step n_i. Denote $\mathcal{Q}_i$ the state of the partitioning set $\mathcal{Q}$ at the moment $n = n_i$. This is a random finite subset of the circle with $\le i$ elements. As before, we denote $\hat{\mathcal{Q}}_i$ the family of half-open intervals into which the circle is partitioned by the finite set $\mathcal{Q}_i$. We will be conditioning on k, i, and $\mathcal{Q}_i$, so from now on, the integers i, k and a finite subset $\mathcal{Q}_i \subseteq \mathbb{S}^1$ (possibly empty) are fixed, while $\mathcal{Q}_k \supseteq \mathcal{Q}_i$ stays random, and we do not know whether a hypothesis update was made at the time n_k. We will further condition on the event (A) "every interval of $\hat{\mathcal{Q}}_k$ contains at least N_i points of the testing sample $\sigma[A_k]$", because given the complementary event, no testing and update were made and $h_{n_k} = h_{n_i} = k_{n_{k-1}}$.

It is now enough to verify, conditionally on the above, that for every interval $I \in \hat{\mathcal{Q}}_i$,

$$P[g_k(X) \ne Y \mid X \in I] \le P[g_i(X) \ne Y \mid X \in I]. \tag{5}$$

Fix such an interval I. Conditioning further on the size of the samples $\sigma[B_i] \restriction I$, $\sigma[A_k] \restriction I$, and $\sigma[B_k] \restriction I$, we see they are conditionally i.i.d., and conditionally jointly independent. The sample $\sigma[A_k] \restriction I$ is conditionally independent on the random partition $\mathcal{P}_k \cap I$. Moreover,

conditionally on the event (A) above, we have $m_k = \sharp\sigma[A_k] \restriction I \geq N(b_{k-1}, \epsilon_k)$, where $b_{k-1} > \sharp\sigma[B_i] \restriction I$. We are under the assumptions of Lemma 11.

Denote $\hat{\mathcal{P}}_k[I]$ the family of all the intervals of the partition $\hat{\mathcal{P}}_k$ contained in I. This is a finite random partition of I (possibly trivial), given by the random set $\mathcal{P}_k \cap I$. For every interval $J \in \hat{\mathcal{P}}_k[I]$, set $m_J = \sharp\sigma[B_k] \restriction I$. According to Lemma 11, conditionally on the event "for all J, $m_J \geq a_k$" the inequality (5) above holds. Since it also holds trivially conditionally on the complementary event (in which case it turns into equality), we are done.

8. Universal consistency

The difficulty here is that the diameter of a random cell (that is, an interval $I = I(X)$ in $\hat{\mathcal{Q}}_k$ containing a random element X) need not converge to zero in probability, and not only because of η. Enough to consider the case where the measure μ is supported on an atom located at 1 and a small arc of length $\epsilon > 0$ around -1. Almost surely, starting with some k, $\hat{\mathcal{Q}}_k$ will contain two intervals of arc length $> 1/2 - \epsilon$ each.

Analysis of the proof of Theorem 6.1 in Devroye et al. (1996) shows that the requirement of the cell diameter going to zero in probability is only needed in order to prove that the sequence of conditional expectations of the regression function η formed with regard to the sequence of random partitions converges to η. This would be, in our case,

$$\mathbb{E}(\eta \mid \hat{\mathcal{Q}}_k) \xrightarrow{p} \eta. \tag{6}$$

We will prove it directly.

Lemma 17 *Let (μ, η) be a learning problem on the circle $\mathbb{S}^1$. Almost surely, starting with some k large enough, at every step n_k every interval of the random partition $\hat{\mathcal{Q}}_k$ will be tested and the hypothesis will be updated.*

Proof By the choice of a_k, the event "every interval of the random partition $\hat{\mathcal{P}}_k$ contains more than $N = N(\epsilon_k, b_{k-1})$ points of $\sigma[A_k]$" occurs with probability $> 1 - \delta_k$, and by the choice of b_k, the event "every interval of the random partition $\hat{\mathcal{P}}_k$ contains more than a_k points of $\sigma[B_k]$" occurs with probability $> 1 - \delta_k$ as well. Since (δ_k) is a summable sequence, we conclude. ■

