Proceedings of Machine Learning Research vol 178:1–49, 2022 35th Annual Conference on Learning Theory

Universal Online Learning: an Optimistically Universal Learning Rule

Moïse Blanchard MOISEB@MIT.EDU
Massachusetts Institute of Technology

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We study the subject of universal online learning with non-i.i.d. processes for bounded losses. The notion of *universally consistent learning* was defined by Hanneke [18] in an effort to study learning theory under minimal assumptions, where the objective is to obtain low long-run average loss for any target function. We are interested in characterizing processes for which learning is possible and whether there exist learning rules guaranteed to be universally consistent given the *only* assumption that such learning is *possible*. The case of unbounded losses is very restrictive since the learnable processes almost surely have to visit a finite number of points and as a result, simple memorization is optimistically universal [18; 4]. We focus on the bounded setting and give a complete characterization of the processes admitting strong and weak universal learning. We further show that the k-nearest neighbor algorithm (kNN) is not optimistically universal and present a novel variant of 1NN which is optimistically universal for general input and value spaces in both strong and weak settings. This closes all the COLT 2021 open problems posed in [19] on universal online learning.

Keywords: online learning, universal consistency, stochastic processes, measurable partitions, statistical learning theory, Borel measure

1. Introduction

We consider the fundamental question of learnability and generalizability for online learning. In this framework, a learner is sequentially given input points $\mathbb{X} := (X_t)_{t\geq 0}$ from a general separable metric *instance space* $(\mathcal{X}, \rho)$ and observes the corresponding values $\mathbb{Y} := (Y_t)_{t\geq 0}$ from a separable near-metric *value space* $(\mathcal{Y}, \ell)$. The learner's goal is to predict the values before their observation. The input points are given according to some stochastic process $\mathbb{X}$ on $\mathcal{X}$ and we assume that the process $\mathbb{Y}$ is generated from $\mathbb{X}$ in a *noiseless* fashion, i.e., that there exists an unknown measurable function $f^* : \mathcal{X} \to \mathcal{Y}$ such that $Y_t = f^*(X_t)$ for all $t \geq 0$. At time step t, the learner outputs a prediction $\hat{Y}_t$ based solely on the historical data $(X_u, Y_u)_{u<t}$ and the new input point X_t. We wish to obtain low long-run average errors $\frac{1}{t}\sum_{u\leq t} \ell(Y_u, \hat{Y}_u)$. Specifically, we consider two types of consistency: strong consistency is achieved when the average error converges to 0 almost surely; and weak consistency is achieved when the expected average error converges to 0. We are interested in *universal* online learning, in which we ask for consistency for any unknown measurable target function f^*. In this framework, the two main questions are 1, to characterize the input processes $\mathbb{X}$ for which universal consistency is achievable, and 2, if possible, provide a learning rule which would guarantee universal consistency whenever such objective is achievable.

Motivation and related work. This work builds upon the stream of papers on universal online learning [18; 4; 2], which aims to study the question of *learnability* under minimal assumptions. A classical objective in statistical learning is to provide learning rules with guarantees for some large class of problem instances. In general, it is not possible to be consistent under all stochastic processes $\mathbb{X}$ and target functions f^*. Therefore, it is necessary to impose instance constraints. In the literature, there is a rich variety on the types of proposed restrictions. The first category of works does not restrict the input sequences $\mathbb{X}$ but instead the target functions f^* [24; 7; 1; 27]. A large portion of the literature belongs to a second category which restricts both input processes and target functions. For instance, if we assume that the input process is independent identically distributed (i.i.d.) and that the target function belongs to a class of finite VC dimension, then there exists an algorithm guaranteeing $O(\log t)$ mistakes in expectation [21]. Other more involved restrictions on $\mathbb{X}$ and f^* have been considered [23; 28; 32; 5]. The subject of this paper is of a third category, in which we impose no assumptions on the set of target functions f^*, but instead, restrict the input sequences $\mathbb{X}$. Specifically, we focus on *universally consistent* algorithms that achieve consistency for all target functions.

Most of the literature on universal learning considers standard *ad-hoc* probabilistic assumptions on the input stochastic process, for instance assuming that the training samples are i.i.d. A classic result in this i.i.d. setting shows that in the Euclidean space, the 1-nearest neighbor rule is universally consistent [9; 30; 11]. Also in the Euclidean case, the k_t-nearest neighbor rule with $k_t/\log t \to \infty$ and $k_t/t \to 0$ is also strongly consistent under mild assumptions in the noisy setting where $(\mathbb{X}, \mathbb{Y})$ is any i.i.d. process [10]. More recently, [20; 17; 31] proposed algorithms which achieve minimal risk for i.i.d. process $(\mathbb{X}, \mathbb{Y})$ in any metric spaces where this is possible—namely *essentially separable* metric spaces which generalize separable metric spaces. This setting is referred to as universal Bayes consistency. Other similar assumptions on the input process $\mathbb{X}$ include stationary ergodic [25; 14; 13] or satisfying the law of large numbers [26; 12; 29]. Instead, we are interested in provably-minimal assumptions rooted in the learning problem itself. Specifically, we follow the so-called optimist's decision theory introduced by Hanneke [18] and frequent in universal learning [31]: to achieve a given objective, the optimist's sole assumption is that this objective is at least achievable by some learning rule. In some sense, this assumption is minimal as it is necessary for any algorithm to have any positive guarantees. In this framework, we are particularly interested in algorithms that would reach the objective without further assumptions. These are named *optimistically universal* learning rules. Such algorithms enjoy the convenient property that if they fail for a particular problem instance, any other learning rule would fail as well. In our case, we are interested in the set of learnable processes $\mathbb{X}$, i.e., for which universal consistency is possible, and aim to provide optimistically universal algorithms if they exist, i.e., learning rules which are universally consistent on all processes $\mathbb{X}$ for which universal consistency is achievable.

In the case of *unbounded* losses ℓ, these questions are settled [18; 4]. Precisely, the learnable processes are exactly the sequences visiting a finite number of input points almost surely, and as a result, simple memorization is optimistically universal. Hence, universal learning with unbounded losses is very restrictive. In this paper, we focus on the bounded loss case for which it is known that i.i.d. and convergent relative frequencies processes are learnable [18]. Recently, [2] provided a reduction from any general bounded output setting $(\mathcal{Y}, \ell)$ to binary classification.

Contributions. We propose a class of learning rule kC1NN for $k \geq 2$, which we prove are strongly and weakly optimistically universal for general separable metric instance spaces $(\mathcal{X}, \rho)$

and separable near-metric value spaces $(\mathcal{Y}, \ell)$ with bounded loss. These learning rules are simple variants of the classical 1-nearest neighbor (1NN). They essentially perform 1NN on a restricted dataset by deleting any input point from the historical dataset whenever it has been used as nearest neighbor at least k times. We also show that any $(k_t)_t-$nearest neighbor fails to be optimistically universal under very mild conditions on the sequence $(k_t)_t$ even for very simple input spaces $\mathcal{X}$ e.g. Euclidean spaces. Finally, we give a complete characterization of processes admitting strong and weak universal learning. This closes all main questions on universal online learning, which are stated as open problems in [19].

Outline of the paper. The rest of this paper is organized as follows. In Section 2, we formally introduce universal learning and present the two main questions of this topic. The main results are then stated in Section 3. In Section 4 we focus on nearest neighbor learning rules and show that they are not universally consistent, by constructing learnable processes for which nearest neighbor methods fail. This example gives motivation for the 2C1NN learning rule, constructed in Section 5, and then show that it is optimistically universal. We further provide a complete characterization of the set of learnable processes. We then turn to weak universal learning in Section 6. Finally, we give open research directions in Section 7.

2. Formal setup and preliminaries

Instance and value space. In this paper, we follow the general framework of online learning where one observes an input sequence $\mathbb{X} = (X_t)_{t\geq 1}$ of points in a separable metric *instance space* $(\mathcal{X}, \rho)$, together with their corresponding target values $\mathbb{Y} = (Y_t)_{t\geq 1}$ coming from a separable near-metric *value space* $(\mathcal{Y}, \ell)$. The loss $\ell : \mathcal{Y}^2 \to [0, \infty)$ is said to be a near metric if it is symmetric $\ell(y_1, y_2) = \ell(y_2, y_1)$, satisfies $\ell(y_1, y_2) = 0$ if and only if $y_1 = y_2$, and also satisfies a relaxed triangle inequality $\forall y_1, y_2, y_3 \in \mathcal{Y}^3 : \ell(y_1, y_3) \leq c_\ell(\ell(y_2, y_1) + \ell(y_2, y_3))$, where c_ℓ is a fixed constant. Note that all metrics are near-metrics with $c_\ell = 1$. As an important example for regression, the squared loss is near-metric with $c_\ell = 2$. We denote by $\bar{\ell} := \sup_{y_1, y_2 \in \mathcal{Y}} \ell(y_1, y_2)$ the loss function supremum and will be particularly interested in *bounded* losses, i.e., $\bar{\ell} < \infty$.

Input and output processes. In an effort to study non-i.i.d. processes, the input sequence of points is a general stochastic process on the Borel space $(\mathcal{X}, \mathcal{B})$ induced by a metric ρ. This is a major difference with a majority of the relevant statistical learning literature imposing ad-hoc hypothesis on $\mathbb{X}$ as discussed in Section 1. We consider a noiseless setting in which the output values $\mathbb{Y}$ are generated from $\mathbb{X}$ through an unknown measurable function $f^* : \mathcal{X} \to \mathcal{Y}$ such that $Y_t = f^*(X_t)$ for all $t \geq 1$.

Online learning and consistency. In *online* learning, the learning process is sequential: at time $t \geq 1$, one observes a new input data-point X_t and outputs a prediction $\hat{Y}_t$ based solely on the historical data $(\mathbb{X}_{\leq t-1}, \mathbb{Y}_{\leq t-1})$ and the new covariate X_t. We measure the performance of the learning rule through the loss function ℓ. *Strong* consistency is achieved when the algorithm obtains asymptotic average loss 0 almost surely. Alternatively, a learning rule is *weakly* consistent when it guarantees 0 asymptotic average loss in expectation. We now formally write these notions. A learning rule is a sequence $f_\cdot = \{f_t\}_{t=1}^{\infty}$ of measurable functions with $f_1 : \mathcal{X} \to \mathcal{Y}$ and $f_t : \mathcal{X}^{t-1} \times \mathcal{Y}^{t-1} \times \mathcal{X} \to \mathcal{Y}$ for $t \geq 2$. Given a history $(X_i, Y_i)_{i<t}$ and a new input point X_t, the rule $f_\cdot$ makes the prediction $f_t(\mathbb{X}_{<t}, \mathbb{Y}_{<t}, X_t)$ for Y_t and $t \geq 2$. For simplicity, for $t = 1$ we may also use

the notation $f_1(\mathbb{X}_{<1}, \mathbb{Y}_{<1}, X_1)$ instead of $f_1(X_t)$. We write the average loss at time T as

$$\mathcal{L}_{\mathbb{X}}(f_\cdot, f^*; T) := \frac{1}{T}\sum_{t=1}^{T} \ell(f_t(\mathbb{X}_{<t}, \mathbb{Y}_{<t}, X_t), f^*(X_t)).$$

We aim to minimize the long-run average loss. The online learning rule $f_\cdot$ is strongly consistent under the input process $\mathbb{X}$ and for the target function f^* when $\mathcal{L}_{\mathbb{X}}(f_\cdot, f^*; T) \to 0 \ \ (a.s.)$. For simplicity, we define $\mathcal{L}_{\mathbb{X}}(f_\cdot, f^*) = \limsup_{T\to\infty} \mathcal{L}_{\mathbb{X}}(f_\cdot, f^*; T)$. Therefore, the above condition can be rewritten as $\mathcal{L}_{\mathbb{X}}(f_\cdot, f^*) = 0 \ \ (a.s.)$. We also consider weak learning: similarly, $f_\cdot$ is weakly consistent under $\mathbb{X}$ and for f^* when $\mathbb{E}\mathcal{L}_{\mathbb{X}}(f_\cdot, f^*; T) \to 0$.

Universal consistency and optimistically universal learning rule. Following [18], we are interested in learning rules which achieve strong (resp. weak) consistency under a specific input sequence $\mathbb{X}$ for all measurable target functions $f^* : \mathcal{X} \to \mathcal{Y}$. Such learning rules are said to be strongly (resp. weakly) *universally consistent* under $\mathbb{X}$. We define SUOL as the set of all stochastic processes $\mathbb{X}$ for which strong universal online learning is achievable by some learning rule. Similarly, we denote by WUOL the set of all processes $\mathbb{X}$ that admit weak universal online learning. These sets may depend on the setup $(\mathcal{X}, \rho), (\mathcal{Y}, \ell)$ so we will specify $\text{SUOL}_{(\mathcal{X},\rho),(\mathcal{Y},\ell)}$ and $\text{WUOL}_{(\mathcal{X},\rho),(\mathcal{Y},\ell)}$ when the spaces are not clear from the context. In this framework, two main areas of research are (1) characterizing the sets SUOL (resp. WUOL) for a given setup in terms of the properties of the stochastic process $\mathbb{X}$, and (2) identifying learning rules which are strongly (resp. weakly) universally consistent for any input process $\mathbb{X}$ in SUOL (resp. WUOL), i.e. that achieve strong (resp. weak) universal consistency whenever it is possible. These are called *optimistically universal* learning rules. In the case of *unbounded* loss functions i.e. $\bar{\ell} = \infty$, both questions are answered for any choice of $(\mathcal{X}, \rho), (\mathcal{Y}, \ell)$ for strong universal consistency [18; 4]. Specifically, [4] shows the stochastic processes $\mathbb{X}$ which admit strong universal online learning are exactly those which visit a *finite* number of distinct input points of $\mathcal{X}$ almost surely. As a consequence, the simple memorization learning rule is optimistically universal. Further, for unbounded losses, strong and weak universal learning are equivalent [18]. These results are rather negative in the sense that unbounded loss results in a very restricted set SUOL.

Bounded loss. The present paper will therefore focus on the *bounded* loss case i.e. $\bar{\ell} < \infty$, for which both questions are open. This is the main case of interest for universal online learning. Contrary to the unbounded case, for bounded losses, the set of learnable processes SUOL contains, in particular, all i.i.d. processes [18]. In fact, the simple 1-nearest neighbor (1NN) learning rule achieves strong (and weak) universal consistency for all i.i.d. processes $\mathbb{X}$ in for the Euclidean space $\mathbb{X} = \mathbb{R}^d$ [10]. It is even known that the $(k_t)_t$-neighbor algorithm ($(k_t)_t$NN) with $k_t/\log t \to \infty$ and $k_t/t \to 0$ achieves Bayes minimal risk in the noisy setting for large classes of input spaces $\mathcal{X}$ [8]. This implies in particular that kNN achieves strong universal consistency in our noiseless setting for these input spaces. However, it is an open question whether there exist simple input spaces $\mathcal{X}$—e.g. Euclidean spaces—for which some kNN algorithms would be optimistically universal. In other terms, does there exist an input process $\mathbb{X}$ such that 1NN fails to achieve consistency for some target function f^* but universal consistency would still be achieved by some other—more sophisticated—learning rule? No characterization of SUOL is known either, although [18] proposed a necessary condition for belonging to SUOL and conjectured that it is also sufficient. We refer to this condition as SMV (sub-linear measurable visits). Intuitively, it asks that for any measurable partition of the

input space $\mathcal{X}$, the process $\mathbb{X}$ only visits a sublinear number of its regions. Note that this condition does not depend on the choice of output setup $(\mathcal{Y}, \ell)$.

Condition SMV *Define the set* $\text{SMV}_{(\mathcal{X},\rho)}$ *as the set of all processes* $\mathbb{X}$ *satisfying the condition that, for every disjoint sequence* $\{A_k\}_{k=1}^{\infty}$ *in* $\mathcal{B}$ *with* $\cup_{k=1}^{\infty} A_k = \mathcal{X}$ *(i.e., every countable measurable partition),* $|\{k \in \mathbb{N} : A_k \cap \mathbb{X}_{<T} \neq \emptyset\}| = o(T) \quad (a.s)$.

For the weak setting we can define a similar condition WSMV (weak sub-linear measurable visits).

Condition WSMV *Define the set* $\text{WSMV}_{(\mathcal{X},\rho)}$ *as the set of all processes* $\mathbb{X}$ *satisfying the condition that, for every countable measurable partition* $\{A_k\}_{k=1}^{\infty}$, $\mathbb{E}[|\{k \in \mathbb{N} : A_k \cap \mathbb{X}_{<T} \neq \emptyset\}|] = o(T)$.

Hanneke [18] showed that these conditions are necessary for strong and weak universal learning.

Proposition 1 (Hanneke [18]) *For any separable Borel space* $\mathcal{X}$ *and separable near-metric output setting* $(\mathcal{Y}, \ell)$ *with* $0 < \bar{\ell} < \infty$ *we have* $\textit{SUOL}_{(\mathcal{X},\rho),(\mathcal{Y},\ell)} \subset \textit{SMV}_{(\mathcal{X},\rho)}$ *and* $\textit{WUOL}_{(\mathcal{X},\rho),(\mathcal{Y},\ell)} \subset \textit{WSMV}_{(\mathcal{X},\rho)}$.

However, it is an open question whether SMV (resp. WSMV) is also a sufficient condition for strong (resp. weak) universal learning. Together with the question of the existence of an optimistically universal learning rule, these are the main objectives for universal online learning. These questions are posed in the COLT 2021 open problems [19], which we now formally restate.

Hanneke's $5000 open problem 1 [19] *Does there exist an optimistically universal online learning algorithm? (in either the weak or strong sense)*

Hanneke's $1000 open problem 2 [19] *Is SMV (resp. WSMV) equal to the set of all* $\mathbb{X}$ *such that strong (resp. weak) universal online learning is possible under* $\mathbb{X}$*?*

It is important to note that these questions are easily solved in the case where $\mathcal{X}$ is countable [18]. Therefore, the main interest is to answer these questions for *any* uncountable $\mathcal{X}$. In fact, Hanneke [19] even announced a $5000 (resp. $1000) reward for solving open problem 1 (resp. 2) for the Euclidean $\mathcal{X} = \mathbb{R}^d$ case. Both questions will be solved in Appendix B for $\mathcal{X} = [0, 1]$ specifically. This is a rather general case because its extension to all standard Borel spaces $\mathcal{X}$ is immediate through an equivalence result from Kuratowski of all uncountable standard Borel spaces. For instance, this solves the question for all Euclidean spaces $\mathbb{R}^d$ for $d \geq 1$. Most importantly, the special case $\mathcal{X} = [0, 1]$ allows for a simplified exposition and provides all useful intuitions. The complete result holds for all separable Borel spaces and is presented in Section 5.

Notations. For any sequence $\boldsymbol{x}$, we will use the following notations when analyzing finite time horizons: $\boldsymbol{x}_{\leq t} := \{x_1, ..., x_t\}$ and $\boldsymbol{x}_{<t} := \{x_1, ..., x_{t-1}\}$ for simplicity. For a metric space $(\mathcal{X}, \rho)$, a point $x \in \mathcal{X}$ and $r \geq 0$, we denote by $B_\rho(x, r) := \{x' \in \mathcal{X},\ \rho(x, x') < r\}$ the open ball centered in x of radius r, and $S_\rho(x, r) = \{x' \in \mathcal{X},\ \rho(x, x') = r\}$ the sphere centered in x of radius r. We might omit the metric ρ in subscript if there is no ambiguity. We also denote by ℓ_{01} the indicator loss function, i.e., $\ell_{01}(i, j) = \mathbb{1}(i \neq j)$. Since it is a metric, it is also a near-metric with $c_\ell = 1$. For simplicity, we will use the same notation ℓ_{01} irrespective of the output space $\mathcal{Y}$. For any measurable set A, we denote by $\mathbb{1}_A$ the function $\mathbb{1}_A(\cdot) := \mathbb{1}_{\cdot \in A}$. We will denote by $|\cdot|$ any norm on $\mathbb{R}$. Recall

that all norms are equivalent on finite dimensional spaces, hence the topology induced by these metrics is identical. When the space $(\mathcal{X}, \rho)$ is obvious from the context, we may reduce the notation $\mathrm{SUOL}_{(\mathcal{X},\rho),(\mathcal{Y},\ell)}$ to $\mathrm{SUOL}_{(\mathcal{Y},\ell)}$. We might omit also the loss ℓ when there is no ambiguity.

3. Main results

We first show that the simple nearest neighbor rule (1NN) is not optimistically universal. The proof generalizes to general $(k_t)_t$-nearest neighbor algorithms under very mild assumptions on $(k_t)_t$.

Theorem 2 *The $(k_t)_t-$nearest neighbor learning rule is not strongly optimistically universal for the input space $\mathcal{X} = [0, 1]$ with usual topology and for binary classification, for any sequence $(k_t)_t$ such that $k_t = o\left(\frac{t}{(\log t)^{1+\delta}}\right)$ for any $\delta > 0$.*

This is obtained by constructing a specific process $\mathbb{X} \in \mathrm{SUOL}_{([0,1],|\cdot|),(\{0,1\},\ell_{01})}$ under which nearest neighbor is not universally consistent. Intuitively, 1NN fails on the process because certain "bad" data points are used an arbitrarily large number of times as nearest neighbor for future input points and hence, induce a large number of mistakes for 1NN. To resolve this issue, we propose a new learning rule 2-Capped-1-Nearest-Neighbor (2C1NN), a variant of the classical 1NN, designed to ensure that the number of times each datapoint is used as nearest neighbor is capped by 2. Specifically, once a datapoint X_t has been used as nearest neighbor twice, it is deleted from the training dataset. We show that this is an optimistically universal learning rule for both strong universal learning and weak universal learning.

Theorem 3 *For any separable Borel space $\mathcal{X}$, and any separable near-metric output setting $(\mathcal{Y}, \ell)$ with bounded loss, i.e., $\sup_{y_1,y_2} \ell(y_1, y_2) < \infty$, 2C1NN is a strongly (resp. weakly) optimistically universal learning rule.*

More generally, we can define learning rules kC1NN for any $k \geq 2$. The proof further shows that all kC1NN is optimistically universal for any $k \geq 2$. Further, we give a characterization of the processes admitting strong and weak universal learning.

Theorem 4 *For any separable Borel space $\mathcal{X}$, and any separable near-metric output setting $(\mathcal{Y}, \ell)$ with $0 < \sup_{y_1,y_2} \ell(y_1, y_2) < \infty$, we have*

$$\mathit{SUOL}_{(\mathcal{X},\rho),(\mathcal{Y},\ell)} = \mathit{SMV}_{(\mathcal{X},\rho)} \quad \textit{and} \quad \mathit{WUOL}_{(\mathcal{X},\rho),(\mathcal{Y},\ell)} = \mathit{WSMV}_{(\mathcal{X},\rho)}.$$

If $\sup_{y_1,y_2} \ell(y_1, y_2) = 0$, then the loss is identically null. Therefore, all stochastic processes are strongly and weakly learnable.

It is worth noting that although the sets SUOL and WUOL differ—the set of weakly learnable processes WUOL is larger than the set of strongly learnable processes SUOL—the same learning rule 2C1NN is optimistically universal in both strong and weak settings. Theorem 3 and Theorem 4 close the two open problems of the existence of an optimistically universal learning rule and a characterization of the set of learnable input sequences, formulated in [19].

4. On nearest neighbor consistency

A natural candidate for good learning rules in general spaces are the nearest neighbor algorithms. We recall that the $(k_t)_t-$nearest neighbor ($(k_t)_t$NN) learning rule, at step t, considers the closest k_t neighbors to the new input point and follows the majority vote to make its prediction. Indeed, for instance, for $\mathcal{X} = \mathbb{R}$ and binary classification, under any process $\mathbb{X} \in$ SUOL which admits universal learning, nearest neighbor successfully learns simple functions—representing union of intervals [2]. Further, the special case of binary classification is not restrictive because if nearest neighbor were optimistically universal for binary classification, it would also be optimistically universal in the general separable bounded case [2]. Additionally, in the Euclidean space, 1NN is universally consistent under all i.i.d. processes [10]. Further, $(k_t)_t$NN learning rules with $k_t/\log t \to \infty$ and $k_t/t \to 0$ are also universally consistent under i.i.d. processes for smooth classes of input spaces $\mathcal{X}$ [8]. However, it is known that there exist separable input spaces for which no $(k_t)_t$NN algorithm achieves universal consistency [6]. In this section, we show that $(k_t)_t$NN learning rules are not optimistically universal even on the interval $\mathcal{X} = [0, 1]$.

Theorem 2 *The $(k_t)_t-$nearest neighbor learning rule is not strongly optimistically universal for the input space $\mathcal{X} = [0, 1]$ with usual topology and for binary classification, for any sequence $(k_t)_t$ such that $k_t = o\left(\frac{t}{(\log t)^{1+\delta}}\right)$ for any $\delta > 0$.*

As a direct consequence, $(k_t)_t-$nearest neighbors are not optimistically universal for any input spaces $\mathcal{X}$ such that there exists a measurable injection $[0, 1] \to \mathcal{X}$ and for any output setting $(\mathcal{Y}, \ell)$ with bounded loss and at least two distinct values $y_1, y_2 \in \mathcal{Y}$ such that $\ell(y_1, y_2) > 0$. In particular, this shows that $(k_t)_t$NN algorithms are not optimistically universal in Euclidean spaces. To prove Theorem 2, we first define the set of processes with convergent relative frequencies CRF as the set of processes $\mathbb{X}$ such that $\forall A \in \mathcal{B}$,

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}_A(X_t) \quad \text{exists (a.s.).}$$

We then explicitly construct a process $\mathbb{X}^{(1)} \in$ CRF on which $(k_t)_t-$nearest neighbor fails. Because convergent relative frequencies processes are learnable CRF $\subset$ SUOL [18], this shows that $(k_t)_t$NN is not optimistically universal for the online learning setting. Note that we have CRF $\subsetneq$ SUOL for any infinite space $\mathcal{X}$. As a remark, it was already known that the self-adaptive/inductive nearest neighbor learning rule is not optimistically universal for the self-adaptive setting [18] (Section 3.2). Inductive learning differs from online learning in that the learner has access to a fixed historical dataset $(X_t, Y_t)_{t<T}$ and from time T has to commit to a (non-adaptive) learning rule. Self-adaptive learning is an intermediate setting between inductive learning and online learning where the learner can be adaptive on observed instances $(X_t)_{t\geq T}$ but not the values $(Y_t)_{t\geq T}$. Hence, the self-adaptive/inductive nearest neighbor learning rule corresponds to performing nearest neighbor with the fixed dataset $(X_u, Y_u)_{u<T}$ for any $t \geq T$. The performance of this learning rule is taken as a double limit: first as $t \to \infty$, then as $T \to \infty$. We refer to [18] for details on these settings. Similarly to the set SUOL, we can define the set SUAL of processes $\mathbb{X}$ admitting strong universal learning in the self-adaptive setting. The proof that self-adaptive nearest neighbor is not optimistically universal is also constructive but not relevant for the online setting because it relies on a completely different process $\mathbb{X}^{(2)} \in$ SUAL under which self-adaptive 1-nearest neighbor fails

but online learning 1-nearest neighbor is universally consistent. Indeed, the set of learnable processes for online learning is larger than the set of learnable processes for self-adaptive learning SUAL $\subset$ SUOL, and strictly larger whenever $\mathcal{X}$ is infinite [18].

