Proceedings of Machine Learning Research vol 178:1–17, 2022 35th Annual Conference on Learning Theory

Universal Online Learning with Bounded Loss: Reduction to Binary Classification

Moïse Blanchard MOISEB@MIT.EDU
Massachusetts Institute of Technology

Romain Cosson COSSON@MIT.EDU
Massachusetts Institute of Technology

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

We study universal consistency of non-i.i.d. processes in the context of online learning. A stochastic process is said to admit *universal consistency* if there exists a learner that achieves vanishing average loss for any measurable response function on this process. When the loss function is unbounded, Blanchard et al. (2022) showed that the only processes admitting strong universal consistency are those taking a finite number of values almost surely. However, when the loss function is bounded, the class of processes admitting strong universal consistency is much richer and its characterization could be dependent on the response setting (Hanneke, 2021b). In this paper, we show that this class of processes is independent from the response setting thereby closing an open question of Hanneke (2021a) (Open Problem 3). Specifically, we show that the class of processes that admit universal online learning is the same for binary classification as for multiclass classification with countable number of classes. Consequently, any output setting with bounded loss can be reduced to binary classification. Our reduction is constructive and practical. Indeed, we show that the *nearest neighbor* algorithm is transported by our construction. For binary classification on a process admitting strong universal learning, we prove that nearest neighbor successfully learns at least all finite unions of intervals.

Keywords: online learning, universal consistency, open problem, statistical learning, invariance, bounded loss, stochastic processes, nearest neighbour

1. Introduction

Problem setup. We consider the online learning framework where we observe a sequence of input points $\mathbb{X} = (X_t)_{t\geq 1}$ from a separable metric *instance space* $(\mathcal{X}, \rho)$ and the associated target values $\mathbb{Y} = (Y_t)_{t\geq 1}$ from some separable near-metric *value space* $(\mathcal{Y}, \ell)$. We assume that the output data stream $\mathbb{Y}$ is generated from $\mathbb{X}$ in a noiseless fashion through an unknown function $f^* : \mathcal{X} \to \mathcal{Y}$, i.e. we have $Y_t = f^*(X_t)$ for all $t \geq 1$. Learning occurs sequentially: at a time step $t \geq 1$, the learner observes a new input X_t (covariates) and outputs a prediction $\hat{Y}_t$ based on the historical data $(\mathbb{X}_{\leq t-1}, \mathbb{Y}_{\leq t-1})$. The performance of the learning rule used by the learner is measured by the long-run average loss $\frac{1}{T}\sum_{t=1}^{T} \ell(Y_t, \hat{Y}_t)$. We say that a learning rule is *universally consistent* under $\mathbb{X}$ if the long-run average loss converges to 0 almost surely, for any measurable function f^*. If such a learning rule exists, we say that $\mathbb{X}$ admits strong universal online learning. Following Hanneke (2021a), we are interested in the set SUOL containing processes $\mathbb{X}$ that admit strong universal online learning. A priori, SUOL may depend on the setup $(\mathcal{X}, \rho, \mathcal{Y}, \ell)$ so we may specify $\text{SUOL}_{(\mathcal{X},\rho,\mathcal{Y},\ell)}$.

© 2022 M. Blanchard & R. Cosson.

Prior work and motivations. In universal learning, the goal is to design learning rules that are consistent with a large variety of data generating processes $(\mathbb{X}, \mathbb{Y})$. A celebrated example, Stone (1977); Devroye et al. (1994) show that the k-nearest neighbour learning rule with $k/\log t \to \infty$ and $k/t \to 0$ is consistent with any i.i.d process under mild hypothesis. More recently, Hanneke et al. (2021); Györfi and Weiss (2021); Cohen and Kontorovich (2022) gave algorithms that are consistent for any i.i.d. process for any metric space $\mathcal{X}$ that admits such an algorithm. In this paper, we primarily focus on *strong* consistency, where we ask the average loss to decay to zero almost surely (Gordon and Olshen, 1978). The literature has also widely investigated *weak* consistency (Müller, 1987), where convergence is in expectation. If randomness or noise is allowed in f^*, consistency is attained when the average loss converges to the loss corresponding to the best deterministic function, i.e. the Bayes loss (Stone, 1977; Devroye et al., 2013). For this reason, universal consistency is sometimes referred to as Bayes risk efficiency (Gordon and Olshen, 1978). For simplicity, this paper assumes that f^* is noiseless and rather focuses on relaxing the assumptions on the input process $\mathbb{X}$.