Lemma 18 *Let (μ, η) be a learning problem on the circle $\mathbb{S}^1$. Let $\epsilon > 0$. Almost surely, starting with some k large enough, for every interval I of the random partition $\hat{\mathcal{Q}}_k$ having the property $p = P[Y = 1 \mid X \in I] \in (\epsilon, 1 - \epsilon)$ we will have $\mathcal{P}_{k+1} \cap I \subseteq \mathcal{Q}_{k+1}$.*

Proof From Lemma 17, we know that almost surely, for all k large enough, the cells of the partition $\mathcal{P}_k$ will be tested. For k' sufficiently large, $\epsilon_{k'} < \epsilon/2$. According to the Chernoff bound,

$$\begin{aligned} P[\text{binomial}(a_{k'}, p) \notin [\epsilon_{k'}, 1 - \epsilon_{k'}]] &\leq P[|\text{binomial}(a_{k'}, p) - p| > \epsilon/2] \\ &< e^{-\epsilon^2 a_{k'}/4}. \end{aligned}$$

The series is summable, and by the Borel–Cantelli lemma, we conclude that the divisibility of I will be certified almost surely from some step on. Consequently, our algorithm prescribes to add the set $\mathcal{P}_{k+1} \cap I$ to the partition $\mathcal{Q}_k$. ■

Lemma 19 *Let (μ, η) be a learning problem on the circle $\mathbb{S}^1$. Let I be a half-open cyclic interval on which η is neither a.e. equal to 1 nor a.e. equal to 0. Almost surely, at some step k we will have $\mathcal{Q}_k \cap I \neq \emptyset$.*

Proof We have $p = P[Y = 1 \mid X \in I] \in (0, 1)$. Every interval J containing I satisfies $P[Y = 1 \mid X \in J] \in (p\mu(I), 1 - p\mu(I))$. Almost surely, if k is large enough, $\mathcal{P}_{k+1} \cap I \neq \emptyset$ (Lemma 16), and either $\mathcal{Q}_k \cap I \neq \emptyset$, or else the interval J_k of the partition $\hat{\mathcal{Q}}_k$ containing I will be tested at the step $k+1$ and the set $\mathcal{P}_{k+1} \cap J_k$ added to the partitioning set (Lemma 18). Thus, almost surely, $\mathcal{Q}_{k+1} \cap I \neq \emptyset$. ■

Denote $\Sigma(\cup_k \hat{\mathcal{Q}}_k)$ the sigma-algebra generated by all the cyclic intervals determined by random partitions $(\mathcal{Q}_k)$, $k \in \mathbb{N}$. Turns out, this random sigma-algebra has a rather transparent structure. We will clarify it now, as well as show that $\Sigma(\cup_k \hat{\mathcal{Q}}_k)$ is a bona fide random variable taking values in a standard Borel space.

Given a subset $A \subseteq \mathbb{S}^1$, denote Σ_A the sigma-algebra on the circle generated by all cyclic intervals $[a, b)$, $a, b \in A$. It is a sub-sigma-algebra of the Borel algebra.

Lemma 20 *A subset $A \subseteq \mathbb{S}^1$ and its closure, $\bar{A}$, generate the same sigma-algebra.*

Proof The inclusion $\Sigma_A \subseteq \Sigma_{\bar{A}}$ is trivial. Now suppose $a \in A$ and $b \in \bar{A}$. If there is a sequence of elements of A with $b_n \uparrow b$ (that is, $[a, b_n, b]$), then $[a, b) = \cup_n [a, b_n)$. If there is a sequence $b_n \downarrow b$ ($[a, b, b_n]$), then $\{b\} = \cap_n [b, b_n)$, and $[a, b) = \cap_n [a, b_n] \setminus \{b\}$. Assume now $a, b \in A$ arbitrary. If there is a sequence of elements of A, $a_n \uparrow a$, then $[a, b) = \cap_n [a_n, b)$; if there is a sequence $a_n \downarrow a$, then $\{a\} = \cap_n [a, a_n)$ and so on. ■