Sketch of proof. For the sake of simplicity, we give the main arguments for the negative result on the $1-$NN learning rule which already provides all necessary intuitions. The process $\mathbb{X}^{(1)}$ is designed so that nearest neighbor fails on the function $f^*(\cdot) = \mathbb{1}_{\mathcal{D}}(\cdot)$ where $\mathcal{D}$ is the set of dyadics. Intuitively, the process alternates between a carefully chosen random dyadic $X_{n_k} \in \mathcal{D}$ and a sequence of random points $X_t,\ n_k < t < n_{k+1}$ which converges exponentially to X_{n_k} but that does not fall in the dyadics almost surely—which is achieved by including uniform noise in X_{n_k+1}. The nearest neighbor algorithm therefore uses the dyadic X_{n_k} as representant for most of the points X_t for $n_k < t < n_{k+1}$ and as a result assigns the wrong category $\hat{Y}_t = 1$. We then impose $n_{k+1} - n_k \to \infty$ so that nearest neighbor makes an asymptotic error rate of 1. A major technical difficulty is to ensure that the process $\mathbb{X}^{(1)}$ is still universally learnable. To do so, we aim to prove the stronger statement that $\mathbb{X}^{(1)} \in$ CRF. We randomly select X_{n_k} in high-order dyadics so that the convergence of the points X_t for $n_k < t < n_{k+1}$ is mild compared to the discretization of $[0, 1]$ of these high-order dyadics.

The generalization to $(k_t)_t-$nearest neighbor follows the same structure. The main difference in the construction of the process $\mathbb{X}^{(1)}$ consists in creating "copies" of the dyadic X_{n_k} using close high-order dyadics. As a result, the nearest neighbors of non-dyadic points are included in the set of copies of X_{n_k} which all provide wrong predictions. Note that keeping the process $\mathbb{X}^{(1)}$ in CRF constraints the possible number of copies. This results in a limitation $k_t = o\left(\frac{t}{\log t^{1+\delta}}\right)$ necessary to use Kolmogorov's convergence criteria for independent random variables.

5. An optimistically universal learning rule

In this section, we present an optimistically universal algorithm and give a characterization of SUOL. We start by defining our new learning rule $k-$Capped $1-$Nearest Neighbor (kC1NN) for any $k \geq 2$. This is a simple variant of the traditional 1NN learning rule where kC1NN performs the 1NN learning rule over a reduced training set. Recall that in the 1NN learning rule, we assign to the new input X_t the value of the nearest neighbor $Y_{NN(t)}$ where $NN(t) = \arg\min_{u<t} \rho(X_t, X_u)$. We refer to the input point $X_{NN(t)}$ as the representant of the input value X_t. In the kC1NN learning rule, we keep in memory the number of times n_t each point X_t is used as a representant for following input data and cap this value at k. Precisely, at each step t we update the dataset $\mathcal{D}_t \subset \{u,\ u < t\}$ containing the indices of data points on which 1NN may be performed. To do so, when n_u reaches k for some $u < t$, we delete u from the current dataset $\mathcal{D}_t$. At each iteration, if the input X_t has already been visited, we use simple memorization to predict Y_t, we do not update the values $(n_u)_{u<t}$ and do not include t in the dataset $\mathcal{D}_{t+1}$. Otherwise, kC1NN performs the 1NN learning rule on the current dataset $(X_u, Y_u)_{u\in\mathcal{D}_t}$, where ties can be broken arbitrarily for instance with minimum index, and updates $(n_u)_{u\in\mathcal{D}_t}$ and the dataset accordingly. In the following, we denote by $\phi(t)$ the index of the representant used for X_t, i.e. of its closest neighbor within the dataset $\mathcal{D}_t$. The rule is formally described in Algorithm 1.

In Section 4 we presented a process $\mathbb{X}$ on which nearest neighbor fails. The main reason for this failure is that some specific input points X_t can be used an arbitrarily large number of times as representant for future points, thereby inducing a large number of prediction errors. The learning

Algorithm 1: kC1NN learning rule

Input: Historical samples $(X_t, Y_t)_{t<T}$ and new input point X_T
Output: Predictions $\hat{Y}_t = kC1NN_t(\boldsymbol{X}_{<t}, \boldsymbol{Y}_{<t}, X_t)$ for $t \leq T$

$\hat{Y}_1 := 0$
$\mathcal{D}_2 := \{1\}$
$n_1 \leftarrow 0$
$t \leftarrow 2$
while $t \leq T$ **do**
  if *exists* $u < t$ *such that* $X_u = X_t$ **then**
    $\hat{Y}_t := Y_u$
    $\mathcal{D}_{t+1} := \mathcal{D}_t$
  else
    $\phi(t) := \arg\min_{u \in \mathcal{D}_t} \rho(X_t, X_u)$
    $\hat{Y}_t := Y_{\phi(t)}$
    $n_{\phi(t)} \leftarrow n_{\phi(t)} + 1$
    $n_t \leftarrow 0$
    if $n_{\phi(t)} = k$ **then**
      $\mathcal{D}_{t+1} := (\mathcal{D}_t \setminus \{\phi(t)\}) \cup \{t\}$
    else
      $\mathcal{D}_{t+1} := \mathcal{D}_t \cup \{t\}$
    end
  end
  $t \leftarrow t + 1$
end

rule kC1NN is designed precisely to tackle this issue by ensuring that any datapoint X_t for $t \geq 1$ is used at most k times as representant, i.e., $|\{u > t : \ \phi(u) = t\}| \leq k$. The goal of this section is to show that 2C1NN is optimistically universal for general separable Borel instance space $(\mathcal{X}, \rho)$ and near-metric separable value space $(\mathcal{Y}, \ell)$ with bounded loss.

We first start with the case of binary classification $(\{0,1\}, \ell_{01})$. This will then be used to prove the result for general output settings using a reduction technique introduced in [2]. Specifically, we show that 2C1NN is universally consistent for binary classification under all processes in $\text{SMV}_{(\mathcal{X},\rho)}$ which yields $\text{SMV}_{(\mathcal{X},\rho)} \subset \text{SUOL}_{(\mathcal{X},\rho),(\{0,1\},\ell_{01})}$. Together with Proposition 1, this shows that we have in fact an equality $\text{SMV}_{(\mathcal{X},\rho)} = \text{SUOL}_{(\mathcal{X},\rho),(\{0,1\},\ell_{01})}$ and as a result, that 2C1NN is optimistically universal. We start by showing that under any process in $\text{SMV}_{(\mathcal{X},\rho)}$, 2C1NN is consistent on functions representing balls of the metric ρ.

Proposition 5 *Let $(\mathcal{X}, \mathcal{B})$ be a separable Borel space constructed from the metric ρ. We consider the binary classification setting $\mathcal{Y} = \{0,1\}$ and the ℓ_{01} binary loss. For any input process $\mathbb{X} \in \text{SMV}_{(\mathcal{X},\rho)}$, for any $x \in \mathcal{X}$, and $r > 0$, the learning rule 2C1NN is consistent for the target function $f^* = \mathbb{1}_{B_\rho(x,r)}$.*

To prove this result, we introduce a tree structure $\mathcal{G}$ for the 2C1NN algorithm on times t such that each new input is linked to its representant which was used to derive the target prediction. Times t corresponding to instances X_t that were previously visited are therefore not considered in

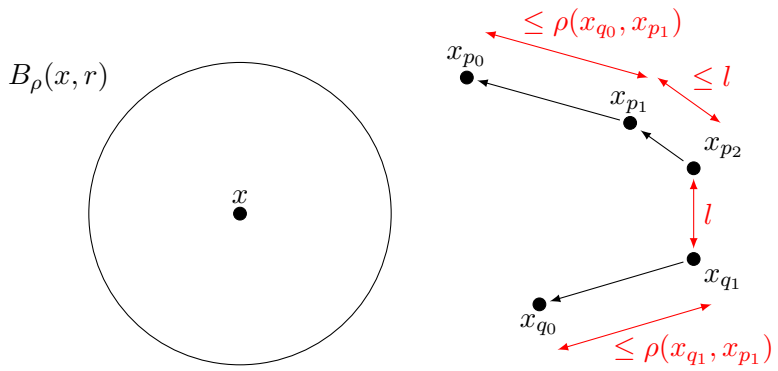

Figure 1: Illustration for Lemma 6 for $d = 2$ and $f = 1$, where the order of appearance is $p_0 < q_0 < p_1 < q_1 < p_2$. The arrows represent the relations of nodes within the tree $\mathcal{G}$, e.g., $\hat{Y}_{p_1} = Y_{p_0}$. If the end points x_{p_2} and x_{q_1} are close, then so are the beginning points x_{p_0} and x_{q_0}. The proof by induction is summarized by the upper bounds in red.

this tree. Precisely, we consider parent relations given by $(t, \phi(t))$ for all times t such that a new input X_t was visited—i.e. memorization was not performed directly. By definition of the 2C1NN learning rule, no time $t \in \mathcal{G}$ has more than 2 children. Further, for any $t, t' \in \mathcal{G}$, if the time $t' < t$ is not present in dataset $\mathcal{D}_t$, it has exactly 2 children. The proof uses the following lemma.

Lemma 6 *Consider two distinct paths $p_d \to p_{d-1} \to \ldots \to p_1 \to p_0$ and $q_f \to q_{f-1} \to \ldots \to q_1 \to q_0$ i.e. $\phi(p_i) = p_{i-1}$ for $1 \leq i \leq d$ and $\phi(q_i) = q_{i-1}$ for $1 \leq i \leq f$. Suppose $p_0 < q_0$ and that there exists $t \geq \max(p_d, q_f)$ such that $p_d, q_f \in \mathcal{D}_t$ (in other words the two end times are in some final dataset). Then, with $v(0) := \max\{0 \leq i \leq d,\ p_i < q_0\}$ we have*

$$\rho(x_{p_{v(0)}}, x_{q_0}) \leq 2^{f+d+1}\rho(x_{p_d}, x_{q_f}) \quad \textit{and} \quad \rho(x_{p_{v(0)}}, x_{p_d}) \leq 2^{f+d+1}\rho(x_{p_d}, x_{q_f}).$$

Sketch of proof of Proposition 5. Having fixed a horizon $T \geq 1$, the proof consists in analyzing the subgraph of $\mathcal{G}$ of times $t \leq T$ corresponding to points x_t falling in $B(x, r)$. These form a collection of disjoint trees where roots correspond to times where the 2C1NN algorithm made a mistake in the prediction. This structure allows to cluster times $\{t \leq T,\ x_t \in B(x, r)\}$ by connected component and show that these connected components are "well separated". Specifically, reasoning by contradiction, if two connected components were close from each other, so would be their roots as shown in Lemma 6. However, the $\text{SMV}_{(\mathcal{X},\rho)}$ property of $\mathbb{X}$ allows to "separate" the parents of all roots—input points which induced a future mistake for points in $B(x, r)$—which provides a contradiction. Hence, all connected components are well separated, and hence there should be a sublinear number of these components, i.e., of mistakes within the ball $B(x, r)$, using the $\text{SMV}_{(\mathcal{X},\rho)}$ property once again. The same analysis can be made for times $\{t \leq T,\ \rho(x_t, x) > r\}$. However, times falling precisely at the border $\{t \leq T,\ \rho(x_t, x) = r\}$ require an additional partition specifically on the border $\{t \leq T,\ \rho(x_t, x) = r\}$.

We are now ready to show that 2C1NN is optimistically universal for the binary classification setting under processes of $\text{SMV}_{\mathcal{X},\rho)}$. Intuitively, given a fixed process $\mathbb{X} \in \text{SMV}_{(\mathcal{X},\rho)}$, we analyze

the set of functions on which 2C1NN is consistent and show that it is a $\sigma-$algebra. Further, Proposition 5 shows that this $\sigma-$algebra contains all open balls, and as a consequence is the complete Borel $\sigma-$algebra, i.e., 2C1NN is universally consistent under $\mathbb{X}$.

Theorem 7 *Let $(\mathcal{X}, \mathcal{B})$ be a separable Borel space. For the binary classification setting, the learning rule 2C1NN is universally consistent for all processes $\mathbb{X} \in SMV_{(\mathcal{X},\rho)}$.*

Sketch of proof. We fix a process $\mathbb{X} \in \text{SMV}_{(\mathcal{X},\rho)}$ and define $\mathcal{S}_{\mathbb{X}}$ as the set of functions on which it is consistent. Because we focus on the binary classification we can write

$$\mathcal{S}_{\mathbb{X}} := \{A \in \mathcal{B},\ \mathcal{L}_{\mathbb{X}}(2C1NN, \mathbb{1}_A) = 0 \quad (a.s.)\}.$$

Proposition 5 implies that all open balls are included in $\mathcal{S}_{\mathbb{X}}$. Thus, it suffices to check that it satisfies the properties of a $\sigma-$algebra. The invariance of 2C1NN to value labels directly shows that $\mathcal{S}_{\mathbb{X}}$ is invariant to complementary. Showing the $\sigma-$additivity is the main technical difficulty of this result. Let $(A_i)_{i\geq 0}$ be a sequence of disjoint set of $\mathcal{S}_{\mathbb{X}}$. Writing $A = \bigcup_{i\geq 0} A_i$, we wish to bound type I errors—$X_{\phi(t)} \in A$ but $X_t \notin A$—and type II errors—$X_{\phi(t)} \notin A$ but $X_t \in A$. The main interest of the 2C1NN rule is that any input point X_t can induce at most 3 prediction errors: potentially a prediction mistake for Y_t and at most 2 children. Hence, we can lower bound the number of errors of type I until time horizon $T \geq 1$ by the number of distinct points falling in A within horizon T. The errors of type II already correspond to times with points falling in A. In other words, the property of 2C1NN that any point can induce at most a finite number of future mistakes implies that to make a linear number of mistakes, 2C1NN must visit the set $A = \bigcup_{i\geq 0} A_i$ a linear number of times. We then use 1. the $\text{SMV}_{(\mathcal{X},\rho)}$ property on the partition $(A_i)_{i\geq 0}$, and 2. the fact that individually on each A_i, 2C1NN makes a sublinear number of errors, to prove that the total number of errors for $\mathbb{1}_A$ is sublinear.

In particular, Theorem 7 shows that $\text{SMV}_{(\mathcal{X},\rho)} \subset \text{SUOL}_{(\mathcal{X},\rho),([0,1],\ell_{01})}$. Together with Proposition 1, this shows that the set of learnable processes for binary classification is exactly $\text{SMV}_{(\mathcal{X},\rho)}$. As a result, 2C1NN is optimistically universal for binary classification. The binary classification setting is not restrictive. Indeed [2] shows that any bounded separable value space $(\mathcal{Y}, \ell)$ can be reduced to binary classification using randomized hypothesis testing:

Theorem 8 (Blanchard and Cosson [2]) *Let $\mathcal{X}$ be a Borel space and $k \geq 2$. For any separable near-metric space $(\mathcal{Y}, \ell)$ with $0 < \bar{\ell} < \infty$, we have $SUOL_{(\mathcal{X},\rho),(\mathcal{Y},\ell)} = SUOL_{(\mathcal{X},\rho),(\{0,1\},\ell_{01})}$. Further, if there exists an optimistically universal for the binary classification setting, then there exists an optimistically universal for the setting $(\mathcal{Y}, \ell)$. Finally, if kC1NN is optimistically universal for binary classification, it is also optimistically universal for the setting $(\mathcal{Y}, \ell)$.*

Applying this reduction from a general bounded output setting to binary classification, we obtain a full characterization of the set of processes admitting strong universal learning for general value spaces and obtain that 2C1NN is optimistically universal for general input and output spaces.

Corollary 9 *For any separable Borel space $\mathcal{X}$ and any separable near-metric space $(\mathcal{Y}, \ell)$ with $0 < \bar{\ell} < \infty$, we have $SUOL_{(\mathcal{X},\rho),(\mathcal{Y},\ell)} = SMV_{(\mathcal{X},\rho)}$.*

Corollary 10 *For any separable Borel space $\mathcal{X}$, and any bounded separable near-metric space $(\mathcal{Y}, \ell)$, 2C1NN is an optimistically universal learning rule.*

Note that the case $\bar{\ell} = 0$ can be treated separately: in this case all processes $\mathbb{X}$ are learnable and any learning rule is optimistically universal. The same proofs imply that the learning rules kC1NN are also optimistically universal for any $k \geq 2$. As a remark, one can note that 2C1NN is the simplest algorithm of this class which is optimistically universal. Indeed, 1C1NN systematically deletes the previous points from the dataset and as a result, at any time, the dataset $\mathcal{D}_t$ is a singleton. Hence, 1C1NN is not optimistically universal. To present a simpler exposition of the optimistical universal consistency of rules from the class kC1NN, we provide in Appendix B a proof of the result for the special case of 4C1NN and the space $\mathcal{X} = [0, 1]$ with the usual topology. This can then directly be generalized to all standard Borel spaces using the Kuratowski theorem. This proof provides the main ideas for the general result without the added technicalities induced by constructing partitions for general separable spaces $(\mathcal{X}, \rho)$, or reduced convergence speeds for 2C1NN compared to 4C1NN.

6. Weak universal learning

We now turn to weak universal learning. In this section, we show that the results for the characterization of learnable processes and the existence of optimistically universal learning rules for the strong setting can also be adapted to the weak setting. Although the set of learnable processes differ—SUOL $\subset$ WUOL in general and SUOL $\subsetneq$ WUOL whenever $\mathcal{X}$ is infinite [18]—we show that the same learning rule 2C1NN is optimistically universal in the weak setting. We recall the necessary condition WSMV for weak learnability for near-metric separable value spaces $(\mathcal{Y}, \ell)$ with $0 < \bar{\ell} < \infty$.

Condition WSMV For every countable measurable partition $\{A_k\}_{k=1}^{\infty}$,

$$\mathbb{E}[|\{k \in \mathbb{N} : A_k \cap \mathbb{X}_{<T} \neq \emptyset\}|] = o(T).$$

Similarly to the strong consistency case, we will show that $\text{WSMV}_{(\mathcal{X},\rho)}$ is also sufficient for weak universal consistency. Note that whenever $\mathcal{X}$ is infinite we have $\text{WSMV}_{(\mathcal{X},\rho)} \subsetneq \text{SMV}_{(\mathcal{X},\rho)}$. We start by adapting Proposition 5 for the weak setting by showing that 2C1NN is weakly consistent on balls under any process $\mathbb{X} \in \text{WSMV}_{(\mathcal{X},\rho)}$.

Proposition 11 *Let $(\mathcal{X}, \mathcal{B})$ be a separable Borel space constructed from some metric ρ. We consider the binary classification setting $\mathcal{Y} = \{0, 1\}$ and the ℓ_{01} binary loss. For any input process $\mathbb{X} \in$ WSMV$_{(\mathcal{X},\rho)}$, for any $x \in \mathcal{X}$, and $r > 0$, the learning rule 2C1NN is weakly consistent for the target function $f^* = \mathbb{1}_{B_\rho(x,r)}$.*

We then show that 2C1NN is weakly consistent under processes of $\text{WSMV}_{(\mathcal{X},\rho)}$ for binary classification adapting the proof of Theorem 7.

Theorem 12 *Let $(\mathcal{X}, \mathcal{B})$ be a separable Borel space constructed from the metric ρ. For the binary classification setting, the learning rule 2C1NN is weakly universally consistent for all processes $\mathbb{X} \in$ WSMV$_{(\mathcal{X},\rho)}$.*

Finally, we turn to the case of a bounded separable output setting $(\mathcal{Y}, \ell)$ and show that 2C1NN is weakly optimistically universal. In the case of weak learning, the reduction from any separable bounded output setting does not require a sophisticated argument as in the proof of Theorem 8 [2], and can be made using the dominated convergence theorem.

Theorem 13 *Let $(\mathcal{X}, \mathcal{B})$ be a separable Borel space constructed from the metric ρ. The learning rule 2C1NN is weakly universally consistent for all processes $\mathbb{X} \in WSMV_{(\mathcal{X},\rho)}$ and any separable bounded output setting $(\mathcal{Y}, \ell)$.*

As an immediate consequence, we have $\text{WSMV}_{(\mathcal{X},\rho)} \subset \text{WUOL}_{(\mathcal{X},\rho),(\mathcal{Y},\ell)}$. Together with Proposition 16 we obtain a complete characterization for weak learnable processes and show that 2C1NN is weakly optimistically universal for general output value spaces.

Corollary 14 *For any separable Borel space $(\mathcal{X}, \mathcal{B})$, and every separable near metric space $(\mathcal{Y}, \ell)$ with $0 < \bar{\ell} < \infty$ we have $WUOL_{(\mathcal{X},\rho),(\mathcal{Y},\ell)} = WSMV_{(\mathcal{X},\rho)}$. In particular, SUOL is invariant from the output setup.*

Corollary 15 *For any separable Borel space $(\mathcal{X}, \mathcal{B})$ and any bounded separable output setting $(\mathcal{Y}, \ell)$, 2C1NN is weakly optimistically universal.*

This completely closes the main questions on universal online learning [18; 19] as we have now proved Theorem 3 (Corollary 10 and 15) and Theorem 4 (Corollary 9 and 14).

7. Conclusion

In this paper, we provided a strong and weak optimistically universal learning rule 2C1NN, which is a simple variant of the nearest neighbor algorithm. We further gave a characterization of the processes admitting strong or weak universal learning, closing the study of universal online learning with bounded losses.

The case of unbounded losses was already settled in [18; 4], which was shown to be very restrictive because the target functions are unrestricted. It would be interesting to bridge the gap between these two cases by considering *restricted* universal learning. Specifically, by adding a constraint on the target functions—for example, moment constraints are fairly common in the literature [15; 16]—one could hope to recover the large set of learnable processes SUOL characterized in this paper, even for the unbounded loss case. We refer to [4] for further motivation of this open direction.

In our setting, we assume that the values are generated from the stochastic process $\mathbb{X}$ through a target function f^* and without noise. Another interesting line of research would be to add noise to the value process $\mathbb{Y}$. This relates to the Bayes consistency literature in which an objective is to reach the minimal risk, known as the Bayes minimal risk; instead of obtaining exact consistency i.e. vanishing average error rate as considered in this paper. A possible direction would be to find mild independence conditions on the noise—generalizing the i.i.d. setting [31]—so that there exist learning rules which are Bayes universally consistent under a large set of processes $\mathbb{X}$. The author notes that subsequently to this paper, [3] used the results of this work to design universal learning rules for arbitrary noise in the responses.

Acknowledgments

The author is very grateful to Patrick Jaillet, Romain Cosson and Steve Hanneke for very useful discussions and for reviewing the manuscript. This work is being partly funded by ONR grant N00014-18-1-2122.

References

[1] Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. In *COLT*, volume 3, page 1, 2009.

[2] Moïse Blanchard and Romain Cosson. Universal online learning with bounded loss: Reduction to binary classification. *arXiv preprint arXiv:2112.14638*, 2021.

[3] Moïse Blanchard and Patrick Jaillet. Universal regression with adversarial responses. *arXiv preprint arXiv:2203.05067*, 2022.

[4] Moïse Blanchard, Romain Cosson, and Steve Hanneke. Universal online learning with unbounded losses: Memory is all you need. In *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*, pages 107–127. PMLR, 2022.

[5] Olivier Bousquet, Steve Hanneke, Shay Moran, Ramon van Handel, and Amir Yehudayoff. A theory of universal learning. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 532–541, 2021.

[6] Frédéric Cérou and Arnaud Guyader. Nearest neighbor classification in infinite dimension. *ESAIM: Probability and Statistics*, 10:340–355, 2006.

[7] Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[8] Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for nearest neighbor classification. *Advances in Neural Information Processing Systems*, 27, 2014.

[9] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[10] Luc Devroye, Laszlo Gyorfi, Adam Krzyzak, and Gábor Lugosi. On the strong universal consistency of nearest neighbor regression function estimates. *The Annals of Statistics*, pages 1371–1385, 1994.

[11] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

[12] Robert M Gray and RM Gray. *Probability, random processes, and ergodic properties*, volume 1. Springer, 2009.

[13] László Gyöfi and Gábor Lugosi. Strategies for sequential prediction of stationary time series. In *Modeling uncertainty*, pages 225–248. Springer, 2002.

[14] L Gyorfi, Gábor Lugosi, and Gusztáv Morvai. A simple randomized algorithm for sequential prediction of ergodic time series. *IEEE Transactions on Information Theory*, 45(7):2642–2650, 1999.

[15] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag New York, 2002.

[16] Lásló Györfi and György Ottucsák. Sequential prediction of unbounded stationary time series. *IEEE Transactions on Information Theory*, 53(5):1866–1872, 2007.

[17] László Györfi and Roi Weiss. Universal consistency and rates of convergence of multiclass prototype algorithms in metric spaces. *Journal of Machine Learning Research*, 22(151):1–25, 2021.

[18] Steve Hanneke. Learning whenever learning is possible: Universal learning under general stochastic processes. *Journal of Machine Learning Research*, 22(130):1–116, 2021.

[19] Steve Hanneke. Open problem: Is there an online learning algorithm that learns whenever online learning is possible? In *Conference on Learning Theory*, pages 4642–4646. PMLR, 2021.

[20] Steve Hanneke, Aryeh Kontorovich, Sivan Sabato, and Roi Weiss. Universal Bayes consistency in metric spaces. *The Annals of Statistics*, 49(4):2129 – 2150, 2021.

[21] David Haussler, Nick Littlestone, and Manfred K Warmuth. Predicting $\{0,1\}$-functions on randomly drawn points. *Information and Computation*, 115(2):248–292, 1994.

[22] Alexander Kechris. *Classical descriptive set theory*, volume 156. Springer Science & Business Media, 2012.

[23] Sanjeev R Kulkarni, Steven E Posner, and Sathyakama Sandilya. Data-dependent k_n-NN and kernel estimators consistent for arbitrary processes. *IEEE Transactions on Information Theory*, 48(10):2785–2788, 2002.

[24] Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2(4):285–318, 1988.

[25] Gusztáv Morvai, Sidney Yakowitz, and László Györfi. Nonparametric inference for ergodic, stationary time series. *The Annals of Statistics*, 24(1):370–379, 1996.

[26] Gusztáv Morvai, Sanjeev R Kulkarni, and Andrew B Nobel. Regression estimation from an individual stable sequence. *Statistics: A Journal of Theoretical and Applied Statistics*, 33(2): 99–118, 1999.

[27] Alexander Rakhlin, Karthik Sridharan, and Ambuj Tewari. Online learning via sequential complexities. *J. Mach. Learn. Res.*, 16(1):155–186, 2015.

[28] Daniil Ryabko and Peter Bartlett. Pattern recognition for conditionally independent data. *Journal of Machine Learning Research*, 7(4), 2006.

[29] Ingo Steinwart, Don Hush, and Clint Scovel. Learning from dependent observations. *Journal of Multivariate Analysis*, 100(1):175–194, 2009.

[30] Charles J Stone. Consistent nonparametric regression. *The Annals of Statistics*, pages 595–620, 1977.

[31] Dan Tsir Cohen and Aryeh Kontorovich. Metric-valued regression. *Submitted to COLT*, 2022.

[32] Ruth Urner and Shai Ben-David. Probabilistic lipschitzness a niceness assumption for deterministic labels. In *Learning Faster from Easy Data-Workshop@ NIPS*, volume 2, page 1, 2013.