Indeed, most of the work on universal learning requires the input $\mathbb{X}$ to be drawn i.i.d from a joint distribution (Stone, 1977; Haussler et al., 1994; Hanneke et al., 2021). Alternatively it is asked to be stationary ergodic (Morvai et al., 1996; Györfi and Ottucsák, 2007; Gyöfi and Lugosi, 2002), to satisfy a law of large numbers (Morvai et al., 1996; Steinwart et al., 2009) or to admit convergent relative frequencies (Hanneke, 2021a). Another line of work, (Littlestone, 1988; Ben-David et al., 2009) makes no assumption on the input data stream $\mathbb{X}$ but restricts the hypothesis class to functions f^*, e.g. to functions admitting finite Littlestone dimension. Many other setups have been considered, mixing restrictions on the pair $(\mathbb{X}, f^*)$ (Ryabko and Bartlett, 2006; Urner and Ben-David, 2013; Bousquet et al., 2021).

Following the work of Hanneke (2021a), we make no assumption on the input data $\mathbb{X}$ other than the fact that it is a stochastic process. We are particularly interested in the set $\text{SUOL}_{(\mathcal{X},\rho,\mathcal{Y},\ell)}$ of processes $\mathbb{X}$ that admit strong universal online learning, i.e. such that there exists a learner which achieves vanishing average loss for any choice of measurable function $f^* : \mathcal{X} \to \mathcal{Y}$. When the loss function is unbounded, i.e. $\sup_{y_1,y_2} \ell(y_1, y_2) = \infty$, this set contains exactly the processes that take a finite number of values almost surely (Blanchard et al., 2022) and is therefore independent of the value space $(\mathcal{Y}, \ell)$. When the loss function is bounded, i.e. $\sup_{y_1,y_2} \ell(y_1, y_2) < \infty$, Hanneke (2021a) conjectured that such processes are characterized by a simple condition that we call SMV, standing for sublinear measurable visits, which is also independent of the setting. He posed as an open question whether $\text{SUOL}_{(\mathcal{X},\rho,\mathcal{Y},\ell)}$ would depend on the setting $(\mathcal{Y}, \ell)$ subject to the loss being bounded (Hanneke (2021a), Open Problem 3).

One interest of characterizing the set SUOL is to identify learning rules which are universally consistent for all processes in SUOL, i.e. that achieve universal consistency whenever it is possible (Hanneke, 2021a). These *optimistically universal* learning rules enjoy the convenient property that if they fail to achieve universal learning for a specific input process $\mathbb{X}$, any other online learning rule would fail as well. For unbounded loss, the simple *memorization* learning rule was shown to be optimistically universal (Blanchard et al., 2022) for any setting $(\mathcal{Y}, \ell)$. For bounded loss, an important question—very related to (Open Problem 3 Hanneke (2021a))—is whether the existence of an optimistically universal learning rule depends on the setting $(\mathcal{Y}, \ell)$.

Contributions. We close a conjecture formulated in Hanneke (2021a) by showing that the set of universally learnable sequences SUOL is invariant with respect to the setting $(\mathcal{Y}, \ell)$ when the loss is bounded. Precisely, we show that any learning task can be reduced to the binary classification setting $(\{0,1\}, \ell_{01})$ where ℓ_{01} is the binary indicator loss. Our main result is stated as follows.

Theorem 1 *For any separable near-metric space* $(\mathcal{Y}, \ell)$ *with* $0 < \bar{\ell} < \infty$*, we have* $\mathrm{SUOL}_{(\mathcal{X},\rho,\mathcal{Y},\ell)} = \mathrm{SUOL}_{(\mathcal{X},\rho,\{0,1\},\ell_{01})}$.