Lemma 21 *On every closed subset F of $\mathbb{S}^1$ the sigma-algebra Σ_F induces the standard Borel structure (as induced from $\mathbb{S}^1$).*

Proof Enough to show that for every $a, b \in F$, $a \neq b$, we have $(a, b) \cap F \in \Sigma_F|_F$. If $(a, b) \cap F = \emptyset$, it is clear; assume the contrary. There is a sequence (a_n) of elements of F with $a_n \in (a, b)$ and $a_n \downarrow \inf_{<_a}(a, b) \cap F$. We have $\{a\} = \cap_n [a, a_n) \cap F \in \Sigma_F|_F$, and finally $(a, b) \cap F = [a, b) \cap F \setminus \{a\} \in \Sigma_F|_F$. ■

It is well-known and easily proved that every open subset U of the real line (hence, of the circle) is uniquely represented as a union of disjoint open intervals (its connected components) whose endpoints belong to the complement of U, see e.g. Alexandroff (1984), §5, Th. 21, or Engelking (1989), Exercise 3.12.4(b).

Lemma 22 *Let F be a closed subset of the circle. Suppose the sigma-algebra Σ_F is non-trivial (equivalently, F contains at least two points). Those atoms of Σ_F that are not singletons are exactly the half-open intervals $[a,b)$ such that (a,b) is a connected component of the complementary set $F^c = \mathbb{S}^1 \setminus F$.*

Proof Let $a, b \in F$, $a \neq b$. We have $[a,b) \in \Sigma_F$. Assume that $(a,b) \subseteq F^c$. The restriction $\Sigma_F|_{[a,b)}$ is generated, as a sigma algebra, by the intersections of the generating sets $[c,d)$, $c,d \in F$, with $[a,b)$. Since every such set either contains $[a,b)$ or is disjoint from it, the sigma-algebra $\Sigma_F|_{[a,b)}$ is trivial. Altogether it means $[a,b)$ is an atom of Σ_F.

Let now $A \in \Sigma_F$ be an atom. Suppose it contains at least two points. For any two $a, b \in F$, $a \neq b$, exactly one of the intervals $[a,b)$ and $[b,a)$ contains A as a subset. Denote I the intersection of all the intervals $[a,b)$, $a,b \in F$ that contain A. Since F is closed, the endpoints c,d of the interval I belong to F. As A is an atom, it must satisfy $A \subseteq [c,d)$, and (c,d) contains no points of F. Since $[c,d)$ is an atom by the first part of the proof, $A = [c,d)$. ■

The map $F \mapsto \Sigma_F$ is not injective even on the closed subsets: for instance, all one-element subsets generate the same trivial sigma-algebra $\{\mathbb{S}^1\}$.

Lemma 23 *If $F \neq G$ are two distinct closed subsets and at least one of them contains two elements, then $\Sigma_F \neq \Sigma_G$.*

Proof Suppose $F \supseteq \{a,b\}$, $a \neq b$. If $F \setminus G \neq \emptyset$, then for any $c \in F \setminus G$ we have $[a,c) \in \Sigma_F \setminus \Sigma_G$. So we can assume $F \subseteq G$. In this case, for any $d \in G \setminus F$, $[a,d) \in \Sigma_G \setminus \Sigma_F$. ■

We can therefore bijectively identify the family of all sigma-algebras of the form Σ_F with the family of all closed subsets of the circle with at least two elements, plus the trivial sigma-algebra $\{\mathbb{S}^1\}$.

The family $\mathcal{F}(K)$ of closed subsets of a compact metric space is itself a compact metric space and therefore a standard Borel space, for example, when equipped with the Hausdorff distance (Kechris (1995), 4.F.):

$$d(F,G) = \inf\{\epsilon > 0 \colon d(x,F) < \epsilon, d(y,G) < \epsilon \text{ for all } x \in G, y \in F\}.$$

The subfamily of sets with at least two elements is open, hence Borel. The union of two standard Borel spaces is a standard Borel space. This gives a standard Borel structure to the family of all sigma-algebras of the form Σ_F, F is a closed subset of $\mathbb{S}^1$.