Appendix A. Proofs of Section 4: On nearest neighbor consistency

A.1. Proof of Theorem 2

Let $\delta > 0$ and a sequence $k_t = o\left(\frac{t}{(\log t)^{1+\delta}}\right)$. To show that $(k_t)_t-$NN is not optimistically universal, we construct a process $\mathbb{X} \in$ SUOL on which $(k_t)_t-$NN has asymptotic error rate 1. We denote by $\mathcal{D}_p := \{\frac{i}{2^p}, 0 \leq i \leq 2^p,\ i \text{ odd}\}$ the set of dyadics of order p i.e. with reduced denominator 2^p, and $\mathcal{D}$ for the set of dyadics. Let $\epsilon > 0$ such that $\frac{1+2\epsilon}{1-2\epsilon} < 1 + \frac{\delta}{2}$. Then pose for $k \geq 1$,

$$n_k = \lfloor e^{k^{1/2-\epsilon}} \rfloor, \quad d_k = \min\left(\left\lfloor \frac{n_k}{(\log n_k)^{1+\delta}} \right\rfloor, n_{k+1} - n_k - 1\right), \quad p_k = 4^k.$$

First note that $n_{k+1} - n_k \sim \left(\frac{1}{2} - \epsilon\right) \frac{n_k}{k^{1/2+\epsilon}} \sim \left(\frac{1}{2} - \epsilon\right) \frac{n_k}{(\log n_k)^{\frac{1/2+\epsilon}{1/2-\epsilon}}}$ therefore we obtain

$$d_k = o\left(\frac{n_k}{(\log n_k)^{1+\delta/2}}\right) = o(n_{k+1} - n_k).$$

Also, for k large enough, $d_k = \left\lfloor \frac{n_k}{(\log n_k)^{1+\delta}} \right\rfloor$. We now construct a process $\mathbb{X}$ on $\mathcal{X}$. Let $(U_k)_{k\geq 1}$ be an i.i.d. sequence of uniforms $\mathcal{U}([0,1])$ and $(D_k)_{k\geq 1}$ a sequence of independent random variables—also independent of $(U_k)_k$—such that $D_k \sim \mathcal{U}(\mathcal{D}_{p_k})$. Additionally, we denote by $D_{k,i}$ the $i-$th closest dyadic of order p_k to D_k. For instance, $D_{k,1} = D_k$, and $|D_{k,i} - D_k| \leq \frac{i}{2^{p_k - 1}}$. For intuition, if D_k is not close to the boundary of $[0,1]$, we have $D_{k,i} = D_k + (-1)^i \cdot \frac{\lfloor i/2 \rfloor}{2^{p_k}}$. We now define the process $\mathbb{X}$ as follows for any $k \geq 1$,

$$X_{n_k+i} = D_{k,i+1}, \quad 0 \leq i \leq d_k \quad \text{and} \quad X_{n_k+d_k+j} = D_k + \frac{U_k - D_k}{2^{n_k} 4^j}, \quad 1 \leq j < n_{k+1} - n_k - d_k.$$

We first prove that $(k_t)_t-$NN is not consistent for the function $f^* = 1_{\mathcal{D}}$. For any $k \geq 1$,

$$\mathbb{P}\left[\min_{t<n_k} |X_t - D_k| < \frac{1}{2^{n_k}}\right] \leq \sum_{t<n_k} \mathbb{P}\left[X_t - \frac{1}{2^{n_k}} < D_k < X_t + \frac{1}{2^{n_k}}\right] \leq \frac{2n_k}{2^{n_k}},$$

because $n_k \leq p_k$. Now note that for all $k \geq 1$ and $0 \leq i \leq d_k$ we have $X_{n_k+i} \in \mathcal{D}_{p_k}$, while almost surely, all other random variables do not fall in $\mathcal{D}$. Then, denote by $\mathcal{E}$ the event of probability 1 where $\mathbb{X}$ does not visit $\mathcal{D}$ except for times $n_k + i$ for $k \geq 1$ and $0 \leq i \leq d_k$. In other words,

$$\mathcal{E} := \{X_{n_k+i} \notin \mathcal{D}, \quad k \geq 1, d_k < i < n_{k+1} - n_k\}$$

and $\mathbb{P}(\mathcal{E}) = 1$. We also denote by $\mathcal{A}_k$ the event $\mathcal{A}_k := \{\min_{t<n_k} |X_t - D_k| \geq 2^{-n_k}\}$ and $\mathcal{B}_k$ the event $\mathcal{B}_k := \{|U_k - D_k| \geq 2^{-k}\}$. We have $\mathbb{P}(\mathcal{B}_k^c) \leq 2^{-k+1}$ and we showed previously $\mathbb{P}(\mathcal{A}_k^c) \leq \frac{2n_k}{2^{n_k}}$. Now note that $\frac{d_k}{2^{p_k - 1}} = o(\frac{1}{2^{n_k + 2n_{k+1} + k+1}})$. Therefore, let k_0 such that for any

$k \geq k_0$, $\frac{d_k}{2^{p_k-1}} \leq \frac{1}{2^{n_k+2n_{k+1}+k+1}}$. Then, for any $k \geq k_0$, on the event $\mathcal{A}_k \cap \mathcal{B}_k \cap \mathcal{E}$, for any $1 \leq j < n_{k+1} - n_k - d_k$, the $d_k + 1$ nearest neighbors of $X_{n_k+d_k+j}$ are exactly the points $\{X_{n_k+i} = D_{k,i+1},\ 0 \leq i \leq d_k\}$. Indeed,

$$|X_{n_k+d_k+j} - D_{k,i}| \leq |X_{n_k+d_k+j} - D_k| + \frac{d_k}{2^{p_k-1}} \leq \frac{1}{2^{n_k}4^j} + \frac{1}{2^{n_k+2j}} < \frac{1}{2^{n_k+2j-1}}.$$

Further, for all $t < n_k$,

$$|X_{n_k+d_k+j} - X_t| \geq |D_k - X_t| - |X_{n_k+d_k+j} - D_k| \geq \frac{1}{2^{n_k}} - \frac{1}{2^{n_k+2}} > \frac{1}{2^{n_k+2j-1}}.$$

and finally, for $1 \leq j' < j$ and any $0 \leq i \leq d_k$, we have

$$\begin{aligned}
|X_{n_k+d_k+j} - X_{n_k+d_k+j'}| &\geq |X_{n_k+d_k+j} - X_{n_k+d_k+j-1}| = 3 \cdot \frac{|U_k - D_k|}{2^{n_k+2j}} \\
&\geq |X_{n_k+d_k+j} - D_k| + 2 \cdot \frac{1}{2^{n_k+2j+k}} \\
&\geq |X_{n_k+d_k+j} - D_k| + 2 \cdot \frac{d_k}{2^{p_k-1}} \\
&> |X_{n_k+d_k+j} - D_k| + |D_k - D_{k,i}| \\
&\geq |X_{n_k+d_k+j} - D_{k,i}|.
\end{aligned}$$

We now observe that

$$\max_{n_k+d_k+1 \leq t < n_{k+1}} k_t = o\left(\frac{n_{k+1}}{(\log n_k)^{1+\delta}}\right) = o(d_k).$$

Therefore, let k_1 such that for any $k \geq k_1$, and any $1 \leq j < n_{k+1} - n_k - d_k$, we have $k_{n_k+d_k+j} \leq d_k$. Now for any $k \geq \max(k_0, k_1)$, on the event $\mathcal{A}_k \cap \mathcal{B}_k \cap \mathcal{E}$, $(k_t)_t$NN makes an error in the prediction of all $X_{n_k+d_k+j}$ for $1 \leq j < n_{k+1} - n_k - d_k$ since its k_t closest neighbors are in the set $\{X_{n_k+i} = D_{k,i+1},\ 0 \leq i \leq d_k\}$ which all have value $\mathbb{1}_{\mathcal{D}}(X_{n_k+i}) = 1$ instead of $\mathbb{1}_{\mathcal{D}}(X_{n_k+d_k+j}) = 0$.

Last, note that the frequency of the times of the form $n_k + i$ for $k \geq 1$ and $0 \leq i \leq d_k$ vanishes to 0, because $d_k = o(n_{k+1} - n_k)$ and $n_{k+1} \sim n_k$. Therefore, on the event $\mathcal{E} \cap \cup_{k' \geq 1} \bigcap_{k \geq k'} (\mathcal{A}_k \cap \mathcal{B}_k)$, the learning rule $(k_t)_t$NN has error rate $\mathcal{L}_{\mathbb{X}}((k_t)_t NN, f^*) = 1$. Now note that $\mathbb{P}[\mathcal{E} \cap \mathcal{A}_k^c \cap \mathcal{B}_k^c] \leq 2^{-k+1} + \frac{2n_k}{2^{n_k}}$. Because we have $\sum_{n \geq 1} 2^{-n+1} < \infty$ and $\sum_{n \geq 1} \frac{2n}{2^n} < \infty$, the Borel-Cantelli lemma implies $\mathbb{P}[\mathcal{E} \cap \cup_{k' \geq 1} \bigcap_{k \geq k'} (\mathcal{A}_k \cap \mathcal{B}_k)] = 1$. To summarize, with probability one, $(k_t)_t$NN has error rate 1 hence is not consistent for process $\mathbb{X}$ and target function $f^* = \mathbb{1}_{\mathcal{D}}$. This ends the proof that $(k_t)_t$NN is not universally consistent for process $\mathbb{X}$.

We now show that $\mathbb{X} \in$ SUOL by showing that in fact $\mathbb{X} \in$ CRF. Let $A \subset [0, 1]$. We will show that the frequencies of falling in A converge almost surely to $\mu(A)$ where μ is the Lebesgue measure. We introduce the random variables

$$Y_k = \sum_{i=0}^{n_{k+1}-n_k-1} \mathbb{1}_A(X_{n_k+i}).$$

Again, for $k \geq 1$, and $1 \leq j < n_{k+1} - n_k - d_k$, $X_{n_k+d_k+j}$ is an absolutely continuous random variable with density $f(x) = \frac{1}{2^{p_k-1}} \sum_{l=0}^{2^{p_k-1}-1} f_l(x)$ where $f_l(x)$ corresponds to the conditional

density to $D_k = \frac{2l+1}{2^{p_k}} =: d_l$, i.e.

$$f_l(x) = 2^{n_k} 4^j \cdot \mathbb{1}\left(x \in \left[d_l - \frac{d_l}{2^{n_k}4^j}, d_l + \frac{1-d_l}{2^{n_k}4^j}\right]\right)$$

But $x \in \left[d_l - \frac{d_l}{2^{n_k}4^i}, d_l + \frac{1-d_l}{2^{n_k}4^i}\right]$ i.if $\frac{2^{p_k-1}(x-\frac{1}{2^{n_k}4^i})}{1-\frac{1}{2^{n_k}4^i}} - \frac{1}{2} \leq l \leq \frac{2^{p_k-1}x}{1-\frac{1}{2^{n_k}4^i}} - \frac{1}{2}$. Therefore, the number $N(x)$ of non-zero terms in the sum $f(x) = \frac{1}{2^{p_k-1}} \sum_{l=0}^{2^{p_k-1}-1} f_l(x)$ is

$$\frac{2^{p_k-1-n_k-2i}}{1-\frac{1}{2^{n_k}4^i}} - 1 \leq N(x) \leq \frac{2^{p_k-1-n_k-2i}}{1-\frac{1}{2^{n_k}4^i}} + 1$$

Hence,

$$\left| f(x) - \frac{1}{1-\frac{1}{2^{n_k}4^i}} \right| = \left| \frac{2^{n_k}4^i N(x)}{2^{p_k-1}} - \frac{1}{1-\frac{1}{2^{n_k}4^i}} \right| \leq \frac{1}{2^{p_k-1-n_k-2i}}.$$

Finally, we obtain

$$|\mathbb{P}(X_{n_k+d_k+j} \in A) - \mu(A)| \leq \frac{1}{2^{p_k-1-n_k-2j}} + \frac{1}{2^{n_k+2j}-1}.$$

Therefore,

$$\begin{aligned}
|\mathbb{E}Y_k - (n_{k+1}-n_k)\mu(A)| &\leq \sum_{i=0}^{d_k} |\mathbb{P}(X_{n_k+i} \in A) - \mu(A)| + \sum_{j=1}^{n_{k+1}-n_k-d_k-1} \mathbb{P}(X_{n_k+d_k+j} \in A) \\
&\leq d_k + 1 + \frac{n_{k+1}-n_k}{2^{p_k-2n_{k+1}}} + \frac{n_{k+1}-n_k}{2^{n_k}-1} \\
&\leq d_k + C
\end{aligned}$$

where $C \geq 1$ is some universal constant, given that $\frac{n_{k+1}-n_k}{2^{p_k-2n_{k+1}}} \to 0$ and $\frac{n_{k+1}-n_k}{2^{n_k}-1} \to 0$ as $k \to \infty$. Now note that because Y_k is a sum of $n_{k+1} - n_k$ random variables bounded by 1, then

$$Var(Y_k) \leq (n_{k+1}-n_k)^2 = \mathcal{O}\left(\frac{n_{k+1}^2}{k^{1+2\epsilon}}\right).$$

Therefore, $\sum_{k\geq 1} \frac{Var(Y_k)}{(n_{k+1}-1)^2} < \infty$. Further, we can note that the random variables $(Y_k)_{k\geq 1}$ are together independent. Thus, by Kolmogorov's Convergence Criteria, we obtain

$$\sum_{l=1}^{k} \frac{Y_l - \mathbb{E}Y_l}{n_{k+1}-1} \to 0 \quad (a.s.)$$

We then apply Kronecker's lemma which gives

$$\epsilon_k := \frac{\sum_{l=1}^{k} Y_l - \mathbb{E}Y_l}{n_{k+1}-1} \to 0 \quad (a.s.)$$

We now compute,

$$\begin{aligned}\left|\frac{1}{n_{k+1}-1}\sum_{t=1}^{n_{k+1}-1}\mathbb{1}_A(X_t)-\mu(A)\right| &= \frac{1}{n_{k+1}-1}\left|\sum_{l=1}^{k}Y_l-(n_{k+1}-n_k)\mu(A)\right| \\ &= \frac{1}{n_{k+1}-1}\left|(n_{k+1}-1)\epsilon_k+\sum_{l=1}^{k}\mathbb{E}Y_l-(n_{k+1}-n_k)\mu(A)\right| \\ &\leq \epsilon_k+\frac{Ck+\sum_{l=1}^{k}d_l}{n_{k+1}-1}.\end{aligned}$$

Because $\frac{k}{n_{k+1}-1}\to 0$ and $\sum_{l=1}^{k}d_l=o(n_{k+1}-1)$, we obtain $\frac{1}{n_{k+1}-1}\sum_{t=1}^{n_{k+1}-1}\mathbb{1}_A(X_t)\to\mu(A)\quad(a.s.)$. We complete the proof by noting that for any $n_k\leq T<n_{k+1}$,

$$\frac{1}{n_{k+1}-1}\sum_{t=1}^{n_k-1}\mathbb{1}_A(X_t)\leq\frac{1}{T}\sum_{t=1}^{T}\mathbb{1}_A(X_t)\leq\frac{1}{n_k-1}\sum_{t=1}^{n_{k+1}-1}\mathbb{1}_A(X_t),$$

and that $\frac{n_k-1}{n_{k+1}-1}\to 1$ as $k\to\infty$. Therefore $\frac{1}{T}\sum_{t=1}^{T}\mathbb{1}_A(X_t)\to\mu(A)\quad(a.s.)$ which shows that $\mathbb{X}\in$ CRF. Because CRF $\subset$ SUOL [18], this ends the proof of the theorem.

Appendix B. An optimistically universal learning rule for standard Borel spaces

To provide a simpler exposition of the main results Theorem 3 and Theorem 4, we now show that kC1NN is in fact optimistically universal for $k\geq 4$ starting with $\mathcal{X}=[0,1]$. This will in turn give the result for general standard Borel space as shown in Appendix B.2 and already provides all the intuitions necessary for the general case presented in Section 5 and proved in Appendix C.

B.1. Universal online learning on $\mathcal{X}=[0,1]$

We will consider the case $\mathcal{X}=[0,1]$ in this section and show that 4C1NN is optimistically universal for this input space. To do so, we prove that 4C1NN is universally consistent under all processes in $\text{SMV}_{([0,1],|\cdot|)}$ which yields $\text{SMV}_{([0,1],|\cdot|)}\subset\text{SUOL}_{([0,1],|\cdot|),(\{0,1\},\ell_{01})}$. Together with Proposition 1, this will show that $\text{SUOL}_{([0,1],|\cdot|),(\{0,1\},\ell_{01})}=\text{SMV}_{([0,1],|\cdot|)}$ and as a result, that 4C1NN is optimistically universal. As a first step, we focus on the simple function f^* represented by the fixed interval $[0,1/2]$ in the binary classification setting, and show that 4C1NN is consistent under any input process for this target function.

Proposition 16 *Let $\mathcal{X}=[0,1]$ with the usual topology. We consider the binary classification setting $\mathcal{Y}=\{0,1\}$ with ℓ_{01} binary loss. Under any input process $\mathbb{X}\in$ SMV$_{([0,1],|\cdot|)}$, the learning rule 4C1NN is strongly consistent for the target function $f^*=\mathbb{1}_{[0,1/2]}$.*

Proof We reason by the contrapositive and suppose that 4C1NN is not consistent on f^*. We will show that the process $\mathbb{X}$ disproves the $\text{SMV}_{([0,1],|\cdot|)}$ condition by considering the partition $\mathcal{P}$ of $\mathcal{X}$ defined by

$$\left\{\frac{1}{2}\right\}\cup\bigcup_{k\geq 1}\left[\frac{1}{2}-\frac{1}{2k};\frac{1}{2}-\frac{1}{2(k+1)}\right)\cup\bigcup_{k\geq 1}\left(\frac{1}{2}+\frac{1}{2(k+1)};\frac{1}{2}+\frac{1}{2k}\right].$$

Precisely, we will show that the process does not visit a sublinear number of sets of this partition with nonzero probability.

Because 4C1NN is not consistent, $\delta := \mathbb{P}(\mathcal{L}_{\mathbb{X}}(4C1NN, f^*) > 0) > 0$. Define

$$\mathcal{A} := \{\mathcal{L}_{\mathbb{X}}(4C1NN, f^*) > 0\}.$$

We now consider a specific realization $\boldsymbol{x} = (x_t)_{t\geq 0}$ of the process $\mathbb{X}$ falling in the event $\mathcal{A}$. Note that $\boldsymbol{x}$ is not random anymore. We now show that $\boldsymbol{x}$ does not visit a sublinear number of sets in the partition $\mathcal{P}$. By construction $\epsilon := \mathcal{L}_{\boldsymbol{x}}(4C1NN, f^*) > 0$. We now denote by $(t_k)_{k\geq 1}$ the increasing sequence of all times when 4C1NN makes an error in the prediction of $f^*(x_t)$. Now define an increasing sequence of times $(T_l)_{l\geq 1}$ such that

$$\frac{1}{T_l}\sum_{t=1}^{T_l} \ell_{01}(4C1NN(\boldsymbol{x}_{<t}, \boldsymbol{y}_{<t}, x_t), f^*(x_t)) > \frac{\epsilon}{2}.$$

For any $l \geq 1$ consider the last index $k = \max\{u, t_u \leq T_l\}$ when 4C1NN makes a mistake. Then we obtain $k > \frac{\epsilon}{2}T_l \geq \frac{\epsilon}{2}t_k$. Considering the fact that $(T_l)_{l\geq 1}$ is an increasing unbounded sequence we therefore obtain an increasing sequence of indices $(k_l)_{l\geq 1}$ such that $t_{k_l} < \frac{2k_l}{\epsilon}$.

At an iteration where the new input x_t has not been previously visited we will denote by $\phi(t)$ the index of the nearest neighbor of the current dataset in the 4C1NN learning rule. Now let $l \geq 1$. We focus on the time t_{k_l}. Consider the tree $\mathcal{G}$ where nodes are times $\mathcal{T} := \{t,\ t \leq t_{k_l},\ x_t \notin \{x_u, u < t\}\}$ for which a new input was visited, where the parent relations are given by $(t, \phi(t))$ for $t \in \mathcal{T} \setminus \{1\}$. In other words, we construct the tree in which a new input is linked to its representant which was used to derive the target prediction. Note that by definition of the 4C1NN learning rule, each node has at most 4 children and a node is not in the dataset at time t_{k_l} when it has exactly 4 children.

By symmetry, we will suppose without loss of generality that the majority of input points on which 4C1NN made a mistake belong to the first half $[0, \frac{1}{2}]$ i.e.

$$|\{t \leq t_{k_l},\ \ell_{01}(4C1NN(\boldsymbol{x}_{<t}, \boldsymbol{y}_{<t}, x_t), f^*(x_t)) = 1,\ x_t \in [0, 1/2]\}| \geq \frac{k_l}{2}$$

or equivalently, $\left|\left\{k \leq k_l,\ x_{t_k} \leq \frac{1}{2}\right\}\right| \geq \frac{k_l}{2}$.

Let us now consider the subgraph $\tilde{\mathcal{G}}$ given by restricting $\mathcal{G}$ only to nodes in the first half-space $[0, 1/2]$ which are mapped to the true value 1 i.e. on times $\{t \in \mathcal{T},\ x_t \leq \frac{1}{2}\}$. In this subgraph, the only times with no parent are times t_k with $k \leq k_l$ and $x_{t_k} \leq \frac{1}{2}$ and possibly time $t = 1$. Indeed, if a time in $\tilde{\mathcal{G}}$ has a parent $\phi(t)$ in $\tilde{\mathcal{G}}$, the prediction of 4C1NN for x_t returned the correct answer 1. The converse is also true except for the root time $t = 1$ which has no parent in $\mathcal{G}$. Therefore, $\tilde{\mathcal{G}}$ is a collection of disjoint trees with roots times $\{t_k,\ k \leq k_l,\ x_{t_k} \leq \frac{1}{2}\}$ (and possibly $t = 1$). For a given time t_k with $k \leq k_l$ and $x_{t_k} \leq \frac{1}{2}$, we will denote by $\mathcal{T}_k$ the corresponding tree in $\tilde{\mathcal{G}}$ with root t_k. We say that the $\mathcal{T}_k$ is a *good* tree if all times $t \in \mathcal{T}_k$ of this tree are parent in $\mathcal{G}$ to at most 1 time from the second half-space $(\frac{1}{2}, 1]$ i.e. if

$$\forall t \in \mathcal{T}_k, \quad \left|\left\{u \leq t_{k_l},\ \phi(u) = t,\ x_u > \frac{1}{2}\right\}\right| \leq 1.$$

We denote by $G = \{k \leq k_l,\ x_{t_k} \leq \frac{1}{2},\ \mathcal{T}_k \text{ good}\}$ the set of indices of good trees. By opposition, we will say that a tree is *bad* otherwise. We now give a simple upper bound on N_{bad} the number of

bad trees. Note that for any time $t \in \mathcal{T}_k$ of a tree, times in $\left\{u \leq t_{k_l},\ \phi(u) = t,\ x_u > \frac{1}{2}\right\}$ are when 4C1NN makes a mistake on the second-half $(\frac{1}{2}, 1]$. Therefore,

$$\sum_{k \leq k_l,\, x_{t_k} \leq \frac{1}{2}} \sum_{t \in \mathcal{T}_k} \left| \left\{ u < t_{k_l},\ \phi(u) = t,\ x_u > \frac{1}{2} \right\} \right| \leq \left| \left\{ k \leq k_l, x_{t_k} > \frac{1}{2} \right\} \right| \leq \frac{k_l}{2}$$

because by hypothesis $\left|\left\{k \leq k_l,\ x_{t_k} \leq \frac{1}{2}\right\}\right| \geq \frac{k_l}{2}$. Therefore, since each bad tree contains a node which is parent to at least 2 times of mistake in $(\frac{1}{2}, 1]$, we obtain

$$N_{\text{bad}} \leq \sum_{k \leq k_l,\, x_{t_k} \leq \frac{1}{2}} \sum_{t \in \mathcal{T}_k} \frac{1}{2} \left| \left\{ u < t_{k_l},\ \phi(u) = t,\ x_u > \frac{1}{2} \right\} \right| \leq \frac{k_l}{4}.$$

Thus, the number of good trees is $|G| = \left|\left\{k \leq k_l,\ x_{t_k} \leq \frac{1}{2}\right\}\right| - N_{\text{bad}} \geq \frac{k_l}{4}$. We now focus on good trees only and analyze their relation with the final dataset $\mathcal{D}_{t_{k_l}}$. Precisely, for a good tree $\mathcal{T}_k$, denote $\mathcal{V}_k = \mathcal{T}_k \cap \mathcal{D}_{t_{k_l}}$ the set of times which are present in the final dataset and belong to the tree induced by error time t_k. One can note that the sets $\{x_u,\ u \in \mathcal{V}_k\}_{k \in G}$ are totally ordered:

$$\forall k_1 < k_2 \in G,\ \forall t_1 \in \mathcal{T}_{k_1},\ \forall t_2 \in \mathcal{T}_{k_2}, \quad x_{t_1} < x_{t_2}.$$

This can be shown by observing that at each iteration t of 4C1NN, the following invariant is conserved: the sets $\{x_u,\ u \in \mathcal{T}_k \cap \mathcal{D}_t\}_{k \in \{l \in G,\, t_l \leq t\}}$ are totally ordered. The induction follows from the fact that when a new input point is visited, 4C1NN performs the 1NN learning rule on the current dataset $\mathcal{D}_l$. Therefore, either the sets $\{x_u,\ u \in \mathcal{T}_k \cap \mathcal{D}_t\}_{k \in \{l \in G,\, t_l \leq t\}}$ are conserved, or a new point is added when $t = t_k$ for some $k \leq k_l$ which forms its own tree and is closest to $(\frac{1}{2}, 1]$ than all other sets $\{x_u,\ u \in \mathcal{T}_k \cap \mathcal{D}_t\}_{k \in \{l \in G,\, t_l \leq t\}}$, or a new point is added to an existing tree $\mathcal{T}_k$ in which case it should be closer to some time of $\mathcal{T}_k \cap \mathcal{D}_t$ than any time in $\mathcal{T}_{k-1} \cap \mathcal{D}_t$ or $\mathcal{T}_{k+1} \cap \mathcal{D}_t$—if $\mathcal{T}_{k-1}$ or $\mathcal{T}_{k+1}$ exist. Additionally, a time may be removed which is still consistent with the invariant. Last, we observe that these sets never run empty because a time is removed only when at least 3 other points were added to the same set.

We now reason by induction to show that the sets $\{x_u,\ u \in \mathcal{V}_k\}_{k \in G}$ are also well separated—in a multiplicative way. Let us order the good trees by $G = \{g_1 < \ldots < g_{|G|}\}$ and start with tree $\mathcal{T}_{g_1}$. Consider any leaf of this tree and the corresponding path to the root $p_l \to p_{l-1} \to p_0 = t_{g_1}$ and define $x^1 = \min_{1 \leq i \leq l} x_{p_i}$. By construction, any point on this path is being replaced by its parent. Therefore, at any step of the algorithm 4C1NN at least one point on this path is available in the dataset $\mathcal{D}_t$ for any $t \geq t_{g_1}$—for instance the last time p_i such that $p_i \leq t$. This point x^1 provides a lower bound for the maximum point in $\{x_u,\ u \in \mathcal{T}_{g_1} \cap \mathcal{D}_t\}$ which in turn will provide a lower bound for all points in $\{x_u,\ u \in \mathcal{T}_{g_2} \cap \mathcal{D}_t\}$.