This shows that to characterize the set SUOL it suffices to focus on universal binary classification. Our work builds upon Hanneke (2021a) which proves that universal learning can be reduced to either binary classification $(\{0,1\}, \ell_{01})$ or *multiclass* classification with countable number of labels $(\mathbb{N}, \ell_{01})$. Thus, we show the invariance of SUOL to the learning setting by proving that universal binary classification and universal countably-many classes classification are equivalent. Further, our proof is constructive and therefore would provide a construction of an optimistically universal learning rule for any setting $(\mathcal{Y}, \ell)$ given an optimistically universal learning rule for binary classification—if such learning rule exists.

Theorem 2 *The existence of an optimistically universal learning rule is invariant to the output space* $(\mathcal{Y}, \ell)$ *when* $0 < \bar{\ell} < \infty$*. In particular, provided an optimistically universal learning rule for binary classification* $(\{0,1\}, \ell_{01})$ *one can construct an optimistically universal learning rule for a general setup* $(\mathcal{Y}, \ell)$ *with* $0 < \bar{\ell} < \infty$.

Last, we make practical use of this construction to analyze the simple *nearest neighbour* learning rule. In the restricted setting $\mathcal{X} = \mathbb{R}$ we show that for processes that admit strong universal learning, the nearest neighbour learning rule successfully learns functions $f^* : \mathbb{R} \to \{0,1\}$ which represent finite union of intervals i.e. is capable of solving simple classification tasks.

Outline of the paper. The paper is organized as follows. In Section 2 we formally introduce the universal online learning setup and recall some useful results from Hanneke (2021a). We then prove the main reduction theorems and present a class of learning rules that are preserved by this reduction in Section 3. This class includes for instance the nearest neighbor rule. Finally, we focus on this learning rule in Section 4 proving that is consistent for simple classification tasks.

Notations. In the following, ℓ_{01} will denote the indicator loss function $\ell_{01}(i,j) = \mathbb{1}(i \neq j)$ irrespective of the output space $\mathcal{Y}$. Note that it satisfies the relaxed triangle inequality with constant $c_\ell = 1$. When the space $\mathcal{X}$ is clear from the context, we simplify the notation $\mathrm{SUOL}_{(\mathcal{X},\mathcal{Y},\ell)}$ to $\mathrm{SUOL}_{(\mathcal{Y},\ell)}$. We might also omit the loss function ℓ when there is no ambiguity.

2. Background and Preliminaries

2.1. Formal Setup

Instance and value space. Recall that the sequence of inputs $\mathbb{X} = (X_t)_{t\geq 1}$ comes from a separable metric *instance space* $(\mathcal{X}, \rho)$ and the targets $\mathbb{Y} = (Y_t)_{t\geq 1}$ belong to some separable near-metric *value space* $(\mathcal{Y}, \ell)$. The near-metric loss function $\ell : \mathcal{Y}^2 \to [0, \infty)$ is assumed to satisfy symmetry $\ell(y_1, y_2) = \ell(y_2, y_1)$, discernibly $\ell(y_1, y_2) = 0$ if and only if $y_1 = y_2$, as well as 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 finite constant. For instance, the squared loss that is classically used in regression settings satisfies this identity with $c_\ell = 2$. In the following, we will denote by $\bar{\ell} = \sup_{y_1,y_2\in\mathcal{Y}} \ell(y_1, y_2)$ the supremum of the loss function. In particular, the loss function is said to be *bounded* when $\bar{\ell} < \infty$.

Data generation process. The stream of input points $\mathbb{X}$ will be modeled as a general stochastic process with respect to the σ-algebra induced by the metric ρ on $\mathcal{X}$. This differs substantially from most of the statistical learning literature which often imposes additional hypothesis such as being i.i.d., or satisfying the law of large numbers. The stream of output data $\mathbb{Y}$ is assumed to be generated from $\mathbb{X}$ in a noiseless fashion through an unknown fixed measurable function $f^* : \mathcal{X} \to \mathcal{Y}$. Precisely, we have $Y_t = f^*(X_t)$ for all $t \geq 1$. When considering bounded time horizon $t \geq 1$, we will use the following notation: $\mathbb{X}_{\leq t} = \{X_1, ..., X_t\}$ and $\mathbb{X}_{<t} = \{X_1, ..., X_{t-1}\}$.