The sigma-algebras $\Sigma(\cup_k \hat{\mathcal{Q}}_k)$ that we are interested in are exactly of the form $\Sigma_{\mathcal{Q}_\infty}$, where we denote $\mathcal{Q}_\infty = \cup_k \mathcal{Q}_k$ the set of all partitioning points added by our algorithm. This inclusion $\Sigma(\cup_k \hat{\mathcal{Q}}_k) \subseteq \Sigma_{\mathcal{Q}_\infty}$ is clear, and if $a, b \in \mathcal{Q}_\infty$, then for some k, $a, b \in \mathcal{Q}_k$, and $[a,b)$ is in the sigma-algebra determined by the partition $\hat{\mathcal{Q}}_k$.

Finally, the random variable with values in the above standard Borel space that we call a random sigma-algebra is realized through a map sending a sample path in $(\mathbb{S}^1 \times \{0,1\})^\infty$ to the sigma-algebra $\mathcal{Q}_\infty$. This map is Borel measurable with regard to the above Borel

structure. Indeed, it is a combination of the sequence of maps $(\mathbb{S}^1 \times \{0,1\})^{[n_k+1,n_{k+1}]}$ to $(\mathbb{S}^1)^k$, produced by the learning rule, each of which can be expressed by a finite first-order formula with relation symbols $[\ ,\ ]$ and $<$ and the real numbers as constants, and so is measurable, and the map sending a sequence (x_k) to the closure of the set $\{x_k\}$. The measurability of the latter map can be seen as follows: the inverse image of the Hausdorff ϵ-neighbourhood of a closed set F consists of all sequences satisfying the formula

$$\exists n \forall k, B_{1/n}(x_k) \subseteq F_\epsilon,$$

making it a Borel set.

Here is a corollary of Lemma 19.

Lemma 24 *Either almost surely the sigma-algebra $\Sigma_{\mathcal{Q}_\infty}$ is trivial (and this is the case if and only if the regression function η is constant a.e., taking value 0 or 1), or almost surely it is non-trivial.*

Lemma 25 *Almost surely,*

1. *on the random closed set $F = \bar{\mathcal{Q}}_\infty$, the random sigma-algebra $\Sigma_{\mathcal{Q}_\infty}$ induces the standard Borel structure $\mathcal{B}_F$ coming from $\mathbb{S}^1$, and*
2. *the regression function η assumes a.e. a constant value 0 or 1 on every atom of $\Sigma_{\mathcal{Q}_\infty}$ that is not a singleton.*

Proof The first claim follows from Lemmas 20 and 21.

For the second claim, according to Lemma 24, it is enough to consider the case where $\Sigma_{\mathcal{Q}_\infty}$ is almost surely non-trivial. It follows from Lemma 19 that almost surely, every interval with rational endpoints on which η does not take a.e. identical value 0 or 1 will be divided at the k-th step for some k large enough. We conclude that, almost surely, on every interval with rational endpoints contained in some atom of $\Sigma_{\mathcal{Q}_\infty}$ the regression function takes a.e. the identical value 0 or the value 1. It follows that almost surely, for every atom A that is non-singleton and so has the form $A = [a, b)$ for $a, b \in \bar{\mathcal{Q}}_\infty$, on the corresponding open interval (a, b) η takes identical value 0 or 1 a.e. For those atoms with $\mu\{a\} = 0$, the proof is over.

Now denote $\mathcal{U}$ the family of all half-open intervals of the form $[a, b)$, where $\mu\{a\} > 0$ and b is rational. The family U is countable, so again applying Lemma 19, we conclude that almost surely, if any such interval is an atom, then η must take the same value at the left endpoint a as a.e. on the rest of the interval (this includes also the case $\mu(a, b) = 0$). ■

Lemma 26 *Almost surely, $\mathbb{E}(\eta \mid \Sigma_{\mathcal{Q}_\infty}) = \eta$.*

Proof Select a Borel measurable version of η. Further, on every nontrivial atom of $\Sigma_{\mathcal{Q}_\infty}$ replace η with a suitable constant value, either identically 0 or identically 1 (Lemma 25,(2)). The union, A, of the countable family of nontrivial atoms belongs to our sigma-algebra, and the restriction of η to A is $\Sigma_{\mathcal{Q}_\infty}$-measurable. We have $A^c \subseteq \bar{\mathcal{Q}}_\infty$, therefore, almost surely

the restriction of $\Sigma_{\mathcal{Q}_\infty}$ induces the standard Borel structure on A^c (Lemma 25,(1)) and the restriction of η to A^c is $\Sigma_{\mathcal{Q}_\infty}$-measurable as well. We conclude: our realization of η is $\Sigma_{\mathcal{Q}_\infty}$-measurable. ■

Lemma 27 *Almost surely,* $\mathbb{E}(\eta \mid \hat{\mathcal{Q}}_k) \to \eta$.