Let us now turn to $\mathcal{T}_{g_2}$. By construction, in a good tree $\mathcal{T}_k$, a time $t \in \mathcal{T}_k$ which is not in the final dataset $\mathcal{D}_{t_{k_l}}$ must be parent to at least 3 other times within $\mathcal{T}_k$. Therefore, until the minimal depth of an available time $\mathcal{V}_{g_2} = \mathcal{T}_{g_2} \cap \mathcal{D}_{t_{k_l}}$ in the current dataset $\mathcal{D}_{t_{k_l}}$, each node of the tree $\mathcal{T}_{g_2}$ has at least 3 parents which correspond necessarily to times $t > t_{g_2}$. Therefore, the minimal depth $d(g_2)$ of an available time $\mathcal{V}_k$ in the current dataset satisfies

$$\sum_{i=0}^{d(g_2)-1} 3^i \leq |\mathcal{T}_{g_2}| \leq t_{k_l}.$$

Therefore $d(g_2) \leq \log_3(2t_{k_l}+1) \leq \log_3 t_{k_l}$. Now consider the specific path from this node in $\mathcal{V}_{g_2}$ of minimal depth to the root t_{g_2}. Denote this path $p_{d(g_2)} \to p_{d(g_2)-1} \to p_0 = t_{g_2}$. Each arc of this path represents the fact that at the corresponding iteration p_i of 4C1NN, the parent $x_{p_{i-1}}$ was closer from x_{p_i} than any other point of the current dataset $\mathcal{D}_{p_i}$, in particular any point of $\{x_u,\ u \in \mathcal{T}_{g_1} \cap \mathcal{D}_{p_i}\}$. This gives $|x_{p_{i-1}} - x_{p_i}| \leq |x^1 - x_{p_{i-1}}| = x_{p_{i-1}} - x^1$ because we have $x_{p_{i-1}}, x_{p_i} > x^1$. Therefore we obtain

$$x_{p_{i-1}} \geq \frac{x^1 + x_{p_i}}{2}.$$

Indeed, if this were not the case we would have $|x_{p_{i-1}} - x_{p_i}| = x_{p_i} - x_{p_{i-1}} > x_{p_{i-1}} - x^1$. Similarly, considering the fact that 4C1NN makes a mistake at time t_{g_2}, the parent of t_{g_2} satisfies $x_{\phi(t_{g_2})} > \frac{1}{2}$ which yields $x_{t_{g_2}} \geq \frac{x^1 + x_{\phi(t_{g_2})}}{2} \geq \frac{x^1+\frac{1}{2}}{2}$. Hence, for any $0 \leq i \leq d(g_2)$,

$$x_{p_i} \geq x^1\left(1 - \frac{1}{2^i}\right) + \frac{x_{t_{g_2}}}{2^i} \geq x^1 + \frac{x_{t_{g_2}} - x^1}{2^{d(g_2)}} \geq x^1 + \left(\frac{1}{2} - x^1\right) t_{k_l}^{-\frac{\log 2}{\log 3}}.$$

Again, at every iteration $t \geq t_{g_2}$ of 4C1NN, at least one of the points x_{p_i} is available in the dataset $\mathcal{D}_t$—for instance the last x_{p_i} such that $p_i \leq t$. By total ordering, this $x^2 := \min_{0 \leq i \leq d(g_2)} x_{p_i}$ provides a lower bound for all points $\{x_u,\ u \in \mathcal{T}_{g_3} \cap \mathcal{D}_t\}$ whenever $t \geq t_{g_3}$. Hence, the lower bound x^2 acts as a new barrier: the equivalent of x^1 for the above argument with $\mathcal{T}_{g_2}$.

For clarity, we precise the next iteration of the induction for $\mathcal{T}_{g_3}$. The minimal depth $d(g_3)$ of an available time $\mathcal{V}_{g_3}$ satisfies $d(g_3) \leq \log_3(t_{k_l} - t_{g_3} + 1) + 1$ using the same argument as above. Now consider the corresponding path in $\mathcal{T}_{g_3}$ from this minimal depth node to the root $p_{d(g_3)} \to \ldots \to p_0 = t_{g_3}$. By definition of the 4C1NN learning rule, the parent $x_{p_{i-1}}$ was closer to x_{p_i} than any point of $\{x_u,\ u \in \mathcal{T}_{g_2} \cap \mathcal{D}_t\}$. By the previous step of the induction, we know that the maximum value of this set is at least x^2. Therefore, we obtain $|x_{p_{i-1}} - x_{p_i}| \leq |x^2 - x_{p_i}| = x_{p_i} - x^2$. We recall that we also have $x_{p_{i-1}} \geq x^2$ and $x_{p_i} \geq x^2$. The same argument as above gives $x_{p_i} \geq \frac{x^2 + x_{p_{i-1}}}{2}$. Further, we obtain similarly $x_{t_{g_3}} \geq \frac{x^2 + x_{\phi(t_{g_3})}}{2} \geq \frac{x^2+\frac{1}{2}}{2}$. Hence, for all $0 \leq i \leq d(g_3)$,

$$x_{p_i} \geq x^2 + \frac{x_{t_{g_3}} - x^2}{2^{d(g_3)}} \geq x^2 + \left(\frac{1}{2} - x^2\right) t_{k_l}^{-\frac{\log 2}{\log 3}}.$$

We denote $x^3 := \min_{0 \leq i \leq d(g_3)} x_{p_i}$, which now acts as a lower barrier for the tree $\mathcal{T}_{g_4}$ and we can apply the induction.

We complete this induction for $\mathcal{T}_{g_3}, \ldots, \mathcal{T}_{g_{|G|}}$. This creates a sequence of distinct visited input points $(x^i)_{1 \leq i \leq |G|}$ with $x^i \leq \frac{1}{2}$ such that for any $1 \leq i < |G|$, $x^{i+1} \geq x^i + \left(\frac{1}{2} - x^i\right) t_{k_l}^{-\frac{\log 2}{\log 3}}$ i.e.

$$\frac{1}{2} - x^{i+1} \leq \left(\frac{1}{2} - x^i\right)\left(1 - t_{k_l}^{-\frac{\log 2}{\log 3}}\right).$$

In particular, we can observe that $0 \leq x^1 < x^2 < \ldots < x^{|G|} \leq \frac{1}{2}$. Further, recalling that we have $t_{k_l} < \frac{2k_l}{\epsilon}$, we get

$$\log\left(\frac{1}{2} - x^{i+1}\right) - \log\left(\frac{1}{2} - x^i\right) \leq \log\left(1 - t_{k_l}^{-\frac{\log 2}{\log 3}}\right) \leq -t_{k_l}^{-\frac{\log 2}{\log 3}} \leq -\left(\frac{\epsilon}{2k_l}\right)^{\frac{\log 2}{\log 3}},$$

for any $1 \leq i \leq |G| - 1$. We will now argue that most of these points x^i fall in distinct sets of the type $[a_k, a_{k+1})$ where $a_k := \frac{1}{2} - \frac{1}{2k}$ for $k \geq 1$. We observe that for any $k \geq 1$, we have by concavity $\log\left(\frac{1}{2} - a_{k+1}\right) - \log\left(\frac{1}{2} - a_k\right) = \log\left(1 - \frac{1}{k+1}\right) \geq -\frac{\log 2}{k+1}$. Therefore, with $k^0 = \left\lceil \log 2 \cdot \left(\frac{2k_l}{\epsilon}\right)^{\frac{\log 2}{\log 3}} \right\rceil$, for any $k \geq k^0$ we have

$$\log\left(\frac{1}{2} - a_{k+1}\right) - \log\left(\frac{1}{2} - a_k\right) > -\left(\frac{\epsilon}{2k_l}\right)^{\frac{\log 2}{\log 3}}.$$

Therefore, for any $1 \leq i \leq |G| - 1$ such that $x^i > a_{k^0}$, x^i and x^{i+1} would lie in different sets of the type $[a_k, a_{k+1})$, $k \geq 1$. In fact because the sequence $(x^i)_{1 \leq i \leq |G|}$ is increasing, if $x^{i^*} > a_{k^0}$ then all points $(x^i)_{i^* \leq i \leq |G|}$ lie in distinct sets of the type $[a_k, a_{k+1})$, $k \geq 1$. Recall that $|G| \geq \frac{k_l}{4}$. Denote $i^* = \lfloor \frac{k_l}{8} \rfloor$. Because $(k_l)_{l \geq 1}$ is an increasing sequence, we have

$$\log\left(\frac{1}{2} - x^{i^*}\right) \leq \log\left(\frac{1}{2}\right) - (i^* - 1)\left(\frac{\epsilon}{2k_l}\right)^{\frac{\log 2}{\log 3}} \underset{l \to \infty}{\sim} -c_\epsilon k_l^{1 - \frac{\log 2}{\log 3}},$$

where $c_\epsilon := \frac{1}{8}\left(\frac{\epsilon}{2}\right)^{\frac{\log 2}{\log 3}}$ is a constant. Therefore,

$$\log\left(\frac{1}{2} - a_{k^0}\right) = -\log(2k^0) \underset{l \to \infty}{\sim} -\frac{\log 2}{\log 3} \log k_l = o\left(\log\left(\frac{1}{2} - x^{i^*}\right)\right)$$

which shows that for some constant l^0 and any $l \geq l^0$ we have $a_{k^0} < x^{i^*} < \frac{1}{2}$. Hence, for any $l \geq l^0$, all the points $(x^i)_{i^* \leq i \leq |G|}$ lie in distinct sets of the partition and there are at least $|G| - \frac{k_l}{8} \geq \frac{k_l}{8}$ such points. Therefore, for any $l \geq l^0$,

$$|\{P \in \mathcal{P}, \quad P \cap \boldsymbol{x}_{\leq t_{k_l}} \neq \emptyset\}| \geq \frac{k_l}{8} \geq \frac{\epsilon}{16} t_{k_l}.$$

Because $t_{k_l} \to \infty$ as $l \to \infty$, this shows that $|\{P \in \mathcal{P}, \quad P \cap \boldsymbol{x}_{<T} \neq \emptyset\}| \neq o(T)$. Because this holds for any realization of the event $\mathcal{A}$, we obtained

$$\mathbb{P}(|\{P \in \mathcal{P}, \quad P \cap \mathbb{X}_{<T} \neq \emptyset\}| = o(T)) \leq \mathbb{P}(\mathcal{A}^c) = 1 - \delta < 1.$$

This shows that $\mathbb{X} \notin \text{SMV}_{([0,1],|\cdot|)}$ and ends the proof of the proposition. ■

Note that using the same proof, we observe that the result from Proposition 16 holds for all learning rules kC1NN with $k \geq 4$.

We are now ready to prove that 4C1NN is universally consistent under processes of $\text{SMV}_{([0,1],|\cdot|)}$ for the binary classification setting. Intuitively, we analyze the set of functions on which 4C1NN is consistent under a fixed process $\mathbb{X} \in \text{SMV}_{([0,1],|\cdot|)}$ and show that this is a σ-algebra. Proposition 16 will be useful to show that this σ-algebra contains all intervals and as a result is the complete Borel σ-algebra $\mathcal{B}$ i.e. 4C1NN is universally consistent under $\mathbb{X}$.

Theorem 17 *Let $\mathcal{X} = [0, 1]$ with the usual topology $\mathcal{B}$. For the binary classification setting, the learning rule 4C1NN is universally consistent for all processes $\mathbb{X} \in SMV_{([0,1],|\cdot|)}$.*

Proof let $\mathbb{X} \in \text{SMV}_{([0,1],|\cdot|)}$. We will show that 4C1NN is universally consistent on $\mathbb{X}$ by considering the set $\mathcal{S}_{\mathbb{X}}$ of functions for which it is consistent. More precisely, since $\mathcal{Y} = \{0, 1\}$ in the binary setting, all target functions can be described as $f = \mathbb{1}_{A_{f^*}}$ where $A_{f^*} = f^{<-1>}(\{1\})$ is a measurable set. In the following, we will refer interchangeably to the function f^* or the set A_{f^*}, and define $\mathcal{S}_{\mathbb{X}}$ using the corresponding sets:

$$\mathcal{S}_{\mathbb{X}} := \{A \in \mathcal{B}, \quad \mathcal{L}_{\mathbb{X}}(4C1NN, \mathbb{1}_A) = 0 \quad (a.s.)\}$$

By construction we have $\mathcal{S}_{\mathbb{X}} \subset \mathcal{B}$. The goal is to show that in fact $\mathcal{S}_{\mathbb{X}} = \mathcal{B}$. To do so, we will show that $\mathcal{S}$ satisfies the following properties

- $\emptyset \in \mathcal{S}_{\mathbb{X}}$ and $\mathcal{S}_{\mathbb{X}}$ contains all intervals $[0, s)$ with $0 < s \leq 1$,
- if $A \in \mathcal{S}_{\mathbb{X}}$ then $A^c \in \mathcal{S}_{\mathbb{X}}$ (stable to complementary),
- if $(A_i)_{i\geq 1}$ is a sequence of disjoint sets of $\mathcal{S}_{\mathbb{X}}$, then $\bigcup_{i\geq 1} A_i \in \mathcal{S}_{\mathbb{X}}$ (stable to $\sigma-$additivity for disjoint sets),
- if $A, B \in \mathcal{S}_{\mathbb{X}}$, then $A \cup B \in \mathcal{S}_{\mathbb{X}}$ (stable to union).

Together, these properties show that $\mathcal{S}_{\mathbb{X}}$ is a $\sigma-$algebra that contains all open intervals of $\mathcal{X} = [0, 1]$. Recall that by definition, $\mathcal{B}$ is the smallest $\sigma-$algebra containing open intervals. Therefore we get $\mathcal{B} \subset \mathcal{S}_{\mathbb{X}}$ which proves the theorem. We now show the four properties.

We start by showing the invariance to complementary. Note that 4C1NN is invariant to labels and that the loss ℓ_{01} is symmetric. Therefore, if it achieves consistency for $\mathbb{1}_A$ it also achieves consistency for $\mathbb{1}_{A^c}$. Indeed, at each step, 4C1NN will use the same representant for the prediction hence for any $t \geq 0$,

$$\ell_{01}(4C1NN(\boldsymbol{x}_{<t}, \mathbb{1}_{\boldsymbol{x}_{<t}\in A}, x_t), \mathbb{1}_{x_t\in A}) = \ell_{01}(4C1NN(\boldsymbol{x}_{<t}, \mathbb{1}_{\boldsymbol{x}_{<t}\in A^c}, x_t), \mathbb{1}_{x_t\in A^c}).$$

4C1NN is clearly consistent for $f^* = 0$. Therefore $\emptyset \in \mathcal{S}_{\mathbb{X}}$. Now let $0 < s \leq 1$. We will show that $[0, s) \in \mathcal{S}_{\mathbb{X}}$. Proposition 16 shows that $[0, \frac{1}{2}] \in \mathcal{S}_{\mathbb{X}}$. In fact, one can note that the same proof shows that $[0, \frac{1}{2}) \in \mathcal{S}_{\mathbb{X}}$. Further, for any $0 < s \leq 1$ using the same proof with the following partition centered in s,

$$\{s\} \cup \bigcup_{k\geq 1} \left[s\left(1 - \frac{1}{k}\right); s\left(1 - \frac{1}{k+1}\right)\right) \cup \bigcup_{k\geq 1} \left(s + \frac{1-s}{k+1}; s + \frac{1-s}{k}\right]$$

shows that $[0, s], [0, s) \in \mathcal{S}_{\mathbb{X}}$.

We now turn to the $\sigma-$additivity for disjoint sets. Let $(A_i)_{i\geq 1}$ is a sequence of disjoint sets of $\mathcal{S}_{\mathbb{X}}$. We denote $A := \bigcup_{i\geq 1} A_i$. We consider the target function $f^* = \mathbb{1}_A$. There are two types of statistical errors: errors of type 1 correspond to $X_t \in A$ and a predicted value 0 while type 2 errors correspond to $X_t \notin A$ and a predicted value 1. We then write the average loss in the following way,

$$\frac{1}{T}\sum_{t=1}^{T} \ell_{01}(4C1NN(\mathbb{X}_{<t}, \mathbb{Y}_{<t}, X_t), f^*(X_t)) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{X_t\in A}\mathbb{1}_{X_{\phi(t)}\notin A} + \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{X_t\notin A}\mathbb{1}_{X_{\phi(t)}\in A},$$

where the first term corresponds to type 1 errors and the second term corresponds to type 2 errors.

We suppose by contradiction that $\mathbb{P}(\mathcal{L}_{\mathbb{X}}(4C1NN, f^*) > 0) := \delta > 0$ Therefore, there exists $\epsilon > 0$ such that $\mathbb{P}(\mathcal{L}_{\mathbb{X}}(4C1NN, f^*) > \epsilon) \geq \frac{\delta}{2}$. We denote this event by $\mathcal{A} := \{\mathcal{L}_{\mathbb{X}}(4C1NN, f^*) > \epsilon\}$. We first analyze the errors induced by one set A_i only. We have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}(\mathbb{1}_{X_t\in A_i}\mathbb{1}_{X_{\phi(t)}\notin A} + \mathbb{1}_{X_t\notin A}\mathbb{1}_{X_{\phi(t)}\in A_i}) &\leq \frac{1}{T}\sum_{t=1}^{T}(\mathbb{1}_{X_t\in A_i}\mathbb{1}_{X_{\phi(t)}\notin A_i} + \mathbb{1}_{X_t\notin A_i}\mathbb{1}_{X_{\phi(t)}\in A_i}) \\ &= \frac{1}{T}\sum_{t=1}^{T}\ell_{01}(4C1NN(\mathbb{X}_{<t}, \mathbb{1}_{\mathbb{X}_{<t}\in A_i}, X_t), \mathbb{1}_{X_t\in A_i}).\end{aligned}$$

Then, because 4C1NN is consistent for $\mathbb{1}_{A_i}$, we have

$$\frac{1}{T}\sum_{t=1}^{T}(\mathbb{1}_{X_t\in A_i}\mathbb{1}_{X_{\phi(t)}\notin A} + \mathbb{1}_{X_t\notin A}\mathbb{1}_{X_{\phi(t)}\in A_i}) \to 0 \quad (a.s.).$$

We take $\epsilon_i = \frac{\epsilon}{4\cdot 2^i}$ and $\delta_i = \frac{\delta}{8\cdot 2^i}$. The above equation gives

$$\mathbb{P}\left[\bigcup_{t_0\geq 1}\bigcap_{T\geq t_0}\left\{\frac{1}{T}\sum_{t=1}^{T}(\mathbb{1}_{X_t\in A_i}\mathbb{1}_{X_{\phi(t)}\notin A} + \mathbb{1}_{X_t\notin A}\mathbb{1}_{X_{\phi(t)}\in A_i}) < \epsilon_i\right\}\right] = 1.$$

Therefore, let T^i such that

$$\mathbb{P}\left[\bigcap_{T\geq T^i}\left\{\frac{1}{T}\sum_{t=1}^{T}(\mathbb{1}_{X_t\in A_i}\mathbb{1}_{X_{\phi(t)}\notin A} + \mathbb{1}_{X_t\notin A}\mathbb{1}_{X_{\phi(t)}\in A_i}) < \epsilon_i\right\}\right] \geq 1 - \delta_i.$$

We will denote by $\mathcal{E}_i$ this event. We now consider the scale of the process $\mathbb{X}_{\leq T^i}$ when falling in A_i, by introducing $\eta_i > 0$ such that

$$\mathbb{P}\left[\min_{\substack{t_1,t_2\leq T^i;\, X_{t_1},X_{t_2}\in A_i;\\ X_{t_1}\neq X_{t_2}}} |X_{t_1} - X_{t_2}| > \eta_i\right] \geq 1 - \delta_i.$$

We denote by $\mathcal{F}_i$ this event. By the union bound, we have $\mathbb{P}(\bigcup_{i\geq 1}\mathcal{E}_i^c \cup \bigcup_{i\geq 1}\mathcal{F}_i^c) \leq \frac{\delta}{4}$. Therefore, we obtain $\mathbb{P}(\mathcal{A}\cap\bigcap_{i\geq 1}\mathcal{E}_i\cap\bigcap_{i\geq 1}\mathcal{F}_i) \geq \mathbb{P}(\mathcal{A}) - \mathbb{P}(\bigcup_{i\geq 1}\mathcal{E}_i^c \cup \bigcup_{i\geq 1}\mathcal{F}_i^c) \geq \frac{\delta}{4}$. We now construct a partition $\mathcal{P}$ obtained by subdividing each set A_i according to scale η_i. For simplicity, we use the notation $N_i = \lfloor\frac{1}{\eta_i}\rfloor$ and construct the partition given of $\mathcal{X} = [0,1]$ given by

$$\mathcal{P} \;:\quad A^c \cup \bigcup_{i\geq 1}\left\{([N_i\eta_i, 1]\cap A_i) \cup \bigcup_{j=0}^{N_i-1}([j\eta_i, (j+1)\eta_i)\cap A_i)\right\}.$$

Let us now consider a realization of $\boldsymbol{x}$ of $\mathbb{X}$ in the event $\mathcal{A}\cap\bigcap_{i\geq 1}\mathcal{E}_i\cap\bigcap_{i\geq 1}\mathcal{F}_i$. The sequence $\boldsymbol{x}$ is now not random anymore. Our goal is to show that $\boldsymbol{x}$ does not visit a sublinear number of sets in the partition $\mathcal{P}$.

By construction, the event $\mathcal{A}$ is satisfied, therefore there exists an increasing sequence of times $(t_k)_{k\geq 1}$ such that for any $k \geq 1$, $\frac{1}{t_k}\sum_{t=1}^{t_k} \ell_{01}(4C1NN(\boldsymbol{x}_{<t}, \mathbb{1}_{\boldsymbol{x}_{<t}\in A}, x_t), \mathbb{1}_{x_t\in A}) > \frac{\epsilon}{2}$. Therefore, we obtain for any $k \geq 1$,

$$\sum_{i\geq 1} \frac{1}{t_k}\sum_{t=1}^{t_k} (\mathbb{1}_{x_t\in A_i}\mathbb{1}_{x_{\phi(t)}\notin A} + \mathbb{1}_{x_t\notin A}\mathbb{1}_{x_{\phi(t)}\in A_i}) > \frac{\epsilon}{2}.$$

Also, because the events $\mathcal{E}_i$ are met, we have

$$\sum_{i\geq 1;\ t_k\geq T^i} \frac{1}{t_k}\sum_{t=1}^{t_k} (\mathbb{1}_{x_t\in A_i}\mathbb{1}_{x_{\phi(t)}\notin A} + \mathbb{1}_{x_t\notin A}\mathbb{1}_{x_{\phi(t)}\in A_i}) < \sum_{i\geq 1, t_k\geq T^i} \epsilon_i \leq \frac{\epsilon}{4}.$$

Combining the two above equations gives

$$\frac{1}{t_k}\sum_{t=1}^{t_k} \sum_{i\geq 1;\ t_k< T^i} (\mathbb{1}_{x_t\in A_i}\mathbb{1}_{x_{\phi(t)}\notin A} + \mathbb{1}_{x_t\notin A}\mathbb{1}_{x_{\phi(t)}\in A_i}) > \frac{\epsilon}{4}. \tag{1}$$

We now consider the set of times such that an input point fell into the set A_i with $T^i > t_k$, either creating a mistake in the prediction of 4C1NN or inducing a later mistake within time horizon t_k: $\mathcal{T} := \bigcup_{i\geq 1;\ T^i > t_k} \mathcal{T}_i$ where

$$\mathcal{T}_i := \left\{ t \leq t_k,\ x_t \in A_i,\ \left(x_{\phi(t)} \notin A \text{ or } \exists t < u \leq t_k \text{ s.t. } \phi(u) = t,\ x_u \notin A\right)\right\}.$$

We now show that all points x_t for $t \in \mathcal{T}$ fall in distinct sets of the partition $\mathcal{P}$. Indeed, because the sets A_i are disjoint, it suffices to check that for any $i \geq 1$ such that $T^i > t_k$, the points x_t for $t \in \mathcal{T}_i$ fall in distinct of the following sets

$$[N_i\eta_i, 1] \cap A_i, \quad [j\eta_i, (j+1)\eta_i) \cap A_i, \quad 0 \leq j \leq N_i - 1.$$

Note that for any $t_1 < t_2 \in \mathcal{T}_i$ we have $x_{t_1}, x_{t_2} \in A_i$ and $x_{t_1} \neq x_{t_2}$. Indeed, we cannot have $x_{t_2} = x_{t_1}$ otherwise 4C1NN would make no mistake at time t_2 and x_{t_2} would induce no future mistake either (recall that if an input point was already visited, we use simple memorization for the prediction and do not add it to the dataset). Therefore, because the event $\mathcal{F}_i$ is satisfied, for any $t_1 < t_2 \in \mathcal{T}_i$ we have $|x_{t_1} - x_{t_2}| > \eta_i$. Hence x_{t_1} and x_{t_2} lie in different sets among $[N_i\eta_i, 1] \cap A_i$ or $[j\eta_i, (j+1)\eta_i) \cap A_i$ for $0 \leq j \leq N_i - 1$. This shows that all points $\{x_t,\ t \in \mathcal{T}\}$ lie in different sets of the partition $\mathcal{P}$. Therefore,

$$|\{P \in \mathcal{P}, P \cap \boldsymbol{x}_{\leq t_k} \neq \emptyset\}| \geq |\mathcal{T}|.$$

We now lower bound $|\mathcal{T}|$, which will uncover the main interest of the learning rule 4C1NN. Intuitively, this learning rule prohibits a single input point x_t to induce a large number of mistakes in the learning process. Indeed, any input point incurs at most $1 + 4 = 5$ mistakes while this number of mistakes incurred by a single point can potentially by unbounded for the traditional 1NN learning rule. We now formalize this intuition.

$$\sum_{t=1}^{t_k} \sum_{i\geq 1;\ t_k<T^i} (\mathbb{1}_{x_t\in A_i}\mathbb{1}_{x_{\phi(t)}\notin A} + \mathbb{1}_{x_t\notin A}\mathbb{1}_{x_{\phi(t)}\in A_i})$$

$$= \sum_{t=1}^{t_k} \sum_{i\geq 1;\ t_k<T^i} \left(\mathbb{1}_{x_t\in A_i}\mathbb{1}_{x_{\phi(t)}\notin A} + \sum_{t<u\leq t_k} \mathbb{1}_{x_u\notin A}\mathbb{1}_{x_t\in A_i}\mathbb{1}_{\phi(u)=t} \right)$$

$$= \sum_{i\geq 1;\ T^i>t_k} \sum_{t\leq t_k,\ x_t\in A_i} \left(\mathbb{1}_{x_{\phi(t)}\notin A} + \sum_{t<u\leq t_k} \mathbb{1}_{x_u\notin A}\mathbb{1}_{\phi(u)=t} \right)$$

$$\leq \sum_{i\geq 1;\ T^i>t_k} \sum_{t\leq t_k,\ x_t\in A_i} 5\max\left(\mathbb{1}_{x_{\phi(t)}\notin A}, \mathbb{1}_{x_u\notin A}\mathbb{1}_{\phi(u)=t},\ t<u\leq t_k\right)$$

$$= 5|\mathcal{T}|$$

where in the last inequality we used the fact that a given time t can have at most 4 children i.e. $|\{u>t, \phi(u)=t\}|\leq 4$ with the 4C1NN learning rule. We now use Equation (1) to obtain

$$|\{P\in\mathcal{P}, P\cap \boldsymbol{x}_{\leq t_k}\neq\emptyset\}| \geq |\mathcal{T}| \geq \frac{\epsilon}{20}t_k.$$

This holds for any $k\geq 1$. Therefore, because $t_k\to\infty$ as $k\to\infty$ we get $|\{P\in\mathcal{P}, P\cap \boldsymbol{x}_{\leq T}\neq\emptyset\}|\neq o(T)$. Finally, this holds for any realization of $\mathbb{X}$ in the event $\mathcal{A}\cap\bigcap_{i\geq 1}\mathcal{E}_i\cap\bigcap_{i\geq 1}\mathcal{F}_i$. Therefore,

$$\mathbb{P}(|\{P\in\mathcal{P}, P\cap \boldsymbol{x}_{\leq T}\neq\emptyset\}| = o(T)) \leq \mathbb{P}\left[\left(\mathcal{A}\cap\bigcap_{i\geq 1}\mathcal{E}_i\cap\bigcap_{i\geq 1}\mathcal{F}_i\right)^c\right] \leq 1-\frac{\delta}{4}<1.$$

Therefore, $\mathbb{X}\notin \text{SMV}_{([0,1],|\cdot|)}$ which contradicts the hypothesis. This concludes the proof that

$$\mathcal{L}_{\mathbb{X}}(4C1NN, \mathbb{1}_{\cdot\in A}) = 0 \quad (a.s.),$$

and hence, $\mathcal{S}_{\mathbb{X}}$ satisfies the $\sigma-$additivity property for disjoint sets.