Online learning. Formally, an online learning rule is defined as a sequence $f. = \{f_t\}_{t=1}^{\infty}$ of measurable functions $f_t : \mathcal{X}^{t-1} \times \mathcal{Y}^{t-1} \times \mathcal{X} \to \mathcal{Y}$. Given $t-1$ training examples of the form $(X_i, f^*(X_i)) \in \mathcal{X} \times \mathcal{Y}$ and a new input sample $X_t \in \mathcal{X}$, the online learning rule f_t makes prediction $f_t(\mathbb{X}_{<t}, \mathbb{Y}_{<t}, X_t)$ for $f^*(X_t)$. We wish to minimize the asymptotic loss,

$$\mathcal{L}_{\mathbb{X}}(f., f^*) = \limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \ell(f_t(\mathbb{X}_{<t}, \mathbb{Y}_{<t}, X_t), f^*(X_t)).$$

We say that the online learning rule $f.$ is consistent under the input process $\mathbb{X}$ and for the target function f^* if $\mathcal{L}_{\mathbb{X}}(f., f^*) = 0 \ \ (a.s.)$.

Processes admitting strong universal online learning. We say that a stochastic process $\mathbb{X}$ admit *strong universal online learning* if there exists a learning rule $\{f_t\}_{t=1}^{\infty}$ that is consistent for all measurable target functions $f^* : \mathcal{X} \to \mathcal{Y}$ on $\mathbb{X}$. We denote by $\text{SUOL}_{(\mathcal{X},\rho,\mathcal{Y},\ell)}$ the set of all processes admitting strong universal online learning. Note that learning rules are allowed to depend on the process $\mathbb{X}$. If a given learning rule is universally consistent under all processes in $\text{SUOL}_{(\mathcal{X},\rho,\mathcal{Y},\ell)}$ we say it is *optimistically universal*.

2.2. Comparing the general setting to binary and countable classification

One of the main contributions of the paper is to show that the set SUOL of input processes $\mathbb{X}$ admitting universal learning is invariant to the choice of value space subject to the loss being bounded. To do so, we compare $\text{SUOL}_{(\mathcal{Y},\ell)}$ for different value spaces $(\mathcal{Y}, \ell)$. Specifically, to show that $\text{SUOL}_{(\mathcal{Y},\ell)} \subset \text{SUOL}_{(\mathcal{Y}',\ell')}$, one aims to construct a universally consistent learning rule for $(\mathcal{Y}', \ell')$ from a universally consistent learning rule for $(\mathcal{Y}, \ell)$ under any fixed process $\mathbb{X} \in \text{SUOL}_{(\mathcal{Y},\ell)}$. In this section, we recall two important known inclusions that hold for any bounded loss setup $(\mathcal{Y}, \ell)$. The first result compares the general setting to binary classification.

Proposition 3 (Hanneke (2021a)) *For any separable near-metric space $(\mathcal{Y}, \ell)$ with $0 < \bar{\ell} < \infty$,*

$$\text{SUOL}_{(\mathcal{Y},\ell)} \subset \text{SUOL}_{(\{0,1\},\ell_{01})}.$$

This shows that binary classification is in essence the easiest setting: whenever universal online learning is achievable for some setting $(\mathcal{Y}, \ell)$, the learning rule that works on this setting should be able to perform binary classification (note that we simply require $\mathcal{Y}$ to contain at least two elements). A formal proof is given in Appendix A, we note that it does not require the boundedness of ℓ.

In the same spirit, we now recall that any process $\mathbb{X}$ admitting strong universal online learning for countable classification $(\mathbb{N}, \ell_{01})$ admits strong universal online learning on any separable value space $(\mathcal{Y}, \ell)$. Hence, countable classification is in essence the hardest setting.

Theorem 4 (Hanneke (2021a)) *For any separable near-metric space $(\mathcal{Y}, \ell)$ with $0 < \bar{\ell} < \infty$,*

$$\text{SUOL}_{(\mathbb{N},\ell_{01})} \subset \text{SUOL}_{(\mathcal{Y},\ell)}.$$

A proof of this theorem is given in (Hanneke (2021a) Theorem 45). It uses a number of intermediary lemmas that are not introduced in this paper. Instead, we provide novel arguments that greatly simplify the proof and that will have practical use in Section 4.