Proof Follows from the forward martingale convergence theorem (Doob (1994), Sect. IX.14) and Lemma 26. ■

And finally, the proof of the universal consistency of our learning rule, g.

Denote h the following variant of g: it is a partitioning rule based on the same sequence of random partitions $\hat{\mathcal{Q}}_k$ and labelling samples $\sigma[B_k]$, but updated at every moment n_k, irrespective of the number of sample points in the cells of the partition:

$$h_{n_k}(\sigma) = h_{\hat{\mathcal{Q}}_k}(\sigma[B_k]).$$

Lemma 27 implies the almost sure convergence of the conditonal expectations $\mathbb{E}(\eta \mid \hat{\mathcal{Q}}_k)$ to η. Because of Lemma 17 and the fact that $a_k \to \infty$, almost surely the smallest number of points of the i.i.d. sample $\sigma[B_k]$ contained in any cell of the random partition $\hat{\mathcal{Q}}_k$ at the step n_k will go to infinity as $k \to \infty$. We are under the assumptions of Theorem 12, and conclude that the rule h is consistent.

The only difference between g and h is that g sometimes delays the hypothesis update. More exactly, we have a certain random function, F, from the natural numbers to itself with the property $F(k) \leq k$ for all k, and the learning rule g is defined from h as follows:

$$g_{n_k}(\sigma) = h_{n_{F(k)}}(\sigma \restriction [n_{F(k)}]).$$

Notice that $F(k) - k \to 0$ almost surely (Lemma 17). We are under the assumptions of Lemma 1 and conclude that the rule g is consistent.

9. Concluding remarks

I am grateful to the two anonymous referees whose comments have helped to improve the readability of the paper.

In connection with the discussion at the start of Sect. 8, it was pointed out by one referee that there are indeed examples of consistent partition-based algorithms without the diameter of the largest cell converging to zero in probability (Scornet et al. (2015)).

A Borel isomorphism between an Euclidean domain and the circle (Sect. 5) is indeed not easy to implement algorithmically. However, already a Borel injection would suffice, and this can be coded in a constructive way, cf. Pestov (2013), Sect. 7. Still, the learning rule described in the present article will be too slow for practical applications: its algorithmic efficiency is admittedly very low. It remains an interesting challenge, to find a "natural" learning algorithm having the monotone expected learning error.

References

P. S. Alexandroff. *Einführung in die Mengenlehre und in die allgemeine Topologie*, volume 85 of *Hochschulbücher für Mathematik [University Books for Mathematics]*. VEB Deutscher Verlag der Wissenschaften, Berlin, 1984. Translated from the Russian by Manfred Peschel, Wolfgang Richter and Horst Antelmann.

Thomas M. Cover and Peter E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

J. L. Doob. *Measure theory*, volume 143 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1994.

Ryszard Engelking. *General topology*, volume 6 of *Sigma Series in Pure Mathematics*. Heldermann Verlag, Berlin, second edition, 1989.

Alexander S. Kechris. *Classical descriptive set theory*, volume 156 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1995.

Vladimir Pestov. Is the k-NN classifier in high dimensions affected by the curse of dimensionality? *Comput. Math. Appl.*, 65(10):1427–1437, 2013.

Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests. *Ann. Statist.*, 43(4):1716–1741, 2015.

S. Świerczkowski. On cyclically ordered groups. *Fund. Math.*, 47:161–166, 1959.

M. Vidyasagar. *Learning and generalization. With applications to neural networks*. Communications and Control Engineering Series. Springer-Verlag London, Ltd., London, second edition, 2003.