Note that the choice of disjoint sets for the proof of $\sigma-$additivity was made for convenience so that the partition defined is not too complex. However to complete the proof of the $\sigma-$additivity of $\mathcal{S}_{\mathbb{X}}$, we have to prove that we can take unions of sets. Let $A_1, A_2\in\mathcal{S}_{\mathbb{X}}$. We consider $A=A_1\cup A_2$ and $f^*(\cdot)=\mathbb{1}_{\cdot\in A}$. Using the same arguments as above, we still have for $T\geq 1$,

$$\frac{1}{T}\sum_{t=1}^{T}(\mathbb{1}_{X_t\in A_i}\mathbb{1}_{X_{\phi(t)}\notin A} + \mathbb{1}_{X_t\notin A}\mathbb{1}_{X_{\phi(t)}\in A_i}) \to 0 \quad (a.s.).$$

for $i \in \{1,2\}$. But note that for any $T \geq 1$,

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\ell_{01}(4C1NN(\mathbb{X}_{<t},\mathbb{Y}_{<t},X_t),f^*(X_t)) \\
&\qquad = \frac{1}{T}\sum_{t=1}^{T}\mathbb{1}_{X_t\in A}\mathbb{1}_{X_{\phi(t)}\notin A} + \frac{1}{T}\sum_{t=1}^{T}\mathbb{1}_{X_t\notin A}\mathbb{1}_{X_{\phi(t)}\in A} \\
&\qquad \leq \frac{1}{T}\sum_{t=1}^{T}(\mathbb{1}_{X_t\in A_1}+\mathbb{1}_{X_t\in A_2})\mathbb{1}_{X_{\phi(t)}\notin A} + \frac{1}{T}\sum_{t=1}^{T}\mathbb{1}_{X_t\notin A}(\mathbb{1}_{X_{\phi(t)}\in A_1}+\mathbb{1}_{X_{\phi(t)}\in A_2}) \\
&\qquad = \sum_{i=1}^{2}\frac{1}{T}\sum_{t=1}^{T}(\mathbb{1}_{X_t\in A_i}\mathbb{1}_{X_{\phi(t)}\notin A} + \mathbb{1}_{X_t\notin A}\mathbb{1}_{X_{\phi(t)}\in A_i}).
\end{aligned}$$

Therefore we obtain directly $\mathcal{L}_{\mathbb{X}}(4C1NN, \mathbb{1}_{\cdot\in A}) = 0 \quad (a.s.)$. This shows that $A_1 \cup A_2 \in \mathcal{S}_{\mathbb{X}}$ and ends the proof of the theorem. ■

As an immediate consequence of Theorem 17 and Proposition 1, we obtain the following results.

Theorem 18 *$SUOL_{([0,1],|\cdot|),(\{0,1\},\ell_{01})} = SMV_{([0,1],|\cdot|)}$.*

Theorem 19 *For $\mathcal{X} = [0,1]$ with usual measure, and for binary classification, 4C1NN is an optimistically universal learning rule.*

B.2. Generalization to standard Borel input spaces and separable bounded output spaces.

The specific choices of input space $\mathcal{X} = [0,1]$ and binary classification for output setting are in fact not very restrictive. Indeed, any standard Borel input space $\mathcal{X}$ can be reduced to either $[0,1]$ or a countable set through the Kuratowski theorem. We recall that two standard Borel spaces i.e. complete separable Borel spaces, are Borel isomorphic if there exists a measurable bijection between them.

Theorem 20 (Kuratowski's theorem) *Any standard Borel space $\mathcal{X}$ is Borel isomorphic to one of (1) $\mathbb{R}$, (2) $\mathbb{N}$ or (3) a finite space.*

This classical result can be found for example in [22] (Section 15.B). Further, any bounded output setting $(\mathcal{Y},\ell)$ can be reduced to binary classification using Theorem 8 [2]. Using these two reductions, we can generalize Theorem 18 and Theorem 19 to any standard Borel space $\mathcal{X}$ and any separable bounded setting $(\mathcal{Y},\ell)$.

Corollary 21 *For any standard Borel space $\mathcal{X}$ and any separable near-metric output space $(\mathcal{Y},\ell)$ with $0 < \bar{\ell} < \infty$, we have $SUOL_{(\mathcal{X},\rho),(\mathcal{Y},\ell)} = SMV_{(\mathcal{X},\rho)}$.*

Corollary 22 *For any standard Borel space $\mathcal{X}$, and any separable near-metric output space $(\mathcal{Y},\ell)$ with bounded loss, there exists an optimistically universal learning rule.*

Proof of Corollary 21 and 22 Using Theorem 8 directly gives the result for $\mathcal{X} = [0,1]$ and any bounded separable near-metric output space. The results are already known when $\mathcal{X}$ is countable and in these cases, memorization is an optimistically universal learning rule [18]. We now fix a bounded separable ouptput setting $(\mathcal{Y}, \ell)$ and a standard Borel space $\mathcal{X}$, Borel isomorphic to $\mathbb{R}$ and as a result Borel isomorphic to $[0,1]$. Let $g : \mathcal{X} \to [0,1]$ be a measurable bijection and a process $\mathbb{X} \in \text{SMV}_{(\mathcal{X},\rho)}$. Note that the process $g(\mathbb{X}) := (g(X_t))_{t\geq 1}$ belongs to $\text{SMV}_{([0,1],|\cdot|)}$ by bi-measurability of g. We can construct the learning rule $f.$ for value setting $\mathcal{X}$ and output setting $(\mathcal{Y}, \ell)$ such that for any $\boldsymbol{x}_{\leq t} \in \mathcal{X}^t$ and $\boldsymbol{y}_{<t} \in \mathcal{Y}^{t-1}$ we define $f_t(x_{<t}, y_{<t}, x_t) = 4C1NN_t(g(x_{<t}), y_{<t}, g(x_t))$. By construction, for target function $f^* : \mathcal{X} \to \mathcal{Y}$ this learning rule under $\mathbb{X}$ has same losses as 4C1NN under $g(\mathbb{X})$ for the target function $f^* \circ g^{-1}$. Therefore, $f.$ is universally consistent under $\mathbb{X}$ which yields $\text{SMV}_{(\mathcal{X},\rho)} \subset \text{SUOL}_{(\mathcal{X},\rho),(\mathcal{Y},\ell)}$. Using Proposition 1 we have $\text{SUOL}_{(\mathcal{X},\rho),(\mathcal{Y},\ell)} = \text{SMV}_{(\mathcal{X},\rho)}$. We can also end the proof of Corollary 22 by noting that $f.$ is an optimistically universal learning rule. ■

Although quite intuitive and direct, this generalization has two limitations. First, it only applies to standard Borel spaces instead of general separable Borel spaces. Second, it does not provide a practical optimistically universal rule in general. Indeed, the constructed optimistically universal learning rule in Corollary 22 uses a bimeasurable bijection between $\mathcal{X}$ and $[0,1]$—in the non-trivial case where $\mathcal{X}$ is Borel isomorphic to $\mathbb{R}$—which can be very complex and non-intuitive. For instance, the constructed learning rule for $[0,1]^2$ is not 4C1NN but instead a complex learning rule using a measurable bijection $[0,1] \to [0,1]^2$. In the next section we solve these two issues by showing that 2C1NN is optimistically universal in the general case.

Appendix C. Proofs of Section 5: An optimistically universal learning rule

C.1. Proof of Lemma 6

Define

$$\begin{aligned} v(j) &:= \max\{0 \leq i \leq d,\ p_i < q_j\}, \quad j = 0, \dots, f. \\ u(i) &:= \max\{0 \leq j \leq f,\ q_j < p_i\}, \quad i = v(0)+1, \dots, d, \end{aligned}$$

Now observe that for any $v(0) + 1 \leq i \leq d$, we have $q_{u(i)} \in \mathcal{D}_{p_i}$ i.e. the datapoint $q_{u(i)}$ is available in the current dataset. Indeed, it is possibly removed after all of its children have been revealed, in particular $q_{u(i)+1}$ if it exists. By definition of $u(i)$, even if $q_{u(i)+1}$ exists, it has not yet been revealed since $p_i < q_{u(i)+1}$. Therefore, we have $\rho(x_{p_i}, x_{p_{i-1}}) \leq \rho(x_{p_i}, x_{q_{u(i)}})$. Similarly, we have for all $1 \leq j \leq f$, $\rho(x_{q_j}, x_{q_{j-1}}) \leq \rho(x_{q_j}, x_{p_{v(j)}})$. We now take $v_0 + 1 \leq i < d$. We have $q_{u(i)} < p_{v(u(i))+1} < \dots < p_i < q_{u(i)+1} < \dots < q_{u(i+1)} < p_{i+1}$ (where some terms might not exist). Therefore,

$$\begin{aligned} \rho(x_{p_i}, x_{q_{u(i)}}) &\leq \rho(x_{p_i}, x_{p_{i+1}}) + \rho(x_{p_{i+1}}, x_{q_{u(i+1)}}) + \rho(x_{q_{u(i)}}, x_{q_{u(i+1)}}) \\ &\leq 2\rho(x_{p_{i+1}}, x_{q_{u(i+1)}}) + \sum_{w=u(i)}^{u(i+1)-1} \rho(x_{q_w}, x_{q_{w+1}}) \\ &\leq 2\rho(x_{p_{i+1}}, x_{q_{u(i+1)}}) + \sum_{w=u(i)}^{u(i+1)-1} \rho(x_{p_i}, x_{q_{w+1}}) \end{aligned}$$

where in the last inequality, we used the fact that for all $u(i) \leq w \leq u(i+1)-1$, we have $v(w+1) = i$. Now observe that for any $u(i)+1 \leq w \leq u(i+1)-1$,

$$\rho(x_{p_i}, x_{q_w}) \leq \rho(x_{p_i}, x_{q_{w+1}}) + \rho(x_{q_w}, x_{q_{w+1}}) \leq 2\rho(x_{p_i}, x_{q_{w+1}}).$$

Therefore we have by induction $\rho(x_{p_i}, x_{q_w}) \leq 2^{u(i+1)-w}\rho(x_{p_i}, x_{q_{u(i+1)}})$. which yields

$$\rho(x_{p_i}, x_{q_{u(i)}}) \leq 2\rho(x_{p_{i+1}}, x_{q_{u(i+1)}}) + (2^{u(i+1)-u(i)} - 1)\rho(x_{p_i}, x_{q_{u(i+1)}}).$$

Finally, we observe that $\rho(x_{p_i}, x_{q_{u(i+1)}}) \leq \rho(x_{p_i}, x_{p_{i+1}}) + \rho(x_{p_{i+1}}, x_{q_{u(i+1)}}) \leq 2\rho(x_{p_{i+1}}, x_{q_{u(i+1)}})$. Hence,

$$\rho(x_{p_i}, x_{q_{u(i)}}) \leq 2^{u(i+1)-u(i)+1}\rho(x_{p_{i+1}}, x_{q_{u(i+1)}}).$$

By recursion, this yields

$$\rho(x_{p_{v(0)+1}}, x_{q_{u(v(0)+1)}}) \leq 2^{u(d)-u(v(0)+1)+(d-v(0)-1)}\rho(x_{p_d}, x_{q_{u(d)}}).$$

We now relate the quantity $\rho(x_{p_{v(0)+1}}, x_{q_{u(v(0)+1)}})$ (resp. $\rho(x_{p_d}, x_{q_{u(d)}})$) to $\rho(x_{p_{v(0)}}, x_{q_0})$ (resp. $\rho(x_{p_d}, x_{q_d})$). We have by construction $p_{v(0)} < q_0 < q_1 < \ldots < q_{u(v(0)+1)} < p_{v(0)+1}$. Therefore, similarly to before,

$$\begin{aligned}\rho(x_{p_{v(0)}}, x_{q_0}) &\leq \rho(x_{p_{v(0)}}, x_{p_{v(0)+1}}) + \rho(x_{p_{v(0)+1}}, x_{q_{u(v(0)+1)}}) + \sum_{w=0}^{u(v(0)+1)-1} \rho(x_{q_w}, x_{q_{w+1}}) \\ &\leq 2\rho(x_{p_{v(0)+1}}, x_{q_{u(v(0)+1)}}) + \sum_{w=1}^{u(v(0)+1)} \rho(x_{p_{v(0)}}, x_{q_w}).\end{aligned}$$

But $\rho(x_{p_{v(0)}}, x_{q_w}) \leq \rho(x_{p_{v(0)}}, x_{q_{w+1}}) + \rho(x_{q_w}, x_{q_{w+1}}) \leq 2\rho(x_{p_{v(0)}}, x_{q_{w+1}})$. Hence $\rho(x_{p_{v(0)}}, x_{q_w}) \leq 2^{u(v(0)+1)-w}\rho(x_{p_{v(0)}}, x_{q_{u(v(0)+1)}}) \leq 2^{u(v(0)+1)-w+1}\rho(x_{p_{v(0)+1}}, x_{q_{u(v(0)+1)}})$. Then,

$$\rho(x_{p_{v(0)}}, x_{q_0}) \leq 2^{u(v(0)+1)+1}\rho(x_{p_{v(0)+1}}, x_{q_{u(v(0)+1)}}) \leq 2^{u(d)+(d-v(0))}\rho(x_{p_d}, x_{q_{u(d)}}).$$

Finally, we have $q_{u(d)} < p_d < q_{u(d)+1} < \ldots < q_f$. Then,

$$\begin{aligned}\rho(x_{p_d}, x_{q_{u(d)}}) &\leq \sum_{w=u(d)}^{f-1} \rho(x_{q_w}, x_{q_{w+1}}) + \rho(x_{p_d}, x_{q_f}) \\ &\leq \sum_{w=u(d)}^{f-1} \rho(x_{p_d}, x_{q_{w+1}}) + \rho(x_{p_d}, x_{q_f}).\end{aligned}$$

Again, note that for $u(d)+1 \leq w \leq f-2$, we have $\rho(x_{p_d}, x_{q_w}) \leq \rho(x_{p_d}, x_{q_{w+1}}) + \rho(x_{q_w}, x_{q_{w+1}}) \leq 2\rho(x_{p_d}, x_{q_{w+1}})$. Hence, $\rho(x_{p_d}, x_{q_w}) \leq 2^{f-w}\rho(x_{p_d}, x_{q_f})$ and we obtain

$$\rho(x_{p_d}, x_{q_{u(d)}}) \leq (2^{f-u(d)} - 1)\rho(x_{p_d}, x_{q_f}) + \rho(x_{p_d}, x_{q_f}) = 2^{f-u(d)}\rho(x_{p_d}, x_{q_f}).$$

Putting everything together yields

$$\rho(x_{p_{v(0)}}, x_{q_0}) \leq 2^{f+d}\rho(x_{p_d}, x_{q_f}).$$

Finally, we compute

$$\begin{aligned}
\rho(x_{p_{v(0)}}, x_{p_d}) &\leq \sum_{i=v(0)+1}^{d} \rho(x_{p_{i-1}}, x_{p_i}) \\
&\leq \sum_{i=v(0)+1}^{d} \rho(x_{p_i}, x_{q_{u(i)}}) \\
&\leq \sum_{i=v(0)+1}^{d} 2^{u(d)-u(i)+d-i} \rho(x_{p_d}, x_{q_{u(d)}}) \\
&\leq \sum_{i=v(0)+1}^{d} 2^{u(d)-u(v(0)+1)+d-i} \rho(x_{p_d}, x_{q_{u(d)}}) \\
&\leq 2^{u(d)-u(v(0)+1)+d-v(0)} \rho(x_{p_d}, x_{q_{u(d)}}) \\
&\leq 2^{f-u(v(0)+1)+d-v(0)} \rho(x_{p_d}, x_{q_f}) \\
&\leq 2^{f+d} \rho(x_{p_d}, x_{q_f}).
\end{aligned}$$

This ends the proof of the lemma.

C.2. Proof of Proposition 5

We fix $\bar{x} \in \mathcal{X}$, $r > 0$ and $f^* = \mathbb{1}_{B(\bar{x},r)}$. We reason by the contrapositive and suppose that 2C1NN is not consistent on f^*. We will show that the process $\mathbb{X}$ disproves the $\text{SMV}_{(\mathcal{X},\rho)}$ condition by considering a partition for which, the process $\mathbb{X}$ does not visit a sublinear number of sets with nonzero probability.

Because 2C1NN is not consistent, $\delta := \mathbb{P}(\mathcal{L}_{\mathbb{X}}(2C1NN, f^*) > 0) > 0$. Therefore, there exists $0 < \epsilon \leq 1$ such that $\mathbb{P}(\mathcal{L}_{\mathbb{X}}(2C1NN, f^*) > \epsilon) > \frac{\delta}{2}$. Denote $\mathcal{A} := \{\mathcal{L}_{\mathbb{X}}(2C1NN, f^*) > \epsilon\}$. We therefore have $\mathbb{P}(\mathcal{A}) > \frac{\delta}{2}$. We now define a partition $\mathcal{P}$. Because $\mathcal{X}$ is separable, there exists a sequence $(x^i)_{i\geq 1}$ of elements of $\mathcal{X}$ which is dense i.e.

$$\forall x \in \mathcal{X}, \quad \inf_{i\geq 1} \rho(x, x^i) = 0.$$

We focus for now on the sphere $S(\bar{x}, r)$ and for any $\tau > 0$ we take $(P_i(\tau))_{i\geq 1}$ the sequence of sets included in $S(\bar{x}, r)$ defined by

$$P_i(\tau) := \big(S(\bar{x}, r) \cap B(x^i, \tau)\big) \setminus \left(\bigcup_{1\leq j<i} B(x^j, \tau)\right).$$

These sets are disjoint. Further, they partition $S(\bar{x}, r)$. Indeed, if $x \in S(\bar{x}, r)$, let $i \geq 1$ such that $\rho(x, x^i) \leq \tau$. Then, $x \in S(\bar{x}, r) \cap B(x^i, \tau) \subset \bigcup_{j\leq i} P_j^\tau$. We now pose

$$\tau_l := c_\epsilon \cdot \frac{r}{2^{l+1}},$$

for $l \geq 1$, where $c_\epsilon := \frac{1}{2 \cdot 2^{2^5/\epsilon}}$ is a constant dependant on ϵ only. We also pose $\tau_0 = r$. Then, because $\mathbb{X} \in \text{SMV}_{(\mathcal{X},\rho)}$, the process visits a sublinear number of sets of $\mathcal{P}_i(\tau_l)$ almost surely. Therefore, there exists an increasing sequence $(n_l)_{l\geq 1}$ such that for any $l \geq 1$,

$$\mathbb{P}\left[\forall n \geq n_l,\ |\{i,\ P_i(\tau_l) \cap \mathbb{X}_{<n} \neq \emptyset\}| \leq \frac{\epsilon}{2^7} n\right] \geq 1 - \frac{\delta}{2 \cdot 2^{l+2}} \quad \text{and} \quad n_{l+1} \geq \frac{2^6}{\epsilon} n_l$$

We denote by $\mathcal{E}_l$ this event. Thus, $\mathbb{P}[\mathcal{E}_l] \leq \frac{\delta}{2 \cdot 2^{l+2}}$. Now, for any $l \geq 1$, we now construct $\mu_l > 0$ such that

$$\mathbb{P}\left[\min_{i<j\leq n_l,\ X_i \neq X_j} \rho(X_i, X_j) > \mu_l\right] \geq 1 - \frac{\delta}{2 \cdot 2^{l+2}}.$$

We denote this event by $\mathcal{F}_l$. Thus $\mathbb{P}[\mathcal{F}_l] \leq \frac{\delta}{2 \cdot 2^{l+2}}$. Note that the sequence $(\mu_l)_{l\geq 1}$ is non-increasing. We now define radiuses $(z^i)_{i\geq 1}$ as follows:

$$z^i = \begin{cases} \mu_{l_i+1} & \text{if } \rho(x^i, \bar{x}) < r, \text{ where } \frac{r}{2^{l_i+1}} < r - \rho(x^i, \bar{x}) \leq \frac{r}{2^{l_i}} \\ 0 & \text{if } \rho(x^i, \bar{x}) \geq r, \end{cases}$$

and consider the sets $R_i := B(x^i, z^i) \cap \left\{x \in \mathcal{X} :\ \rho(x, \bar{x}) < r - \frac{r}{2^{l_i+2}}\right\}$. We construct

$$P_i := R_i \setminus \left(\bigcup_{k<i} R_k\right),$$

for $i \geq 1$. As shown in the following lemma, $(P_i)_{i\geq 1}$ forms a partition of $B(\bar{x}, r)$.

Lemma 23 *$(P_i)_{i\geq 1}$ forms a partition of $B(\bar{x}, r)$.*

We now define a second partition. We start by defining a sequence of radiuses $(r^i)_{i\geq 1}$ as follows

$$r^i = \begin{cases} c_\epsilon \inf_{x:\ \rho(x,\bar{x}) \leq r} \rho(x^i, x) & \text{if } \rho(x^i, \bar{x}) > r, \\ c_\epsilon \inf_{x:\ \rho(x,\bar{x}) \geq r} \rho(x^i, x) & \text{if } \rho(x^i, \bar{x}) < r, \\ 0 & \text{if } \rho(x^i, \bar{x}) = r. \end{cases}$$

We consider the sets $(A_i)_{i\geq 0}$ given by $A_0 = S(\bar{x}, r)$ and for $i \geq 1$,

$$A_i = B(x^i, r^i) \setminus \left(\bigcup_{1\leq j<i} B(x^j, r^j)\right).$$

We now show that these sets form a partition in the following lemma.

Lemma 24 *$(A_i)_{i\geq 0}$ forms a partition of $\mathcal{X}$.*

We now formally consider the product partition of $(P_i)_{i\geq 1}$ and $(A_i)_{i\geq 0}$ i.e.

$$\mathcal{Q} : \bigcup_{i\geq 0,\ A_i \subset B(\bar{x},r)} \bigcup_{j\geq 1} (A_i \cap P_j) \cup \bigcup_{i\geq 0,\ A_i \subset \mathcal{X}\setminus B(\bar{x},r)} A_i.$$

where we used the fact that sets A_i satisfy either $A_i \subset B(\bar{x}, r)$ or $A_i \subset \mathcal{X} \setminus B(\bar{x}, r)$. We will show that this partition disproves the $\text{SMV}_{(\mathcal{X},\rho)}$ hypothesis on $\mathbb{X}$. In practice, we will either prove that the process visits many sets from partition $(A_i)_{i\geq 0}$ or $(P_i)_{i\geq 1}$ and use the fact that the same analysis would work for $\mathcal{Q}$, the product partition as well.

We now consider a specific realization $\boldsymbol{x} = (x_t)_{t\geq 0}$ of the process $\mathbb{X}$ falling in the event $\mathcal{A} \bigcap_{l\geq 1}(\mathcal{E}_l \cap \mathcal{F}_l)$. This event has probability

$$\mathbb{P}\left[\mathcal{A}\bigcap_{l\geq 1}(\mathcal{E}_l \cap \mathcal{F}_l)\right] \geq \mathbb{P}[\mathcal{A}] - \sum_{l\geq 1}(\mathbb{P}[\mathcal{E}_l^c] + \mathbb{P}[\mathcal{F}_l^c]) \geq \frac{\delta}{2} - \frac{\delta}{4} = \frac{\delta}{4}.$$

Note that $\boldsymbol{x}$ is not random anymore. We now show that $\boldsymbol{x}$ does not visit a sublinear number of sets in the partition $\mathcal{Q}$.

We now denote by $(t_k)_{k\geq 1}$ the increasing sequence of all times when 2C1NN makes an error in the prediction of $f^*(x_t)$. Because the event $\mathcal{A}$ is satisfied, $\mathcal{L}_{\boldsymbol{x}}(2C1NN, f^*) > \epsilon$, therefore, we can define an increasing sequence of times $(T_l)_{l\geq 1}$ such that

$$\frac{1}{T_l}\sum_{t=1}^{T_l} \ell_{01}(2C1NN(\boldsymbol{x}_{<t}, \boldsymbol{y}_{<t}, x_t), f^*(x_t)) > \frac{\epsilon}{2}.$$

For any $l \geq 1$ consider the last index $k = \max\{u, t_u \leq T_l\}$ when 2C1NN makes a mistake. Then we obtain $k > \frac{\epsilon}{2}T_l \geq \frac{\epsilon}{2}t_k$. Considering the fact that $(T_l)_{l\geq 1}$ is an increasing unbounded sequence we therefore obtain an increasing sequence of indices $(k_l)_{l\geq 1}$ such that $t_{k_l} < \frac{2k_l}{\epsilon}$.

At an iteration where the new input x_t has not been previously visited we will denote by $\phi(t)$ the index of the nearest neighbor of the current dataset in the 2C1NN learning rule. Now let $l \geq 1$. We focus on the time t_{k_l}. Consider the tree $\mathcal{G}$ where nodes are times $\mathcal{T} := \{t,\ t \leq t_{k_l},\ x_t \notin \{x_u, u < t\}\}$ for which a new input was visited, where the parent relations are given by $(t, \phi(t))$ for $t \in \mathcal{T} \setminus \{1\}$. In other words, we construct the tree in which a new input is linked to its representant which was used to derive the target prediction. Note that by definition of the 2C1NN learning rule, each node has at most 2 children and a node is not in the dataset at time t_{k_l} when it has exactly 2 children.

Step 1. We now suppose that the majority of input points on which 2C1NN made a mistake belong to $B(\bar{x}, r)$ i.e.

$$|\{t \leq t_{k_l},\ \ell_{01}(2C1NN(\boldsymbol{x}_{<t}, \boldsymbol{y}_{<t}, x_t), f^*(x_t)) = 1,\ x_t \in B(\bar{x}, r)\}| \geq \frac{k_l}{2},$$

or equivalently $|\{k \leq k_l,\ x_{t_k} \in B(\bar{x}, r)\}| \geq \frac{k_l}{2}$.