Proof We fix a process $\mathbb{X} \in \text{SUOL}_{(\mathbb{N},\ell_{0,1})}$, and let $f_\cdot^{\mathbb{N}}$ be the corresponding strongly consistent learning rule. By separability, there exists a dense countable sequence $(y^i)_{i\geq 1}$ of $\mathcal{Y}$ i.e. such that $\forall y \in \mathcal{Y} : \inf_{i\in\mathbb{N}} \ell(y^i, y) = 0$. Following Hanneke (2021a), given a prediction task on $(\mathcal{Y}, \ell)$ and $\epsilon > 0$, we reduce it to a countable classification using the function $h_\epsilon : y \in \mathcal{Y} \mapsto \inf\{i \in \mathbb{N} : \ell(y^i, y) < \epsilon\} \in \mathbb{N}$. This allows to define the ϵ-learning rule $f_\cdot^\epsilon$ as follows: given $x_{\leq t} \in \mathcal{X}^t$ and $y_{<t} \in \mathcal{Y}^{t-1}$,

$$f_t^\epsilon(x_{\leq t}, y_{<t}, x_t) = y^{f_t^{\mathbb{N}}(x_{\leq t}, h_\epsilon(y_{<t}), x_t)}.$$

By construction, at each step if the prediction on h_ϵ is successful, the loss of f_t^ϵ is at most ϵ. If the prediction of h_ϵ fails, we can upper bound the loss by $\bar{\ell}$:

$$\ell(f_t^\epsilon(x_{\leq t}, y_{<t}, x_t), y_t) \leq \epsilon + \bar{\ell} \cdot \ell_{01}(f_t^{\mathbb{N}}(x_{\leq t}, h_\epsilon(y)_{<t}, x_t), h_\epsilon(y)_t)$$

where $h_\epsilon(y) := (h_\epsilon(y_t))_{t\geq 1}$. Therefore, for any target measurable function $f^* : \mathcal{X} \to \mathcal{Y}$, we obtain $\mathcal{L}_{\mathbb{X}}^{(\mathcal{Y},\ell)}(f_\cdot^\epsilon, f^*; T) \leq \epsilon + \bar{\ell}\mathcal{L}_{\mathbb{X}}^{(\mathbb{N},\ell_{01})}(f_\cdot^\epsilon, h_\epsilon \circ f^*; T)$, where $\mathcal{L}_{\mathbb{X}}^{(\mathbb{N},\ell_{01})}(f_\cdot^\epsilon, h_\epsilon \circ f^*; T) \to 0 \ (a.s.)$. Unfortunately, using the learning rule $f_\cdot^\epsilon$ only ensures $\mathcal{L}_{\mathbb{X}}^{(\mathcal{Y},\ell)}(f_\cdot^\epsilon, f^*) \leq \epsilon$ almost surely. Thus, the final learning rule will use the learning rules $f_\cdot^{\epsilon_k}$ for a sequence of ϵ_k decreasing to 0 e.g. $\epsilon_k = 2^{-k}$. Intuitively, each learning rule $f_\cdot^{\epsilon_k}$ with prediction y^i effectively predicts that the output y_t belongs to the set $\mathcal{B}_i^{\epsilon_k} := B_\ell(y^i, \epsilon_k) \setminus \bigcup_{1\leq j<i} B_\ell(y^j, \epsilon_k)$ where we used the notation $B_\ell(y, \epsilon) = \{y' \in \mathcal{Y}, \ell(y, y') < \epsilon\}$ for the "ball" induced by the loss ℓ. We now consider the learning rule on $(\mathcal{Y}, \ell)$ denoted $\hat{f}_\cdot^{(\mathcal{Y},\ell)}$ which successively checks consistency of these set predictions $f_\cdot^{\epsilon_1}, f_\cdot^{\epsilon_2}$ etc. and outputs a point $\hat{y} \in \mathcal{Y}$ close to the consistent intersection of these sets. Formally,

$$\hat{f}_t^{(\mathcal{Y},\ell)}(x_{\leq t}, y_{<t}, x_t) = f_t^{\epsilon_{\hat{p}}}(x_{\leq t}, y_{<t}, x_t) \text{ for } \hat{p} = \max\left\{1 \leq p \leq t, \bigcap_{1\leq k\leq p} \mathcal{B}^{\epsilon_k}_{f_t^{\epsilon_k}(x_{\leq t}, y_{<t}, x_t)} \neq \emptyset\right\}.$$

In this definition, the upper bound $\hat{p} \leq t$ is put for simplicity only to ensure that there is a finite maximum. We can now show that this learning rule is universally consistent.