Let us now consider the subgraph $\tilde{\mathcal{G}}$ given by restricting $\mathcal{G}$ only to nodes in the ball $B(\bar{x}, r)$—which are mapped to the true value 1—i.e. on times $\{t \in \mathcal{T},\ x_t \in B(\bar{x}, r)\}$. In this subgraph, the only times with no parent are times t_k with $k \leq k_l$ and $x_{t_k} \in B(\bar{x}, r)$ and possibly time $t = 1$. Indeed, if a time in $\tilde{\mathcal{G}}$ has a parent $\phi(t)$ in $\tilde{\mathcal{G}}$, the prediction of 2C1NN for x_t returned the correct answer 1. The converse is also true except for the root time $t = 1$ which has no parent in $\mathcal{G}$. Therefore, $\tilde{\mathcal{G}}$ is a collection of disjoint trees with roots times $\{t_k,\ k \leq k_l,\ x_{t_k} \in B(\bar{x}, r)\}$—and possibly $t = 1$ if $x_1 \in B(\bar{x}, r)$. For a given time t_k with $k \leq k_l$ and $x_{t_k} \in B(\bar{x}, r)$, we will denote

by $\mathcal{T}_k$ the corresponding tree in $\tilde{\mathcal{G}}$ with root t_k. We will say that the $\mathcal{T}_k$ is a *good* tree if all times $t \in \mathcal{T}_k$ of this tree are parent in $\mathcal{G}$ to at most 1 time from $\mathcal{X} \setminus B(\bar{x}, r)$ i.e. if

$$\forall t \in \mathcal{T}_k, \quad |\{u \leq t_{k_l},\ \phi(u) = t,\ \rho(x_u, \bar{x}) \geq r\}| \leq 1.$$

We denote by $G = \{k \leq k_l,\ x_{t_k} \in B(\bar{x}, r),\ \mathcal{T}_k \text{ good}\}$ the set of indices of good trees. By opposition, we will say that a tree is *bad* otherwise. We now give a simple upper bound on N_{bad} the number of bad trees. Note that for any $t \in \mathcal{T}_k$, times in $\{u \leq t_{k_l},\ \phi(u) = t,\ \rho(x_u, \bar{x}) \geq r\}$ are times when 2C1NN makes a mistake on $\mathcal{X} \setminus B(\bar{x}, r)$. Therefore,

$$\sum_{k \leq k_l,\ x_{t_k} \in B(\bar{x}, r)} \sum_{t \in \mathcal{T}_k} |\{u < t_{k_l},\ \phi(u) = t,\ \rho(x_u, \bar{x}) \geq r\}| \leq |\{k \leq t_{k_l}, \rho(x_{t_k}, \bar{x}) \geq r\}| \leq \frac{k_l}{2}$$

because by hypothesis $|\{k \leq k_l,\ x_{t_k} \in B(\bar{x}, r)\}| \geq \frac{k_l}{2}$. Therefore, since each bad tree contains a node which is parent to at least 2 times of mistake in $\mathcal{X} \setminus B(\bar{x}, r)$, we obtain

$$N_{\text{bad}} \leq \sum_{k \leq k_l,\ x_{t_k} \in B(\bar{x}, r)} \sum_{t \in \mathcal{T}_k} \frac{1}{2} |\{u < t_{k_l},\ \phi(u) = t,\ \rho(x_u, \bar{x}) \geq r\}| \leq \frac{k_l}{4}.$$

Thus, the number of good trees is $|G| \geq |\{k \leq k_l,\ x_{t_k} \in B(\bar{x}, r)\}| - N_{\text{bad}} \geq \frac{k_l}{4}$. Now note that trees are disjoint, therefore, $\sum_{k \in G} |\mathcal{T}_k| \leq t_{k_l} < \frac{2k_l}{\epsilon}$. Therefore,

$$\sum_{k \in G} \mathbb{1}_{|\mathcal{T}_k| \leq \frac{16}{\epsilon}} = |G| - \sum_{k \in G} \mathbb{1}_{|\mathcal{T}_k| > \frac{16}{\epsilon}} > |G| - \frac{\epsilon}{16} \sum_{k \in G} |\mathcal{T}_k| \geq \frac{k_l}{8}.$$

We will say that a tree $|\mathcal{T}_k|$ is *sparse* if it is good and has at most $\frac{\epsilon}{16}$ nodes. With $S := \{k \in G,\ |\mathcal{T}_k| \leq \frac{16}{\epsilon}\}$ the set of sparse trees, the above equation we have $|S| \geq \frac{k_l}{8}$. We now focus only on sparse trees $\mathcal{T}_k$ for $k \in S$ and analyze their relation with the final dataset $\mathcal{D}_{t_{k_l}}$. Precisely, for a sparse tree $\mathcal{T}_k$, denote $\mathcal{V}_k = \mathcal{T}_k \cap \mathcal{D}_{t_{k_l}}$ the set of times which are present in the final dataset and belong to the tree induced by error time t_k. Because each node of $\mathcal{T}_k$ and not present in $\mathcal{D}_{t_{k_l}}$ has at least 1 children in $\mathcal{T}$, we note that $\mathcal{V}_k \neq \emptyset$. We now consider the path from a node of $\mathcal{V}_k$ to the root t_k. We denote by $d(k)$ the depth of this node in $\mathcal{V}_k$ and denote the path by $p^k_{d(k)} \to p^k_{d(k)-1} \to p^k_0 = t_k$ where $p^k_{d(k)} \in \mathcal{V}_k$. Then we have,

$$d(k) \leq |\mathcal{T}_k| - 1 \leq \frac{16}{\epsilon} - 1.$$

Each arc of this path represents the fact that at the corresponding iteration p^k_i of 2C1NN, the parent $x_{p^k_{i-1}}$ was closer from $x_{p^k_i}$ than any other point of the current dataset $\mathcal{D}_{p^k_i}$. We will now show that all the points $\{p^k_{d(k)},\ k \in S\}$ fall in distinct sets of the partition $(A_i)_{i \geq 0}$. Suppose by contradiction that we have $k_1 \neq k_2 \in S$ falling into the same set A_i. Note that because $x_{p^{k_1}_{d(k_1)}}, x_{p^{k_2}_{d(k_2)}} \in B(\bar{x}, r)$, we obtain $A_i \cap B(\bar{x}, r) \neq \emptyset$. However, the partition $(A_i)_{i \geq 0}$ was constructed so that sets are included totally in either $B(\bar{x}, r)$, $S(\bar{x}, r)$ or $\{x \in \mathcal{X},\ \rho(x, \bar{x}) > r\}$. Therefore, we obtain $A_i \subset B(\bar{x}, r)$ and $x^i \in B(\bar{x}, r)$. We can now apply Lemma 6 to $p^{k_1}_{d(k_1)} \to p^{k_1}_{d(k_1)-1} \to \ldots \to p^{k_1}_0$ and $p^{k_2}_{d(k_2)} \to p^{k_2}_{d(k_2)-1} \to \ldots \to p^{k_2}_0$—which we write by convenience $p_d \to p_{d-1} \to \ldots \to p_1 \to p_0$ and

$q_f \to q_{f-1} \to \ldots \to q_1 \to q_0$—assuming without loss of generality that $p_0 < q_0$. Therefore, $\rho(x_{p_{v(0)}}, x_{q_0}) \leq 2^{f+d}\rho(x_{p_d}, x_{q_f}) \leq 2^{f+d+1}r^i$ and $\rho(x_{p_{v(0)}}, x_{p_d}) \leq 2^{f+d}\rho(x_{p_d}, x_{q_f}) \leq 2^{f+d+1}r^i$. But recall that these two paths come from sparse trees, so $d, f \leq \frac{16}{\epsilon} - 1$. Hence, $2^{f+d+1} \leq \frac{1}{2}2^{2^5/\epsilon} = \frac{1}{4c_\epsilon}$. Let us now consider $x_{\phi(q_0)}$ the point which induced a mistake in the prediction of x_{q_0}, i.e. $\rho(x_{\phi(q_0)}, \bar{x}) \geq r$. Then,

$$\begin{aligned}\rho(x_{q_0}, x_{\phi(q_0)}) &\geq \rho(x_{\phi(q_0)}, x^i) - \rho(x^i, x_{p_d}) - \rho(x_{p_d}, x_{p_{v(0)}}) - \rho(x_{p_{v(0)}}, x_{q_0}) \\ &\geq \frac{r^i}{c_\epsilon} - r^i - \frac{r^i}{4c_\epsilon} - \frac{r^i}{4c_\epsilon} \\ &\geq \frac{r^i}{4c_\epsilon}\end{aligned}$$

where in the last inequality we used the fact that $c_\epsilon < \frac{1}{4}$. Recall that we also proved $\rho(x_{p_{v(0)}}, x_{q_0}) \leq \frac{r^i}{4c_\epsilon} < \rho(x_{q_0}, x_{\phi(q_0)})$. However, datapoint $x_{p_{v(0)}}$ is available in dataset $\mathcal{D}_{q_0}$. This contradicts the fact that $x_{\phi(t)}$ was chosen as representant for x_{q_0}. This ends the proof that all the points $\{p^k_{d(k)},\ k \in S\}$ fall in distinct sets of the partition $(A_i)_{i\geq 0}$. Therefore,

$$|\{i,\ A_i \cap \boldsymbol{x}_{\leq t_{k_l}} \neq \emptyset\}| \geq |S| \geq \frac{k_l}{8} \geq \frac{\epsilon}{16}t_{k_l}.$$

Step 2. We now turn to the case when the majority of input points on which 2C1NN made a mistake are not in the ball $B(\bar{x}, r)$ i.e.

$$|\{t \leq t_{k_l},\ \ell_{01}(2C1NN(\boldsymbol{x}_{<t}, \boldsymbol{y}_{<t}, x_t), f^*(x_t)) = 1,\ \rho(x_t, \bar{x}) \geq r\}| \geq \frac{k_l}{2},$$

or equivalently $|\{k \leq k_l,\ \rho(x_{t_k}, \bar{x}) \geq r\}| \geq \frac{k_l}{2}$. Similarly as the previous case, we consider the graph $\tilde{G}$ given by restricting $\mathcal{G}$ only to nodes outside the ball $B(\bar{x}, r)$ i.e. on times $\{t \in \mathcal{T}, \rho(x_t, \bar{x}) \geq r\}$. Again, $\tilde{\mathcal{G}}$ is a collection of disjoint trees with root times $\{t_k,\ k \leq k_l,\ \rho(x_{t_k}, \bar{x}) \geq r\}$ (and possibly $t = 1$). We denote $\mathcal{T}_k$ the corresponding tree of $\tilde{\mathcal{G}}$ rooted in t_k. Similarly to above, a tree is *sparse* if

$$\forall t \in \mathcal{T}_k, \quad |\{u \leq t_{k_l},\ \phi(u) = t,\ \rho(x_u, \bar{x}) < r\}| \leq 1 \quad \text{and} \quad |\mathcal{T}_k| \leq \frac{16}{\epsilon}.$$

If $S = \{k \leq k_l, ; \rho(x_{t_k}, \bar{x}) \geq r,\ \mathcal{T}_k \text{ sparse}\}$ denotes the set of sparse trees, the same proof as above shows that $|S| \geq \frac{k_l}{8}$. Again, for any $k \in S$, if $d(k)$ denotes the depth of some node from $\mathcal{V}_k := \mathcal{T}_k \cap \mathcal{D}_{t_{k_l}}$ in $\mathcal{T}_k$ we have $d(k) \leq \frac{16}{\epsilon} - 1$. For each $k \in S$ we consider the path from this node of $\mathcal{V}_k$ to the root t_k: $p^k_{d(k)} \to p^k_{d(k)-1} \to \ldots \to p^k_0 = t_k$ where $p^k_{d(k)} \in \mathcal{V}_k$. The same proof as above shows that all the points $\{p^k_{d(k)},\ k \in S,\ \rho(x_{p^k_{d(k)}}, \bar{x}) > r\}$ lie in distinct sets of the partition $(A_i)_{i\geq 0}$.

Indeed, let $p_d \to p_{d-1} \to \ldots \to p_1 \to p_0$ and $q_f \to q_{f-1} \to \ldots \to q_1 \to q_0$ two such paths with $\rho(x_{p_d}, \bar{x}) > r$ and $\rho(x_{q_f}, \bar{x}) > r$ and suppose by contradiction that $x_{p_d}, x_{q_f} \in A_i$ for some $i \geq 0$. Necessarily, $i \geq 1$ and $\rho(x^i, \bar{x}) > r$. Lemma 6 gives again $\rho(x_{p_{v(0)}}, x_{q_0}), \rho(x_{p_{v(0)}}, x_{p_d}) \leq 2^{f+d}\rho(x_{p_d}, x_{q_f}) \leq 2^{f+d+1}r^i \leq \frac{r^i}{4c_\epsilon}$. Then, if $x_{\phi(q_0)}$ is the point that induced a mistake in the prediction of x_{q_0}, we have $\rho(x_{\phi(q_0)}, \bar{x}) < r$. Using the definition of r^i we obtain the same computations

$$\rho(x_{q_0}, x_{\phi(q_0)}) \geq \rho(x_{\phi(q_0)}, x^i) - \rho(x^i, x_{p_d}) - \rho(x_{p_d}, x_{p_{v(0)}}) - \rho(x_{p_{v(0)}}, x_{q_0}) \geq \frac{r^i}{4c_\epsilon} > \rho(x_{p_{v(0)}}, x_{q_0})$$

which contradicts the fact that $x_{\phi(q_0)}$ was used as representant for x_{q_0}. This ends the proof that all the points $\{p^k_{d(k)},\ k \in S,\ \rho(x_{p^k_{d(k)}}, \bar{x}) > r\}$ lie in distinct sets of the partition $(A_i)_{i\geq 0}$. Suppose $|\{k \in S,\ \rho(x_{p^k_{d(k)}}, \bar{x}) > r\}| \geq \frac{|S|}{2}$, then we have

$$|\{i,\ A_i \cap \boldsymbol{x}_{\leq t_{k_l}} \neq \emptyset\}| \geq |\{k \in S,\ \rho(x_{p^k_{d(k)}}, \bar{x}) > r\}| \geq \frac{|S|}{2} \geq \frac{k_l}{16} \geq \frac{\epsilon}{32} t_{k_l}.$$

Step 3. In this last step, we suppose again that the majority of input points on which 2C1NN made a mistake are not in the ball $B(\bar{x}, r)$ and that $|\{k \in S,\ \rho(x_{p^k_{d(k)}}, \bar{x}) > r\}| < \frac{|S|}{2}$. Therefore, we obtain

$$|\{k \in S,\ \rho(x_{p^k_{d(k)}}, \bar{x}) = r\}| = |S| - |\{k \in S,\ \rho(x_{p^k_{d(k)}}, \bar{x}) > r\}| \geq \frac{|S|}{2} \geq \frac{k_l}{16} \geq \frac{\epsilon}{32} t_{k_l}.$$

We will now make use of the partition $(P_i)_{i\geq 1}$. Because $(n_u)_{u\geq 1}$ is an increasing sequence, let $u \geq 1$ such that $n_{u+1} \leq t_{k_l} \leq n_{u+2}$ (we can suppose without loss of generality that $t_{k_0} > n_2$). Note that we have $n_u \leq \frac{\epsilon}{2^6} n_{u+1} \leq \frac{\epsilon}{2^6} t_{k_l}$. Let us now analyze the process between times n_u and t_{k_l}. In particular, we are interested in the indices $T = \{k \in S,\ \rho(x_{p^k_{d(k)}}, \bar{x}) = r\}$ and times $\mathcal{U}_u = \{p^k_{d(k)} :\ n_u < p^k_{d(k)} \leq k_l,\ k \in T\}$. In particular, we have

$$|\mathcal{U}_u| \geq |\{k \in S,\ \rho(x_{p^k_{d(k)}}, \bar{x}) = r\}| - n_u \geq \frac{\epsilon}{32} t_{k_l} - \frac{\epsilon}{2^6} t_{k_l} = \frac{\epsilon}{2^6} t_{k_l}.$$

Because the event $\mathcal{E}_u$ is met, we have

$$|\{i,\ P_i(\tau_u) \cap \boldsymbol{x}_{\mathcal{U}_u} \neq \emptyset\}| \leq |\{i,\ P_i(\tau_u) \cap \boldsymbol{x}_{\leq t_{k_l}} \neq \emptyset\}| \leq \frac{\epsilon}{2^7} t_{k_l}.$$

Note that $\boldsymbol{x}_{\mathcal{U}_u} \subset S(\bar{x}, r)$. Therefore, each of the points in $\boldsymbol{x}_{\mathcal{U}_u}$ falls into one of the sets $(P_i(\tau_u))_{i\geq 1}$. Let $i \geq 1$ such that the set $P_i(\tau_u)$ was visited by $\boldsymbol{x}_{\mathcal{U}_u}$ and consider $T_i = \{k \in T,\ x_{p^k_{d(k)}} \in A_i\}$. We will show that at least $|T_i| - 1$ of the points $\{x_{\phi(t_k)},\ k \in T_i\}$ fall in $B(\bar{x}, r) \setminus B(\bar{x}, r - \frac{r}{2^{u+2}})$.

To do so, let $k_1, k_2 \in T_i$. Similarly as above, for simplicity, we will refer to the path $p^{k_1}_{d(k_1)} \to p^{k_1}_{d(k_1)-1} \to \ldots \to p^{k_1}_0$ (resp. $p^{k_2}_{d(k_2)} \to p^{k_2}_{d(k_2)-1} \to \ldots \to p^{k_2}_0$) as $p_d \to p_{d-1} \to \ldots \to p_1 \to p_0$ (resp. $q_f \to q_{f-1} \to \ldots \to q_1 \to q_0$), and assume without loss of generality that $p_0 < q_0$. Note that by hypothesis, $k_1, k_2 \in T_i$, therefore, $\rho(x_{p_d}, x^i), \rho(x_{q_f}, x^i) \leq \tau_u$ Then, using the above computations yields

$$\rho(x_{p_{v(0)}}, x_{q_0}) \leq 2^{f+d} \rho(x_{p_d}, x_{q_f}) \leq 2^{f+d}(\rho(x_{p_d}, x^i) + \rho(x_{q_f}, x^i)) \leq 2^{f+d+1} \tau_u \leq \frac{\tau_u}{4c_\epsilon},$$

where in the last inequality we used the fact that $f, d \leq \frac{16}{\epsilon} - 1$ hence $2^{f+d+1} \leq \frac{1}{4c_\epsilon}$. Now by definition of a representant, we obtain

$$\rho(x_{\phi(q_0)}, x_{q_0}) \leq \rho(x_{p_{v(0)}}, x_{q_0}) \leq \frac{r}{8 \cdot 2^u}.$$

Therefore, $\rho(x_{\phi(q_0)}, \bar{x}) \geq \rho(x_{q_0}, \bar{x}) - \rho(x_{\phi(q_0)}, x_{q_0}) \geq r - \frac{r}{8 \cdot 2^u}$. Because $x_{\phi(q_0)}$ induced a mistake in the prediction for x_{q_0} we have $x_{\phi(q_0)} \in B(\bar{x}, r)$. Now order $T_i = \{k_1 < \ldots < k_{|T_i|}\}$. We then have $t_{k_1} < \ldots < t_{k_{|T_i|}}$. The argument above then shows that for any $2 \leq j \leq |T_i|$, we have

$x_{\phi(t_{k_j})} \in B(\bar{x}, r) \backslash B(\bar{x}, r - \frac{r}{2^{u+3}})$. Therefore, defining $T' := \{k \in T,\ r - \frac{r}{2^{u+3}} \leq \rho(x_{\phi(t_k)}, \bar{x}) < r\}$ we obtain

$$\left|T'\right| \geq |\mathcal{U}_u| - |\{i,\ P_i(\tau_u) \cap \boldsymbol{x}_{\mathcal{U}_u} \neq \emptyset\}| \geq \frac{\epsilon}{2^7} t_{k_l}.$$

We will now show that all the points in $\{x_{t_k},\ k \in T'\}$ lie in distinct sets of $(P_i)_{i \geq 1}$. Note that because we have $t_{k_l} \leq n_{u+2}$ and because the event $\mathcal{F}_{u+2}$ is met, we have that for any $p, q \in T'$ that $\rho(x_{\phi(t_p)}, x_{\phi(t_q)}) > \mu_{u+2}$. Now suppose by contradiction that $x_{\phi(t_p)}, x_{\phi(t_q)} \in P_i$ for some $i \geq 1$. Then, with l_i such that $r - \frac{r}{2^{l_i}} \leq \rho(x^i, \bar{x}) < r - \frac{r}{2^{l_i+1}}$ we have that

$$x_{\phi(t_p)}, x_{\phi(t_q)} \in \left\{x \in \mathcal{X} :\ \rho(x, \bar{x}) < r - \frac{r}{2^{l_i+2}}\right\}$$

But we know that $\rho(x_{\phi(t_p)}, \bar{x}) \geq r - \frac{r}{2^{u+3}}$. Therefore we obtain $r - \frac{r}{2^{l_i+2}} > r - \frac{r}{2^{u+3}}$ and hence $l_i \geq u + 1$. Recall that $P_i \subset B(x^i, \mu_{l_i+1})$. Therefore, we obtain

$$\rho(x_{\phi(t_p)}, x_{\phi(t_q)}) \leq \mu_{l_i+1} \leq \mu_{u+2},$$

which contradicts the fact that $\rho(x_{t_p}, x_{t_q}) > \mu_{u+2}$. This ends the proof that all points of $\{x_{t_k},\ k \in T'\}$ lie in distinct subsets of $(P_i)_{i \geq 1}$. Now we obtain

$$|\{i,\ P_i \cap \boldsymbol{x}_{\leq t_{k_l}} \neq \emptyset\}| \geq |T'| \geq \frac{\epsilon}{2^7} t_{k_l}.$$

Step 4. In conclusion, in all cases, we obtain

$$|\{Q \in \mathcal{Q},\ Q \cap \boldsymbol{x}_{\leq t_{k_l}} \neq \emptyset\}| \geq \max(|\{i,\ A_i \cap \boldsymbol{x}_{\leq t_{k_l}} \neq \emptyset\}|, |\{i,\ P_i \cap \boldsymbol{x}_{\leq t_{k_l}} \neq \emptyset\}|) \geq \frac{\epsilon}{2^7} t_{k_l}.$$

Because this is true for all $l \geq 1$ and t_{k_l} is an increasing sequence, we conclude that $\boldsymbol{x}$ disproves the $\text{SMV}_{(\mathcal{X},\rho)}$ condition for $\mathcal{Q}$. Recall that this holds whenever the event $\mathcal{A} \bigcap_{l \geq 1} (\mathcal{E}_l \cap \mathcal{F}_l)$ is met. Thus,

$$\mathbb{P}[|\{Q \in \mathcal{Q},\ Q \cap \mathbb{X}_{<T}\}| = o(T)] \leq 1 - \mathbb{P}[\mathcal{A} \bigcap_{l \geq 1} (\mathcal{E}_l \cap \mathcal{F}_l)] \leq 1 - \frac{\delta}{4} < 1.$$

This shows that $\mathbb{X} \notin \text{SMV}_{(\mathcal{X},\rho)}$ which is absurd. Therefore 2C1NN is consistent on f^*. This ends the proof of the proposition.

C.3. Missing proofs from Section C.2

Lemma 25 *$(P_i)_{i \geq 1}$ forms a partition of $B(\bar{x}, r)$.*

Proof These sets are clearly disjoint. Now let $x \in B(\bar{x}, r)$ and consider $j \geq 0$ such that $\frac{r}{2^{j+1}} < r - \rho(x, \bar{x}) \leq \frac{r}{2^j}$. Then, let $i \geq 1$ such that

$$\rho(x^i, x) < \min\left(\mu_{j+1}, r - \frac{r}{2^{j+1}} - \rho(x, \bar{x}), \rho(x, \bar{x}) - r + \frac{r}{2^{j-1}}\right).$$

We have $\rho(x^i, \bar{x}) \leq \rho(x^i, x) + \rho(x, \bar{x}) < r - \frac{r}{2^{j+1}}$, hence $r - \frac{r}{2^{l_i}} < r - \frac{r}{2^{j+1}}$ i.e. $l_i \leq j$. Then, we obtain $\rho(x^i, x) < \mu_{j+1} \leq \mu_{l_i+1}$ which gives $x \in B(x^i, z^i)$. Last, we observe that

$\rho(x^i,\bar{x}) \geq \rho(x,\bar{x}) - \rho(x^i,\bar{x}) > r - \frac{r}{2^{j-1}}$. Therefore, $r - \frac{r}{2^{l_i+1}} > r - \frac{r}{2^{j-1}}$ i.e. $l_i + 1 \geq j$. Therefore, we have

$$\rho(x,\bar{x}) < r - \frac{r}{2^{j+1}} \leq r - \frac{r}{2^{l_i+2}},$$

which shows $x \in R_i = \bigcup_{k\leq i} P_k$. This ends the proof that $(P_i)_{i\geq 1}$ forms a partition of $B(\bar{x},r)$. ■

Lemma 26 *$(A_i)_{i\geq 0}$ forms a partition of $\mathcal{X}$.*

Proof We start by proving that the sets are disjoint. By construction, if $1 \leq j < i$, we have $A_i \subset B(x^j, r^j)$, therefore $A_i \cap A_j = \emptyset$ by construction. Further, for $i \geq 1$, if $\rho(x^i,\bar{x}) > r$, we first note that $r^i > 0$. Indeed, if $r^i = 0$, then there exists a sequence of points x_j for $j \geq 1$ such that $\rho(x_j,\bar{x}) \leq r$ and $\rho(x^i,x_j) \to 0$ as $j \to \infty$. By triangle inequality,

$$\rho(x^i,\bar{x}) \leq \rho(x^i,x_j) + \rho(x_j,\bar{x}) \leq \rho(x^i,x_j) + r.$$

This holds for any $j \geq 1$, therefore we obtain $\rho(x^i,\bar{x}) \leq r$ which contradicts our hypothesis. Therefore $r^i > 0$. Further, we have $r^i < \inf_{x:\ \rho(x,\bar{x})\leq r} \rho(x^i,x)$. Therefore, for any $x \in A_0 = S(\bar{x},r)$, we have $\rho(x^i,x) > r^i$ which implies $x \notin B(x^i,r^i)$. Hence, $A_0 \cap A_i = \emptyset$. Now if $\rho(x^i,\bar{x}) < r$ we show again that $r^i > 0$. Similarly, if this is not the case, we have a sequence x_j for $j \geq 1$ such that $\rho(x_j,\bar{x}) \geq r$ and $\rho(x^i,x_j) \to 0$ as $j \to \infty$. Then, observing that

$$\rho(x^i,\bar{x}) \geq \rho(x^i,x_j) - \rho(x^i,x_j) \geq r - \rho(x^i,x_j).$$

This holds for any $j \geq 1$, therefore we obtain $\rho(x^i,\bar{x}) \geq r$ which contradicts our hypothesis. This shows $r^i > 0$. Now for $x \in A_0$, we have by construction $r^i < \rho(x^i,x)$ which gives $x \notin A_i$. Hence $A_0 \cap A_i = \emptyset$. Finally, if $\rho(x^i,\bar{x}) = r$, we have $r^i = 0$ so $A_i = \emptyset$ and we obtain direly $A_0 \cap A_i = \emptyset$. This ends the proof that for any $0 \leq i < j$, we have $A_i \cap A_j = \emptyset$.

We now prove that $\cup_{i\geq 0} A_i = \mathcal{X}$. Let $x \in \mathcal{X}$. If $\rho(x,\bar{x}) = r$ then $x \in A_0$. If $\rho(x,\bar{x}) > r$ (resp. $\rho(x,\bar{x}) < r$), using the same arguments as above, we can show that $\inf_{\tilde{x}:\ \rho(\tilde{x},\bar{x})\leq r} \rho(x,\tilde{x}) > 0$ (resp. $\inf_{\tilde{x}:\ \rho(\tilde{x},\bar{x})\geq r} \rho(x,\tilde{x}) > 0$). Therefore, we let $i \geq 1$ such that $\rho(x^i,x) < \frac{1}{1+\frac{2}{c_\epsilon}} \inf_{\tilde{x}:\ \rho(\tilde{x},\bar{x})\leq r} \rho(x,\tilde{x})$ (resp. $\rho(x^i,x) < \frac{1}{1+\frac{2}{c_\epsilon}} \inf_{\tilde{x}:\ \rho(\tilde{x},\bar{x})\geq r} \rho(x,\tilde{x})$). This is possible because the sequence $(x^i)_{i\geq 1}$ is dense in $\mathcal{X}$. Then, we have for any $\tilde{x}$ such that $\rho(\tilde{x},\bar{x}) \leq r$ (resp. $\rho(\tilde{x},\bar{x}) \geq r$),

$$\rho(x^i,\tilde{x}) \geq \rho(x,\tilde{x}) - \rho(x^i,x) > \left(1 + \frac{2}{c_\epsilon} - 1\right) \rho(x^i,x) = \frac{2}{c_\epsilon} \rho(x^i,x).$$

Therefore, $r^i \geq 2\rho(x^i,x) > \rho(x^i,x)$ which gives $x \in B(x^i,r^i)$. Now note that $\bigcup_{1\leq j\leq i} A_i = \bigcup_{1\leq j\leq i} B(x^i,r^i)$, therefore we obtain $x \in \bigcup_{1\leq j\leq i} A_i$. This ends the proof that $(A_i)_{i\geq 0}$ forms a partition of $\mathcal{X}$. ■

C.4. Proof of Theorem 7

let $\mathbb{X} \in \text{SMV}_{(\mathcal{X},\rho)}$. We will show that 2C1NN is universally consistent on $\mathbb{X}$ by considering the set $\mathcal{S}_{\mathbb{X}}$ of functions for which it is consistent. More precisely, since $\mathcal{Y} = \{0,1\}$ in the binary setting, all target functions can be described as $f^* = \mathbb{1}_{A_{f^*}}$ where $A_{f^*} = f^{<-1>}(\{1\})$. We define $\mathcal{S}_{\mathbb{X}}$ using the corresponding sets:

$$\mathcal{S}_{\mathbb{X}} := \{A \in \mathcal{B}, \quad \mathcal{L}_{\mathbb{X}}(2C1NN, \mathbb{1}_{.\in A}) = 0 \quad (a.s.)\}$$

By construction we have $\mathcal{S}_{\mathbb{X}} \subset \mathcal{B}$. The goal is to show that in fact $\mathcal{S}_{\mathbb{X}} = \mathcal{B}$. To do so, we will show that $\mathcal{S}$ satisfies the following properties

- $\emptyset \in \mathcal{S}_{\mathbb{X}}$ and $\mathcal{S}_{\mathbb{X}}$ contains all balls $B(x,r)$ with $x \in \mathcal{X}$ and $r \geq 0$,
- if $A \in \mathcal{S}_{\mathbb{X}}$ then $A^c \in \mathcal{S}_{\mathbb{X}}$ (stable to complementary),
- if $(A_i)_{i\geq 1}$ is a sequence of disjoint sets of $\mathcal{S}_{\mathbb{X}}$, then $\bigcup_{i\geq 1} A_i \in \mathcal{S}_{\mathbb{X}}$ (stable to $\sigma-$additivity for disjoint sets),
- if $A, B \in \mathcal{S}_{\mathbb{X}}$, then $A \cup B \in \mathcal{S}_{\mathbb{X}}$ (stable to union).

Together, these properties show that $\mathcal{S}_{\mathbb{X}}$ is a $\sigma-$algebra that contains all open intervals of $\mathcal{X}$. Recall that by definition, $\mathcal{B}$ is the smallest $\sigma-$algebra containing open intervals. Therefore we get $\mathcal{B} \subset \mathcal{S}_{\mathbb{X}}$ which proves the theorem. We now show the four properties.

The invariance to complementary and to finite union can be shown with the same proof as Theorem 17. Further, we clearly have $\emptyset \in \mathcal{S}_{\mathbb{X}}$. Now let $x \in \mathcal{X}$ and $r \geq 0$, Proposition 5 shows that $B(x,r) \in \mathcal{S}_{\mathbb{X}}$.

We now turn to the $\sigma-$additivity for disjoint sets. Let $(A_i)_{i\geq 1}$ is a sequence of disjoint sets of $\mathcal{S}_{\mathbb{X}}$. We denote $A := \bigcup_{i\geq 1} A_i$. We consider the target function $f^* = \mathbb{1}_A$. We write the average loss in the following way,

$$\frac{1}{T}\sum_{t=1}^{T} \ell_{01}(2C1NN(\mathbb{X}_{<t}, \mathbb{Y}_{<t}, X_t), f^*(X_t)) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{X_t\in A}\mathbb{1}_{X_{\phi(t)}\notin A} + \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{X_t\notin A}\mathbb{1}_{X_{\phi(t)}\in A}.$$

where the first term corresponds to type 1 errors and the second term corresponds to type 2 errors.

We suppose by contradiction that $\mathbb{P}(\mathcal{L}_{\mathbb{X}}(2C1NN, f^*) > 0) := \delta > 0$ Therefore, there exists $\epsilon > 0$ such that $\mathbb{P}(\mathcal{L}_{\mathbb{X}}(2C1NN, f^*) > \epsilon) \geq \frac{\delta}{2}$. We denote this event by $\mathcal{A} := \{\mathcal{L}_{\mathbb{X}}(2C1NN, f^*) > \epsilon\}$. We first analyze the errors induced by one set A_i only. We have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}(\mathbb{1}_{X_t\in A_i}\mathbb{1}_{X_{\phi(t)}\notin A} + \mathbb{1}_{X_t\notin A}\mathbb{1}_{X_{\phi(t)}\in A_i}) &\leq \frac{1}{T}\sum_{t=1}^{T}(\mathbb{1}_{X_t\in A_i}\mathbb{1}_{X_{\phi(t)}\notin A_i} + \mathbb{1}_{X_t\notin A_i}\mathbb{1}_{X_{\phi(t)}\in A_i}) \\ &= \frac{1}{T}\sum_{t=1}^{T} \ell_{01}(2C1NN(\mathbb{X}_{<t}, \mathbb{1}_{\mathbb{X}_{<t}\in A_i}, X_t), \mathbb{1}_{X_t\in A_i}).\end{aligned}$$

Then, because 2C1NN is consistent for $\mathbb{1}_{.\in A_i}$, we get

$$\frac{1}{T}\sum_{t=1}^{T}(\mathbb{1}_{X_t\in A_i}\mathbb{1}_{X_{\phi(t)}\notin A} + \mathbb{1}_{X_t\notin A}\mathbb{1}_{X_{\phi(t)}\in A_i}) \to 0 \quad (a.s.).$$

We take $\epsilon_i = \frac{\epsilon}{4 \cdot 2^i}$. The above equation gives T^i such that

$$\mathbb{P}\left[\bigcap_{T \geq T^i} \left\{ \frac{1}{T} \sum_{t=1}^{T} (\mathbb{1}_{X_t \in A_i} \mathbb{1}_{X_{\phi(t)} \notin A} + \mathbb{1}_{X_t \notin A} \mathbb{1}_{X_{\phi(t)} \in A_i}) < \epsilon_i \right\}\right] \geq 1 - \frac{\delta}{8 \cdot 2^i}.$$

We will denote by $\mathcal{E}_i$ this event. We now consider the scale of the process $\mathbb{X}_{\leq T^i}$ when falling in A_i, by introducing $\eta_i > 0$ such that

$$\mathbb{P}\left[\min_{\substack{t_1, t_2 \leq T^i;\ X_{t_1}, X_{t_2} \in A_i; \\ X_{t_1} \neq X_{t_2}}} \rho(X_{t_1}, X_{t_2}) > \eta_i \right] \geq 1 - \frac{\delta}{8 \cdot 2^i}.$$

We denote by $\mathcal{F}_i$ this event. By the union bound, we have $\mathbb{P}(\bigcup_{i \geq 1} \mathcal{E}_i^c \cup \bigcup_{i \geq 1} \mathcal{F}_i^c) \leq \frac{\delta}{4}$. Therefore, we obtain $\mathbb{P}(\mathcal{A} \cap \bigcap_{i \geq 1} \mathcal{E}_i \cap \bigcap_{i \geq 1} \mathcal{F}_i) \geq \mathbb{P}(\mathcal{A}) - \mathbb{P}(\bigcup_{i \geq 1} \mathcal{E}_i^c \cup \bigcup_{i \geq 1} \mathcal{F}_i^c) \geq \frac{\delta}{4}$. We now construct a partition $\mathcal{P}$ obtained by subdividing each set A_i according to scale η_i. Because $\mathcal{X}$ is separable, there exists a sequence of points $(x^j)_{j \geq 1}$ in $\mathcal{X}$ such that $\forall x \in \mathcal{X}, \inf_{j \geq 1} \rho(x, x^j) = 0$. We construct the following partition of $\mathcal{X}$ given by

$$\mathcal{P} \ : \quad A^c \cup \bigcup_{i \geq 1} \bigcup_{j \geq 1} \left\{ \left(B\left(x^j, \frac{\eta_i}{2}\right) \cap A_i \right) \setminus \bigcup_{k < j} B\left(x^k, \frac{\eta_i}{2}\right) \right\}.$$

Let us now consider a realization of $\boldsymbol{x}$ of $\mathbb{X}$ in the event $\mathcal{A} \cap \bigcap_{i \geq 1} \mathcal{E}_i \cap \bigcap_{i \geq 1} \mathcal{F}_i$. The sequence $\boldsymbol{x}$ is now not random anymore. Our goal is to show that $\boldsymbol{x}$ does not visit a sublinear number of sets in the partition $\mathcal{P}$.

By construction, the event $\mathcal{A}$ is satisfied, therefore there exists an increasing sequence of times $(t_k)_{k \geq 1}$ such that for any $k \geq 1$, $\frac{1}{t_k} \sum_{t=1}^{t_k} \ell_{01}(2C1NN(\boldsymbol{x}_{<t}, \mathbb{1}_{\boldsymbol{x}_{<t} \in A}, x_t), \mathbb{1}_{x_t \in A}) > \frac{\epsilon}{2}$. Therefore, we obtain for any $k \geq 1$,

$$\sum_{i \geq 1} \frac{1}{t_k} \sum_{t=1}^{t_k} (\mathbb{1}_{x_t \in A_i} \mathbb{1}_{x_{\phi(t)} \notin A} + \mathbb{1}_{x_t \notin A} \mathbb{1}_{x_{\phi(t)} \in A_i}) > \frac{\epsilon}{2}.$$

Also, because the events $\mathcal{E}_i$ are met, we have

$$\sum_{i \geq 1;\ t_k \geq T^i} \frac{1}{t_k} \sum_{t=1}^{t_k} (\mathbb{1}_{x_t \in A_i} \mathbb{1}_{x_{\phi(t)} \notin A} + \mathbb{1}_{x_t \notin A} \mathbb{1}_{x_{\phi(t)} \in A_i}) < \sum_{i \geq 1, t_k \geq T^i} \epsilon_i \leq \frac{\epsilon}{4}.$$

Combining the two above equations gives

$$\frac{1}{t_k} \sum_{t=1}^{t_k} \sum_{i \geq 1;\ t_k < T^i} (\mathbb{1}_{x_t \in A_i} \mathbb{1}_{x_{\phi(t)} \notin A} + \mathbb{1}_{x_t \notin A} \mathbb{1}_{x_{\phi(t)} \in A_i}) > \frac{\epsilon}{4}. \tag{2}$$

We now consider the set of times such that an input point fell into the set A_i with $T^i > t_k$, either creating a mistake in the prediction of 4C1NN or inducing a later mistake within time horizon t_k: $\mathcal{T} := \bigcup_{i \geq 1;\ T^i > t_k} \mathcal{T}_i$ where

$$\mathcal{T}_i := \left\{ t \leq t_k,\ x_t \in A_i,\ \left(x_{\phi(t)} \notin A \text{ or } \exists t < u \leq t_k \text{ s.t. } \phi(u) = t,\ x_u \notin A \right) \right\}.$$

We now show that all points x_t for $t \in \mathcal{T}$ fall in distinct sets of the partition $\mathcal{P}$. Indeed, because the sets A_i are disjoint, it suffices to check that for any $i \geq 1$ such that $T^i > t_k$, the points x_t for $t \in \mathcal{T}_i$ fall in distinct of the following sets

$$P_{i,j} := \left(B\left(x^j, \frac{\eta_i}{2}\right) \cap A_i\right) \setminus \bigcup_{k<j} B\left(x^k, \frac{\eta_i}{2}\right), \quad j \geq 1.$$

Note that for any $t_1 < t_2 \in \mathcal{T}_i$ we have $x_{t_1}, x_{t_2} \in A_i$ and $x_{t_1} \neq x_{t_2}$. Indeed, we cannot have $x_{t_2} = x_{t_1}$ otherwise 2C1NN would make no mistake at time t_2 and x_{t_2} would induce no future mistake either (recall that if an input point was already visited, we use simple memorization for the prediction and do not add it to the dataset). Therefore, because the event $\mathcal{F}_i$ is satisfied, for any $t_1 < t_2 \in \mathcal{T}_i$ we have $\rho(x_{t_1}, x_{t_2}) > \eta_i$. Now suppose that x_{t_1}, x_{t_2} fall in the same set $P_{i,j}$ for $j \geq 1$, then we have $\rho(x_{t_1}, x_{t_2}) \leq \rho(x^i, x_{t_1}) + \rho(x^i, x_{t_2}) < \eta_i$, which is absurd. Therefore, all points $\{x_t,\ t \in \mathcal{T}\}$ lie in different sets of the partition $\mathcal{P}$. Therefore,

$$|\{P \in \mathcal{P}, P \cap \boldsymbol{x}_{\leq t_k} \neq \emptyset\}| \geq |\mathcal{T}|.$$

We now lower bound $|\mathcal{T}|$, which will uncover the main interest of the learning rule 2C1NN. Intuitively, any input point incurs at most $1 + 2 = 3$ mistakes, contrary to the traditional 1NN learning rule. We now formalize this intuition.

$$\begin{aligned}
&\sum_{t=1}^{t_k} \sum_{i\geq 1;\ t_k<T^i} (\mathbb{1}_{x_t\in A_i}\mathbb{1}_{x_{\phi(t)}\notin A} + \mathbb{1}_{x_t\notin A}\mathbb{1}_{x_{\phi(t)}\in A_i}) \\
&\qquad = \sum_{t=1}^{t_k} \sum_{i\geq 1;\ t_k<T^i} \left(\mathbb{1}_{x_t\in A_i}\mathbb{1}_{x_{\phi(t)}\notin A} + \sum_{t<u\leq t_k} \mathbb{1}_{x_u\notin A}\mathbb{1}_{x_t\in A_i}\mathbb{1}_{\phi(u)=t}\right) \\
&\qquad = \sum_{i\geq 1;\ T^i>t_k} \sum_{t\leq t_k,\ x_t\in A_i} \left(\mathbb{1}_{x_{\phi(t)}\notin A} + \sum_{t<u\leq t_k} \mathbb{1}_{x_u\notin A}\mathbb{1}_{\phi(u)=t}\right) \\
&\qquad \leq \sum_{i\geq 1;\ T^i>t_k} \sum_{t\leq t_k,\ x_t\in A_i} 3\max\left(\mathbb{1}_{x_{\phi(t)}\notin A}, \mathbb{1}_{x_u\notin A}\mathbb{1}_{\phi(u)=t},\ t<u\leq t_k\right) \\
&\qquad = 3|\mathcal{T}|
\end{aligned}$$

where in the last inequality we used the fact that a given time t can have at most 2 children i.e. $|\{u > t, \phi(u) = t\}| \leq 2$ with the 2C1NN learning rule. We now use Equation (2) to obtain

$$|\{P \in \mathcal{P}, P \cap \boldsymbol{x}_{\leq t_k} \neq \emptyset\}| \geq |\mathcal{T}| \geq \frac{\epsilon}{12} t_k.$$

This holds for any $k \geq 1$. Therefore, because $t_k \to \infty$ as $k \to \infty$ we get $|\{P \in \mathcal{P}, P \cap \boldsymbol{x}_{\leq T} \neq \emptyset\}| \neq o(T)$. Finally, this holds for any realization of $\mathbb{X}$ in the event $\mathcal{A} \cap \bigcap_{i\geq 1} \mathcal{E}_i \cap \bigcap_{i\geq 1} \mathcal{F}_i$. Therefore,

$$\mathbb{P}(|\{P \in \mathcal{P}, P \cap \boldsymbol{x}_{\leq T} \neq \emptyset\}| = o(T)) \leq \mathbb{P}\left[\left(\mathcal{A} \cap \bigcap_{i\geq 1} \mathcal{E}_i \cap \bigcap_{i\geq 1} \mathcal{F}_i\right)^c\right] \leq 1 - \frac{\delta}{4} < 1.$$

Therefore, $\mathbb{X} \notin \text{SMV}_{(\mathcal{X},\rho)}$ which contradicts the hypothesis. This concludes the proof that

$$\mathcal{L}_{\mathbb{X}}(2C1NN, \mathbb{1}_A) = 0 \quad (a.s.),$$

and hence, $\mathcal{S}_{\mathbb{X}}$ satisfies the disjoint $\sigma-$additivity property. This ends the proof of the theorem.

Appendix D. Proofs of Section 6: Weak universal learning

D.1. Proof of Proposition 11

The proof uses a similar structure to the proof of Proposition 5. We fix $\bar{x} \in \mathcal{X}$, $r > 0$ and $f^*(\cdot) = \mathbb{1}_{B(\bar{x},r)}$. We reason by the contrapositive and suppose that 2C1NN is not weakly consistent on f^*. We will show that the process $\mathbb{X}$ disproves the $\text{WSMV}_{(\mathcal{X},\rho)}$ condition.

Because 2C1NN is not weakly consistent for f^*, there exists ϵ and an increasing sequence of times $(T_l)_{l\geq 1}$ such that for any $l \geq 1$,

$$\mathbb{E}\mathcal{L}_{\mathbb{X}}(f_\cdot, f^*; T_l) \geq \epsilon T_l.$$

We now define a partition $\mathcal{P}$. Because $\mathcal{X}$ is separable, there exists a sequence $(x^i)_{i\geq 1}$ of elements of $\mathcal{X}$ which is dense. We focus for now on the sphere $S(\bar{x}, r)$ and for any $\tau > 0$ we take $(P_i(\tau))_{i\geq 1}$ the sequence of sets included in $S(\bar{x}, r)$ defined by

$$P_i(\tau) := \left(S(\bar{x}, r) \cap B(x^i, \tau)\right) \setminus \left(\bigcup_{1\leq j<i} B(x^j, \tau)\right).$$

These sets form a partition of $S(\bar{x}, r)$ as shown in the proof of Proposition 5. We now pose $\tau_l := c_\epsilon \cdot \frac{r}{2^{l+1}}$, for $l \geq 1$, where $c_\epsilon := \frac{1}{2\cdot 2^{2^5/\epsilon}}$ is a constant dependant on ϵ only. We also pose $\tau_0 = r$. Then, because $\mathbb{X} \in \text{WSMV}_{(\mathcal{X},\rho)}$, the expected number of sets visited of $\mathcal{P}_i(\tau_l)$ tends to 0. Therefore, there exists an increasing sequence $(n_l)_{l\geq 1}$ such that for any $l \geq 1$,

$$\forall n \geq n_l, \quad \mathbb{E}[|\{i,\ P_i(\tau_l) \cap \mathbb{X}_{<n} \neq \emptyset\}|] \leq \frac{\epsilon^2}{2^{10}} n \quad \text{and} \quad n_{l+1} \geq \frac{2^6}{\epsilon} n_l$$

Now, for any $l \geq 1$, we now construct $\mu_l > 0$ such that

$$\mathbb{P}\left[\min_{i<j\leq n_l,\ X_i \neq X_j} \rho(X_i, X_j) > \mu_l\right] \geq 1 - \frac{\epsilon}{2^{l+3}}.$$

We denote by $\mathcal{F}_l$ this event. Therefore $\mathbb{P}[\mathcal{F}_l^c] \leq \frac{\epsilon}{2^{l+3}}$. Note that the sequence $(\mu_l)_{l\geq 1}$ is non-increasing. We now define radiuses $(z^i)_{i\geq 1}$ as follows:

$$z^i = \begin{cases} \mu_{l_i+1} & \text{if } \rho(x^i, \bar{x}) < r, \text{ where } \frac{r}{2^{l_i+1}} < r - \rho(x^i, \bar{x}) \leq \frac{r}{2^{l_i}} \\ 0 & \text{if } \rho(x^i, \bar{x}) \geq r, \end{cases}$$

and consider the sets $R_i := B(x^i, z^i) \cap \left\{x \in \mathcal{X} :\ \rho(x, \bar{x}) < r - \frac{r}{2^{l_i+2}}\right\}$. We construct $P_i := R_i \setminus \left(\bigcup_{k<i} R_k\right)$, for $i \geq 1$. By Lemma 23, $(P_i)_{i\geq 1}$ forms a partition of $B(\bar{x}, r)$. We now define a second partition $(A_i)_{i\geq 1}$ similarly as in the proof of Proposition 5. We start by defining a sequence of radiuses $(r^i)_{i\geq 1}$ as follows

$$r^i = \begin{cases} c_\epsilon \inf_{x:\ \rho(x,\bar{x})\leq r} \rho(x^i, x) & \text{if } \rho(x^i, \bar{x}) > r, \\ c_\epsilon \inf_{x:\ \rho(x,\bar{x})\geq r} \rho(x^i, x) & \text{if } \rho(x^i, \bar{x}) < r, \\ 0 & \text{if } \rho(x^i, \bar{x}) = r, \end{cases}$$

and consider the sets $(A_i)_{i\geq 0}$ given by $A_0 = S(\bar{x}, r)$ and for $i \geq 1$, $A_i = B(x^i, r^i) \backslash \Big(\bigcup_{1\leq j<i} B(x^j, r^j) \Big)$. By Lemma 24, this forms a partition of $\mathcal{X}$. We now formally consider the product partition of $(P_i)_{i\geq 1}$ and $(A_i)_{i\geq 0}$ i.e.

$$\mathcal{Q} : \bigcup_{i\geq 0,\ A_i \subset B(\bar{x},r)} \bigcup_{j\geq 1} (A_i \cap P_j) \cup \bigcup_{i\geq 0,\ A_i \subset \mathcal{X}\backslash B(\bar{x},r)} A_i.$$

where we used the fact that sets A_i satisfy either $A_i \subset B(\bar{x}, r)$ or $A_i \subset \mathcal{X} \setminus B(\bar{x}, r)$. We will show that this partition disproves the $\text{WSMV}_{(\mathcal{X},\rho)}$ hypothesis on $\mathbb{X}$.

We now fix $l_0 \geq 1$ such that $T_{l_0} \geq n_2$ and consider $l \geq l_0$. We focus on time T_l. Define the event $\mathcal{A} := \{\mathcal{L}_{\mathbb{X}}(f_\cdot, f^*; T_l) \geq \frac{\epsilon}{2} T_l\}$. Note that we have

$$\mathbb{E}\mathcal{L}_{\mathbb{X}}(f_\cdot, f^*; T_l) \leq \frac{\epsilon}{2} T_l + \mathbb{P}[\mathcal{A}] T_l.$$

Therefore, $\mathbb{P}[\mathcal{A}] \geq \frac{\epsilon}{2}$. Also, because $(n_u)_{u\geq 1}$ is an increasing sequence, let $u \geq 1$ such that $n_{u+1} \leq T_l \leq n_{u+2}$. We define the event $\mathcal{E} = \{|\{i\ P_i(\tau_u) \cap \mathbb{X}_{\leq T_l} \neq \emptyset\}| \leq \frac{\epsilon}{2^7} T_l\}$. Then, we have by construction

$$\frac{\epsilon^2}{2^{10}} T_l \geq \mathbb{E}|\{i\ P_i(\tau_u) \cap \mathbb{X}_{\leq T_l} \neq \emptyset\}| \geq \frac{\epsilon}{2^7} T_l \mathbb{P}[\mathcal{E}^c].$$

Therefore, we have $\mathbb{P}[\mathcal{E}^c] \leq \frac{\epsilon}{8}$. Consider a specific realization $\boldsymbol{x} = (x_t)_{t\geq 0}$ of the process $\mathbb{X}$ falling in the event $\mathcal{A} \cap \mathcal{E} \cap \bigcap_{l\geq 1} \mathcal{F}_l$. This event has probability

$$\mathbb{P}\left[\mathcal{A} \cap \mathcal{E} \cap \bigcap_{l\geq 1} \mathcal{F}_l\right] \geq \mathbb{P}[\mathcal{A}] - \mathbb{P}[\mathcal{E}^c] - \sum_{l\geq 1} \mathbb{P}[\mathcal{F}_l^c] \geq \frac{\epsilon}{2} - \frac{\epsilon}{8} - \frac{\epsilon}{8} = \frac{\epsilon}{4}.$$

Note that $\boldsymbol{x}$ is not random anymore. We now show that $\boldsymbol{x}$ visits a large number of sets in the partition $\mathcal{Q}$. We now denote by $(t_k)_{k\geq 1}$ the increasing sequence of all times when 2C1NN makes an error in the prediction of $f^*(x_t)$. Define k_l such the last time of error before T_l i.e. $k_l = \max\{k \geq 1,\ t_k \leq T_l\}$. By construction, because $\mathcal{A}$ is met we have $k_l \geq \frac{\epsilon}{2} T_l$.

At an iteration where the new input x_t has not been previously visited we will denote by $\phi(t)$ the index of the nearest neighbor of the current dataset in the 2C1NN learning rule. Now let $l \geq 1$. Consider the tree $\mathcal{G}$ where nodes are times $\mathcal{T} := \{t,\ t \leq T_l,\ x_t \notin \{x_u, u < t\}\}$ for which a new input was visited, where the parent relations are given by $(t, \phi(t))$ for $t \in \mathcal{T} \setminus \{1\}$. Again, each node has at most 2 children and a node is not in the dataset at time T_l when it has exactly 2 children.

Step 1. We now suppose that the majority of input points on which 2C1NN made a mistake belong to the $B(\bar{x}, r)$ i.e.

$$|\{t \leq T_l,\ \ell_{01}(2C1NN(\boldsymbol{x}_{<t}, \boldsymbol{y}_{<t}, x_t), f^*(x_t)) = 1,\ x_t \in B(\bar{x}, r)\}| \geq \frac{k_l}{2},$$

or equivalently $|\{k \leq k_l,\ x_{t_k} \in B(\bar{x}, r)\}| \geq \frac{k_l}{2}$.