Let $k \geq 1$. Note that if the predictions at step $t \geq k$ of $f_t^{\epsilon_l}$ were correct for all $1 \leq l \leq k$, then the true output y_t belongs to each set prediction $y_t \in \bigcap_{1\leq l\leq k} \mathcal{B}^{\epsilon_l}_{f_t^{\epsilon_l}(x_{\leq t}, y_{<t}, x_t)}$, thus $\hat{p} \geq k$. Now let any $\bar{y} \in \bigcap_{1\leq l\leq \hat{p}} \mathcal{B}^{\epsilon_l}_{f_t^{\epsilon_l}(x_{\leq t}, y_{<t}, x_t)}$, by relaxed triangle inequality we would have

$$\ell(\hat{f}_t^{(\mathcal{Y},\ell)}(x_{\leq t}, y_{<t}, x_t), y_t) \leq c_\ell(\ell(\hat{f}_t^{(\mathcal{Y},\ell)}(x_{\leq t}, y_{<t}, x_t), \bar{y}) + \ell(y_t, \bar{y})) \leq c_\ell(\epsilon_{\hat{p}} + \epsilon_k) \leq 2c_\ell\epsilon_k.$$

Hence,

$$\ell(\hat{f}_t^{(\mathcal{Y},\ell)}(x_{\leq t}, y_{<t}, x_t), y_t) \leq 2c_\ell\epsilon_k + \bar{\ell} \cdot \sum_{l=1}^{k} \ell_{01}(f_t^{\mathbb{N}}(x_{\leq t}, h_{\epsilon_k}(y)_{<t}, x_t), h_{\epsilon_k}(y)_t),$$

and for any measurable function $f^*:\mathcal{X}\to\mathcal{Y}$, we have $\mathcal{L}_{\mathbb{X}}^{(\mathcal{Y},\ell)}(f_\cdot^{(\mathcal{Y},\ell)},f^*)\leq 2c_\ell\epsilon_k$ $(a.s.)$. By union bound, almost surely this holds for any $k\geq 1$ simultaneously. Therefore, almost surely $\mathcal{L}_{\mathbb{X}}^{(\mathcal{Y},\ell)}(f_\cdot^{(\mathcal{Y},\ell)},f^*)=0$ and the learning rule $f_\cdot^{(\mathcal{Y},\ell)}$ is universally consistent. ■

The results of Hanneke (2021a) offer more details that are not required in the rest of the paper but can be found in Appendix D.

2.3. Open problem 3

For any near-metric space $(\mathcal{Y},\ell)$, the inclusions $\text{SUOL}_{(\mathbb{N},\ell_{01})}\subset\text{SUOL}_{(\mathcal{Y},\ell)}\subset\text{SUOL}_{(\{0,1\},\ell_{01})}$ given in Proposition 3 and Theorem 4 do not answer whether $\text{SUOL}_{(\mathcal{Y},\ell_{01})}$ is invariant to the setup when the loss is bounded. The remaining question is whether $\text{SUOL}_{(\{0,1\},\ell_{01})}\subset\text{SUOL}_{(\mathbb{N},\ell_{01})}$ holds or not. We answer positively to this question in the next section, thereby providing a solution to the following open problem.

Open Problem 3 (Hanneke (2021a)): *Is the set* SUOL *invariant to the specification of* $(\mathcal{Y},\ell)$*, subject to* $(\mathcal{Y},\ell)$ *being separable with* $0<\bar{\ell}<\infty$*?*

Remarks on Open Problem 3. In words, the open problem asks whether any universal learning task is achievable whenever universal binary classification is possible. In order to answer affirmatively it would suffice to show that the countable classification setting can be reduced to the binary classification setting. Given a process $\mathbb{X}$ admitting universal learning for binary classification and a countable classification task $f^*:\mathcal{X}\to\mathbb{N}$, a natural idea would be to solve separately each of the binary classification tasks $f^{*,i}(\cdot)=\mathbb{1}(f^*(\cdot)=i)$ for $i\in\mathbb{N}$ and to merge the results together. This proof technique works when f^* takes only a finite number of values, giving rise to the following lemma. Its proof can be found in Appendix B.