Let us now consider the subgraph $\tilde{\mathcal{G}}$ given by restricting $\mathcal{G}$ only to nodes in the the ball $B(\bar{x}, r)$ which are mapped to the true value 1 i.e. on times $\{t \in \mathcal{T},\ x_t \in B(\bar{x}, r)\}$. As in the proof of Proposition 5, $\tilde{\mathcal{G}}$ is a collection of disjoint trees with roots times $\{t_k,\ k \leq k_l,\ x_{t_k} \in B(\bar{x}, r)\}$—and

possibly $t = 1$ if $x_1 \in B(\bar{x}, r)$. For a given time t_k with $k \leq k_l$ and $x_{t_k} \in B(\bar{x}, r)$, denote $\mathcal{T}_k$ the corresponding tree in $\tilde{\mathcal{G}}$ with root t_k. We will say that the tree $\mathcal{T}_k$ is *sparse* if

$$\forall t \in \mathcal{T}_k, \quad |\{u \leq T_l,\ \phi(u) = t,\ \rho(x_u, \bar{x}) < r\}| \leq 1 \quad \text{and} \quad |\mathcal{T}_k| \leq \frac{16}{\epsilon}.$$

We denote by $S = \{k \leq k_l,\ \rho(x_{t_k}, \bar{x}) < r,\ \mathcal{T}_k \text{ sparse}\}$ the set of sparse trees. Similarly as in the proof of Proposition 5, we have $|S| \geq \frac{k_l}{8}$. We now focus only on sparse trees $\mathcal{T}_k$ for $k \in S$ and analyze their relation with the final dataset $\mathcal{D}_{T_l+1}$. Precisely, for a sparse tree $\mathcal{T}_k$, denote $\mathcal{V}_k = \mathcal{T}_k \cap \mathcal{D}_{T_l+1}$ the set of times which are present in the final dataset and belong to the tree induced by error time t_k. Because each node of $\mathcal{T}_k$ and not present in $\mathcal{D}_{T_l+1}$ has at least 1 children in $\mathcal{T}$, we note that $\mathcal{V}_k \neq \emptyset$. We now consider the path from a node of $\mathcal{V}_k$ to the root t_k. We denote by $d(k)$ the depth of this node in $\mathcal{V}_k$ and denote the path by $p^k_{d(k)} \to p^k_{d(k)-1} \to p^k_0 = t_k$ where $p^k_{d(k)} \in \mathcal{V}_k$. Then we have, $d(k) \leq |\mathcal{T}_k| - 1 \leq \frac{16}{\epsilon} - 1$. The same arguments as in the proof of Proposition 5 show that all the points $\{p^k_{d(k)},\ k \in S\}$ fall in distinct sets of the partition $(A_i)_{i \geq 0}$. Therefore,

$$|\{i,\ A_i \cap \boldsymbol{x}_{\leq T_l} \neq \emptyset\}| \geq |S| \geq \frac{k_l}{8} \geq \frac{\epsilon}{16} T_l.$$

Step 2. We now turn to the case when the majority of input points on which 2C1NN made a mistake are not in the ball $B(\bar{x}, r)$ i.e.

$$|\{t \leq t_{k_l},\ \ell_{01}(2C1NN(\boldsymbol{x}_{<t}, \boldsymbol{y}_{<t}, x_t), f^*(x_t)) = 1,\ \rho(x_t, \bar{x}) \geq r\}| \geq \frac{k_l}{2},$$

or equivalently $|\{k \leq k_l,\ \rho(x_{t_k}, \bar{x}) \geq r\}| \geq \frac{k_l}{2}$. Similarly as the previous case, we consider the graph $\tilde{G}$ given by restricting $\mathcal{G}$ only to nodes outside the ball $B(\bar{x}, r)$ i.e. on times $\{t \in \mathcal{T}, \rho(x_t, \bar{x}) \geq r\}$. Again, $\tilde{\mathcal{G}}$ is a collection of disjoint trees with root times $\{t_k,\ k \leq k_l,\ \rho(x_{t_k}, \bar{x}) \geq r\}$—and possibly $t = 1$. We denote $\mathcal{T}_k$ the corresponding tree of $\tilde{\mathcal{G}}$ rooted in t_k. Similarly to above, a tree is *sparse* if

$$\forall t \in \mathcal{T}_k, \quad |\{u \leq T_l,\ \phi(u) = t,\ \rho(x_u, \bar{x}) < r\}| \leq 1 \quad \text{and} \quad |\mathcal{T}_k| \leq \frac{16}{\epsilon}.$$

If $S = \{k \leq k_l, ; \rho(x_{t_k}, \bar{x}) \geq r,\ \mathcal{T}_k \text{ sparse}\}$ denotes the set of sparse trees, the same proof as above shows that $|S| \geq \frac{k_l}{8}$. Again, for any $k \in S$, if $d(k)$ denotes the depth of some node from $\mathcal{V}_k := \mathcal{T}_k \cap \mathcal{D}_{t_{k_l}}$ in $\mathcal{T}_k$ we have $d(k) \leq \frac{16}{\epsilon} - 1$. For each $k \in S$ we consider the path from this node of $\mathcal{V}_k$ to the root t_k: $p^k_{d(k)} \to p^k_{d(k)-1} \to \ldots \to p^k_0 = t_k$ where $p^k_{d(k)} \in \mathcal{V}_k$. The same proof as above shows that all the points $\{p^k_{d(k)},\ k \in S,\ \rho(x_{p^k_{d(k)}}, \bar{x}) > r\}$ lie in distinct sets of the partition $(A_i)_{i \geq 0}$. Suppose $|\{k \in S,\ \rho(x_{p^k_{d(k)}}, \bar{x}) > r\}| \geq \frac{|S|}{2}$, then we have

$$|\{i,\ A_i \cap \boldsymbol{x}_{\leq T_l} \neq \emptyset\}| \geq |\{k \in S,\ \rho(x_{p^k_{d(k)}}, \bar{x}) > r\}| \geq \frac{|S|}{2} \geq \frac{k_l}{16} \geq \frac{\epsilon}{32} T_l.$$

Step 3. In this last step, we suppose again that the majority of input points on which 2C1NN made a mistake are not in the ball $B(\bar{x}, r)$ and that $|\{k \in S,\ \rho(x_{p^k_{d(k)}}, \bar{x}) > r\}| < \frac{|S|}{2}$. Therefore, we obtain

$$|\{k \in S,\ \rho(x_{p^k_{d(k)}}, \bar{x}) = r\}| = |S| - |\{k \in S,\ \rho(x_{p^k_{d(k)}}, \bar{x}) > r\}| \geq \frac{|S|}{2} \geq \frac{k_l}{16} \geq \frac{\epsilon}{32} T_l.$$

We will now make use of the partition $(P_i)_{i\geq 1}$. Recall that $u \geq 1$ was defined such that $n_{u+1} \leq T_l \leq n_{u+2}$. Note that we have $n_u \leq \frac{\epsilon}{2^6} n_{u+1} \leq \frac{\epsilon}{2^6} T_l$. Let us now analyze the process between times n_u and T_l. In particular, we are interested in the indices $T = \{k \in S,\ \rho(x_{p^k_{d(k)}}, \bar{x}) = r\}$ and times $\mathcal{U}_u = \{p^k_{d(k)} :\ n_u < p^k_{d(k)} \leq k_l,\ k \in T\}$. We have

$$|\mathcal{U}_u| \geq |\{k \in S,\ \rho(x_{p^k_{d(k)}}, \bar{x}) = r\}| - n_u \geq \frac{\epsilon}{32} T_l - \frac{\epsilon}{2^6} T_l = \frac{\epsilon}{2^6} T_l.$$

Because the event $\mathcal{E}_u$ is met, we have

$$|\{i,\ P_i(\tau_u) \cap \boldsymbol{x}_{\mathcal{U}_u} \neq \emptyset\}| \leq |\{i,\ P_i(\tau_u) \cap \boldsymbol{x}_{\leq T_l} \neq \emptyset\}| \leq \frac{\epsilon}{2^7} T_l.$$

The same arguments as in the proof of Proposition 5 show that defining $T' := \{k \in T,\ r - \frac{r}{2^{u+2}} \leq \rho(x_{t_k}, \bar{x}) < r\}$ we obtain

$$\left|T'\right| \geq |\mathcal{U}_u| - |\{i,\ P_i(\tau_u) \cap \boldsymbol{x}_{\mathcal{U}_u} \neq \emptyset\}| \geq \frac{\epsilon}{2^7} T_l.$$

We will now show that all the points in $\{x_{t_k},\ k \in T'\}$ lie in distinct sets of $(P_i)_{i\geq 1}$. Note that because we have $T_l \leq n_{u+2}$ and because the event $\mathcal{F}_{u+2}$ is met, we have that for any $p, q \in T'$ that $\rho(x_{\phi(t_p)}, x_{\phi(t_q)}) > \mu_{u+2}$. Now suppose by contradiction that $x_{\phi(t_p)}, x_{\phi(t_q)} \in P_i$ for some $i \geq 1$. Then, with l_i such that $r - \frac{r}{2^{l_i}} \leq \rho(x^i, \bar{x}) < r - \frac{r}{2^{l_i+1}}$ we have that

$$x_{\phi(t_p)}, x_{\phi(t_q)} \in \left\{x \in \mathcal{X} :\ \rho(x, \bar{x}) < r - \frac{r}{2^{l_i+2}}\right\}$$

But we know that $\rho(x_{\phi(t_p)}, \bar{x}) \geq r - \frac{r}{2^{u+2}}$. Therefore we obtain $r - \frac{r}{2^{l_i+2}} > r - \frac{r}{2^{u+2}}$ and hence $l_i \geq u + 1$. Recall that $P_i \subset B(x^i, \mu_{l_i+1})$. Therefore, we obtain $\rho(x_{\phi(t_p)}, x_{\phi(t_q)}) \leq \mu_{l_i+1} \leq \mu_{u+2}$, which contradicts the fact that $\rho(x_{\phi(t_p)}, x_{\phi(t_q)}) > \mu_{u+2}$. This ends the proof that all points of $\{x_{\phi(t_k)},\ k \in T'\}$ lie in distinct subsets of $(P_i)_{i\geq 1}$. Now we obtain

$$|\{i,\ P_i \cap \boldsymbol{x}_{\leq T_l} \neq \emptyset\}| \geq |T'| \geq \frac{\epsilon}{2^7} T_l.$$

Step 4. In conclusion, in all cases, we obtain

$$|\{Q \in \mathcal{Q},\ Q \cap \boldsymbol{x}_{\leq T_l} \neq \emptyset\}| \geq \max(|\{i,\ A_i \cap \boldsymbol{x}_{\leq T_l} \neq \emptyset\}|, |\{i,\ P_i \cap \boldsymbol{x}_{\leq T_l} \neq \emptyset\}|) \geq \frac{\epsilon}{2^7} T_l.$$

Recall that this holds for any realization $\boldsymbol{x}$ in the event $\mathcal{A} \cap \mathcal{E} \cap \bigcap_{l\geq 1} \mathcal{F}_l$. Therefore,

$$\mathbb{E}[|\{Q \in \mathcal{Q},\ Q \cap \mathbb{X}_{\leq T_l} \neq \emptyset\}|] \geq \mathbb{P}\left[\mathcal{A} \cap \mathcal{E} \cap \bigcap_{l\geq 1} \mathcal{F}_l\right] \frac{\epsilon}{2^7} T_l \geq \frac{\epsilon^2}{2^9} T_l.$$

Because this is true for all $l \geq l_0$ and T_l is an increasing sequence, we conclude that $\mathbb{X} \notin \text{WSMV}_{(\mathcal{X},\rho)}$ which is absurd. Therefore 2C1NN is consistent on f^*.

D.2. Proof of Theorem 12

Again, we follow a similar proof to that of Theorem 7. Let $\mathbb{X} \in \text{WSMV}_{(\mathcal{X},\rho)}$ and consider the set $\mathcal{S}_{\mathbb{X}}$ of functions for which it is weakly consistent $\mathcal{S}_{\mathbb{X}} := \{A \in \mathcal{B}, \quad \mathbb{E}\mathcal{L}_{\mathbb{X}}(2C1NN, \mathbb{1}_A) \to 0\}$. By construction we have $\mathcal{S}_{\mathbb{X}} \subset \mathcal{B}$. The goal is to show that in fact $\mathcal{S}_{\mathbb{X}} = \mathcal{B}$. To do so, we will show that $\mathcal{S}$ satisfies the following properties

- $\emptyset \in \mathcal{S}_{\mathbb{X}}$ and $\mathcal{S}_{\mathbb{X}}$ contains all balls $B(x,r)$ with $x \in \mathcal{X}$ and $r \geq 0$,
- if $A \in \mathcal{S}_{\mathbb{X}}$ then $A^c \in \mathcal{S}_{\mathbb{X}}$ (stable to complementary),
- if $(A_i)_{i\geq 1}$ is a sequence of disjoint sets of $\mathcal{S}_{\mathbb{X}}$, then $\bigcup_{i\geq 1} A_i \in \mathcal{S}_{\mathbb{X}}$ (stable to $\sigma-$additivity for disjoint sets),
- if $A, B \in \mathcal{S}_{\mathbb{X}}$, then $A \cup B \in \mathcal{S}_{\mathbb{X}}$ (stable to union).

Together, these properties show that $\mathcal{S}_{\mathbb{X}}$ is a $\sigma-$algebra that contains all open intervals of $\mathcal{X}$. The invariance to complementary is again due to the fact that 2C1NN is invariant to relabeling. Further, we clearly have $\emptyset \in \mathcal{S}_{\mathbb{X}}$. Now let $x \in \mathcal{X}$ and $r \geq 0$, Proposition 5 shows that $B(x,r) \in \mathcal{S}_{\mathbb{X}}$.

We now turn to the $\sigma-$additivity for disjoint sets. Let $(A_i)_{i\geq 1}$ is a sequence of disjoint sets of $\mathcal{S}_{\mathbb{X}}$. We denote $A := \bigcup_{i\geq 1} A_i$. We consider the target function $f^* = \mathbb{1}_A$. We write the average loss in the following way,

$$\frac{1}{T}\sum_{t=1}^{T} \ell_{01}(2C1NN(\mathbb{X}_{<t}, \mathbb{Y}_{<t}, X_t), f^*(X_t)) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{X_t \in A}\mathbb{1}_{X_{\phi(t)} \notin A} + \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{X_t \notin A}\mathbb{1}_{X_{\phi(t)} \in A}.$$

We suppose by contradiction that 2C1NN is not weakly consistent on f^*. Then there exists $\epsilon > 0$ and an increasing sequence of times $(T_l)_{l\geq 1}$ such that $\mathbb{E}\mathcal{L}_{\mathbb{X}}(2C1NN, f^*; T_l) \geq \epsilon T_l$. We first analyze the errors induced by one set A_i only. Simililarly to the proof of Theorem 7 we have

$$\frac{1}{T}\sum_{t=1}^{T}(\mathbb{1}_{X_t \in A_i}\mathbb{1}_{X_{\phi(t)} \notin A} + \mathbb{1}_{X_t \notin A}\mathbb{1}_{X_{\phi(t)} \in A_i}) \leq \frac{1}{T}\sum_{t=1}^{T} \ell_{01}(2C1NN(\mathbb{X}_{<t}, \mathbb{1}_{\mathbb{X}_{<t} \in A_i}, X_t), \mathbb{1}_{X_t \in A_i}).$$

Then, because 2C1NN is consistent for $\mathbb{1}_{.\in A_i}$, we get

$$\mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(\mathbb{1}_{X_t \in A_i}\mathbb{1}_{X_{\phi(t)} \notin A} + \mathbb{1}_{X_t \notin A}\mathbb{1}_{X_{\phi(t)} \in A_i})\right] \to 0.$$

We take $\epsilon_i = \frac{\epsilon}{4\cdot 2^i}$ and T^i such that

$$\forall T \geq T^i, \quad \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}(\mathbb{1}_{X_t \in A_i}\mathbb{1}_{X_{\phi(t)} \notin A} + \mathbb{1}_{X_t \notin A}\mathbb{1}_{X_{\phi(t)} \in A_i})\right] < \frac{\epsilon_i^2}{2}.$$

We now consider the scale of the process $\mathbb{X}_{\leq T^i}$ when falling in A_i, by introducing $\eta_i > 0$ such that

$$\mathbb{P}\left[\min_{\substack{t_1,t_2 \leq T^i;\ X_{t_1},X_{t_2} \in A_i; \\ X_{t_1} \neq X_{t_2}}} \rho(X_{t_1}, X_{t_2}) > \eta_i\right] \geq 1 - \frac{\epsilon_i}{2}.$$

We denote by $\mathcal{F}_i$ this event. Thus, $\mathbb{P}[\mathcal{F}_i^c] \leq \frac{\epsilon_i}{2}$. We now construct a partition $\mathcal{P}$ obtained by subdividing each set A_i according to scale η_i. Because $\mathcal{X}$ is separable, there exists a sequence of points $(x^j)_{j\geq 1}$ in $\mathcal{X}$ such that $\forall x \in \mathcal{X}, \inf_{j\geq 1} \rho(x, x^j) = 0$. We construct the following partition of $\mathcal{X}$ given by

$$\mathcal{P} \ : \quad A^c \cup \bigcup_{i\geq 1}\bigcup_{j\geq 1}\left\{\left(B\left(x^j, \frac{\eta_i}{2}\right)\cap A_i\right) \setminus \bigcup_{k<j} B\left(x^k, \frac{\eta_i}{2}\right)\right\}.$$

We now fix $l \geq 1$ and consider the event $\mathcal{A} := \{\mathcal{L}_{\mathbb{X}}(2C1NN, f^*; T_l) \geq \frac{\epsilon}{2}\}$. Note that

$$\epsilon T_l \leq \mathbb{E}\mathcal{L}_{\mathbb{X}}(2C1NN, f^*; T_l) \leq \frac{\epsilon}{2}T_l + \mathbb{P}[\mathcal{A}]T_l,$$

which gives $\mathbb{P}[\mathcal{A}] \geq \frac{\epsilon}{2}$. We also define the following event

$$\mathcal{E}_i = \left\{\frac{1}{T_l}\sum_{t=1}^{T_l}(\mathbb{1}_{X_t\in A_i}\mathbb{1}_{X_{\phi(t)}\notin A} + \mathbb{1}_{X_t\notin A}\mathbb{1}_{X_{\phi(t)}\in A_i}) < \epsilon_i\right\},$$

for any $i \in I := \{i \geq 1,\ T_l \geq T^i\}$. Then, we have

$$\frac{\epsilon_i^2}{2} \geq \mathbb{E}\left[\frac{1}{T_l}\sum_{t=1}^{T_l}(\mathbb{1}_{X_t\in A_i}\mathbb{1}_{X_{\phi(t)}\notin A} + \mathbb{1}_{X_t\notin A}\mathbb{1}_{X_{\phi(t)}\in A_i})\right] \geq \epsilon_i\mathbb{P}[\mathcal{E}_i^c],$$

which yields $\mathbb{P}[\mathcal{E}_i^c] \leq \frac{\epsilon_i}{2}$. We will now focus on the event $\mathcal{A}\cap\bigcap_{i\in I}\mathcal{E}_i\cap\bigcap_{i\geq 1}\mathcal{F}_i$, which has probability $\mathbb{P}[\mathcal{A}\cap\bigcap_{i\in I}\mathcal{E}_i\cap\bigcap_{i\geq 1}\mathcal{F}_i] \geq \mathbb{P}(\mathcal{A}) - \sum_{i\in I}\mathbb{P}[\mathcal{E}_i^c] - \sum_{i\geq 1}\mathbb{P}[\mathcal{F}_i^c] \geq \frac{\epsilon}{2} - \frac{\epsilon}{4} = \frac{\epsilon}{4}$. Let us now consider a realization of $\boldsymbol{x}$ of $\mathbb{X}$ in the event $\mathcal{A}\cap\bigcap_{i\in I}\mathcal{E}_i\cap\bigcap_{i\geq 1}\mathcal{F}_i$. The sequence $\boldsymbol{x}$ is now not random anymore. We will show that $\boldsymbol{x}$ does visits a linear number of sets in the partition $\mathcal{P}$.

Because the event $\mathcal{A}$ is met, we have

$$\sum_{i\geq 1}\frac{1}{T_l}\sum_{t=1}^{T_l}(\mathbb{1}_{x_t\in A_i}\mathbb{1}_{x_{\phi(t)}\notin A} + \mathbb{1}_{x_t\notin A}\mathbb{1}_{x_{\phi(t)}\in A_i}) \geq \frac{\epsilon}{2}.$$

Also, because the events $\mathcal{E}_i$ are met, we have

$$\sum_{i\in I}\frac{1}{T_l}\sum_{t=1}^{T_l}(\mathbb{1}_{x_t\in A_i}\mathbb{1}_{x_{\phi(t)}\notin A} + \mathbb{1}_{x_t\notin A}\mathbb{1}_{x_{\phi(t)}\in A_i}) \leq \sum_{i\in I}\epsilon_i \leq \frac{\epsilon}{4}.$$

Combining the two above equations gives

$$\frac{1}{T_l}\sum_{t=1}^{T_l}\sum_{i\notin I}(\mathbb{1}_{x_t\in A_i}\mathbb{1}_{x_{\phi(t)}\notin A} + \mathbb{1}_{x_t\notin A}\mathbb{1}_{x_{\phi(t)}\in A_i}) > \frac{\epsilon}{4}. \tag{3}$$

We now consider the set of times such that an input point fell into the set A_i with $i \notin I$, either creating a mistake in the prediction of 4C1NN or inducing a later mistake within time horizon T_l: $\mathcal{T} := \bigcup_{i\notin I}\mathcal{T}_i$ where

$$\mathcal{T}_i := \left\{t \leq T_l,\ x_t \in A_i,\ \left(x_{\phi(t)} \notin A \text{ or } \exists t < u \leq T_l \text{ s.t. } \phi(u) = t,\ x_u \notin A\right)\right\}.$$

Because the events $\mathcal{F}_i$ are met, the same arguments as in the proof of Theorem 7 show that all points x_t for $t \in \mathcal{T}$ fall in distinct sets of the partition $\mathcal{P}$, i.e. $|\{P \in \mathcal{P}, P \cap \boldsymbol{x}_{\leq t_k} \neq \emptyset\}| \geq |\mathcal{T}|$. We also obtain with the same arguments

$$\sum_{t=1}^{t_k} \sum_{i \notin I} (\mathbb{1}_{x_t \in A_i} \mathbb{1}_{x_{\phi(t)} \notin A} + \mathbb{1}_{x_t \notin A} \mathbb{1}_{x_{\phi(t)} \in A_i}) \leq 3|\mathcal{T}|.$$

We now use Equation (3) to obtain $|\{P \in \mathcal{P}, P \cap \boldsymbol{x}_{\leq t_k} \neq \emptyset\}| \geq |\mathcal{T}| \geq \frac{\epsilon}{12} T_l$. Therefore, because this holds for any realization in $\mathcal{A} \cap \bigcap_{i \in I} \mathcal{E}_i \cap \bigcap_{i \geq 1} \mathcal{F}_i$ we obtain

$$\mathbb{E}[|\{P \in \mathcal{P},\ P \cap \mathbb{X}_{\leq T_l} \neq \emptyset\}|] \geq \mathbb{P}\left[\mathcal{A} \cap \bigcap_{i \in I} \mathcal{E}_i \cap \bigcap_{i \geq 1} \mathcal{F}_i\right] \frac{\epsilon}{12} T_l \geq \frac{\epsilon^2}{48} T_l.$$

This holds for any $l \geq 1$. Therefore, because $(T_l)_{l \geq 1}$ is an increasing sequence, this shows that $\mathbb{X} \notin \text{WSMV}_{(\mathcal{X},\rho)}$ which contradicts the hypothesis. This concludes the proof that $A \in \mathcal{S}_{\mathbb{X}}$ and hence, $\mathcal{S}_{\mathbb{X}}$ satisfies the disjoint $\sigma-$additivity property.

We now show that $\mathcal{S}_{\mathbb{X}}$ is invariant to finite unions. Let $A_1, A_2 \in \mathcal{S}_{\mathbb{X}}$. We consider $A = A_1 \cup A_2$ and $f^*(\cdot) = \mathbb{1}_{\cdot \in A}$. Using the same arguments as above, we still have for $T \geq 1$,

$$\mathbb{E}\left[\frac{1}{T} \sum_{t=1}^{T} (\mathbb{1}_{X_t \in A_i} \mathbb{1}_{X_{\phi(t)} \notin A} + \mathbb{1}_{X_t \notin A} \mathbb{1}_{X_{\phi(t)} \in A_i})\right] \to 0.$$

for $i \in \{1, 2\}$. But note that for any $T \geq 1$,

$$\begin{aligned}
\frac{1}{T}\mathcal{L}_{\mathbb{X}}(2C1NN, f^*; T) &= \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}_{X_t \in A} \mathbb{1}_{X_{\phi(t)} \notin A} + \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}_{X_t \notin A} \mathbb{1}_{X_{\phi(t)} \in A} \\
&\leq \frac{1}{T} \sum_{t=1}^{T} (\mathbb{1}_{X_t \in A_1} + \mathbb{1}_{X_t \in A_2}) \mathbb{1}_{X_{\phi(t)} \notin A} + \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}_{X_t \notin A} (\mathbb{1}_{X_{\phi(t)} \in A_1} + \mathbb{1}_{X_{\phi(t)} \in A_2}) \\
&= \sum_{i=1}^{2} \frac{1}{T} \sum_{t=1}^{T} (\mathbb{1}_{X_t \in A_i} \mathbb{1}_{X_{\phi(t)} \notin A} + \mathbb{1}_{X_t \notin A} \mathbb{1}_{X_{\phi(t)} \in A_i}).
\end{aligned}$$

Therefore we obtain directly $\mathbb{E}\left[\frac{1}{T}\mathcal{L}_{\mathbb{X}}(2C1NN, f^*; T)\right] \to 0$. This shows that $A_1 \cup A_2 \in \mathcal{S}_{\mathbb{X}}$ and ends the proof of the theorem.

D.3. Proof of Theorem 13

We fix an output setting $(\mathcal{Y}, \ell)$ and let $\mathbb{X} \in \text{WSMV}_{(\mathcal{X},\rho)}$. We will show that 2C1NN is weakly universally consistent on $\mathbb{X}$ for $(\mathcal{Y}, \ell)$.

We first start by showing that it is weakly universally consistent for classification with countable number of classes $(\mathbb{N}, \ell_{01})$. We fix a target function $f^* : \mathcal{X} \to \mathbb{N}$. For any $i \in \mathbb{N}$ we define the binary function $f_i^* := \mathbb{1}(f^*(\cdot) = i)$. We define

$$\mathcal{L}_i(T) := \sum_{t=1}^{T} \mathbb{1}_{f^*(x_t)=i} \ell_{01}(f^*(x_{\phi(t)}), f^*(x_t))$$

for all $i \geq 0$. Then,

$$\mathcal{L}_i(T) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{f^*(x_t)=i}\ell_{01}(f_i^*(x_{\phi(t)}), f_i^*(x_t)) \leq \mathcal{L}_{\mathbb{X}}(2C1NN, f_i^*; T)$$

Therefore, because 2C1NN is weakly universally consistent, we have $\mathbb{E}\mathcal{L}_{\mathbb{X}}(2C1NN, f_i^*; T) \to 0$, hence $\mathbb{E}\mathcal{L}_i(T) \to 0$ for all $i \geq 0$. Since $\mathcal{L}_i(T) \geq 0$ and $\sum_{i\geq 0} \mathcal{L}_i(T) = \mathcal{L}_{\mathbb{X}}(2C1NN, f^*; T) \leq 1$, we can apply the dominated convergence theorem and obtain

$$\mathbb{E}\mathcal{L}_{\mathbb{X}}(2C1NN, f^*; T) \to 0,$$

which proves that 2C1NN is weakly universally consistent for classification with countable number of classes.

We now turn to the general setting $(\mathcal{Y}, \ell)$. Let $(y^i)_{i\geq 1}$ be a a dense sequence on $\mathcal{Y}$ with respect to ℓ, let $\epsilon > 0$ and consider the function $h(y) := \inf\{i \geq 1 :\ \ell(y^i, y) < \epsilon\}$. Then, we have

$$\begin{aligned}
\ell(y_{\phi(t)}, y_t) &\leq \bar{\ell} \cdot \mathbb{1}_{h(y_{\phi(t)}) \neq h(y_t)} + \ell(y_{\phi(t)}, y_t)\mathbb{1}_{h(y_{\phi(t)})=h(y_t)} \\
&\leq \bar{\ell} \cdot \ell_{01}\mathbb{1}_{h\circ f^*(x_{\phi(t)}) \neq h\circ f^*(x_t)} + c_\ell(\ell(y_{\phi(t)}, y^{h(y_{\phi(t)})}) + \ell(y^{h(y_t)}, y_t)) \\
&\leq \bar{\ell} \cdot \ell_{01}\mathbb{1}_{h\circ f^*(x_{\phi(t)}) \neq h\circ f^*(x_t)} + 2c_\ell\epsilon.
\end{aligned}$$

This yields $\mathcal{L}_{\mathbb{X}}(2C1NN, f^*; T) \leq \bar{\ell}\mathcal{L}_{\mathbb{X}}(2C1NN, h \circ f^*; T) + 2c_\ell\epsilon$. Because 2C1NN is weakly universally consistent for countably-many classification, we have $\mathbb{E}\mathcal{L}_{\mathbb{X}}(2C1NN, h \circ f^*; T) \to 0$. Therefore, we obtain

$$\limsup_T \mathbb{E}\mathcal{L}_{\mathbb{X}}(2C1NN, f^*; T) \leq 2c_\ell\epsilon.$$

This holds for any $\epsilon > 0$ therefore, $\mathbb{E}\mathcal{L}_{\mathbb{X}}(2C1NN, f^*; T) \to 0$, which ends the proof that 2C1NN is weakly universally consistent on $\mathbb{X}$ for the setting $(\mathcal{Y}, \ell)$.