Lemma 5 *For any $k\geq 2$,* $\text{SUOL}_{([k],\ell_{01})}=\text{SUOL}_{(\{0,1\},\ell_{01})}$.

Unfortunately, the proof technique used to show that finitely-many classification reduces to binary classification does not extend to countably-many classification. Indeed, the rate of convergence of the average loss on the tasks $f^{*,i}(\cdot)=\mathbb{1}(f^*(\cdot)=i)$ is not uniform across $i\in\mathbb{N}$. Thus, although we can wait for the convergence of a fixed number of these predictors—say the predictions for functions $f^{*,1},\ldots,f^{*,k}$—we do not have any guarantee on the average losses of the predictions for the next functions $f^{*,i}$ for $i>k$. Essentially, we can only guarantee low average loss for a finite number of predictors which use binary classification.

Our proof differs substantially from this approach by considering instead a very large set of predictors—uncountably many. However, we introduce a probability distribution on these predictors, which allows to have guarantees on the average loss for the predictor with high probability on both the stochastic process $\mathbb{X}$ and the predictor. More precisely, instead of learning the individual label i, $f^{*,i}(\cdot)=\mathbb{1}(f^*(\cdot)=i)$, we use predictors of sets of labels $\sigma\in\mathcal{P}(\mathbb{N})$ as follows: $f^*_\sigma(\cdot)=\mathbb{1}(f^*(\cdot)\in\sigma)$. We can now introduce a uniform distribution for the variable σ and test the hypothesis $f^*(x_t)=i$ by analysing the probability (in σ) of the prediction for f^*_σ to be consistent with this hypothesis i.e. $f^*_\sigma(x_t)=1$ if $f^*(x_t)\in\sigma$ and $f^*_\sigma(x_t)=0$ if $f^*(x_t)\notin\sigma$. Intuitively, for the right hypothesis $i^*=y_t$, this probability will be close to 1, while for a wrong hypothesis $i^*\neq y_t$ consistency either results from errors in the predictors, or that both $i,i^*\in\sigma$ or both $i,i^*\notin\sigma$ which happens with probability $1/2$. This discrepancy in probability will allow to discriminate which is the true hypothesis with sublinear number of mistakes.

3. Reduction of countable classification to binary classification

We now present the proof of the main technical result of this paper.

Theorem 6 $\text{SUOL}_{(\{0,1\},\ell_{01})} \subset \text{SUOL}_{(\mathbb{N},\ell_{01})}$.

Proof Suppose you have a process $\mathbb{X} \in \text{SUOL}_{\{0,1\}}$. We want to show that there exists some universal learner for the input process $\mathbb{X}$ and the setting $(\mathbb{N}, \ell_{01})$. Denote by $f_\cdot := \{f_t\}_{t=1}^\infty$ the universal learner in the binary classification setting $(\{0,1\}, \ell_{01})$ for sequence $\mathbb{X}$ and by $f^* : \mathcal{X} \to \mathbb{N}$ the unknown function to learn. For some subsets of outputs $S \subset \mathbb{N}$ we will consider learning the binary valued function $f_S^*(\cdot) = \mathbb{1}(f^*(\cdot) \in S)$.

Specifically, we introduce a random set $\sigma \subset \mathbb{N}$ defined on the product topology of independent Bernoullis. Let $(B_j)_{j\geq 0}$ a sequence of i.i.d. Bernouilli $\mathcal{B}(1/2)$, we define $\sigma = \{j \geq 1 : B_j = 1\}$. Based on learning the functions f_σ^* we now define a statistical test which we will use to define a learning rule for the countable classification. Precisely, given a time $t \geq 0$, define for all $i \in \mathbb{N}$,

$$p^t(x_{<t}, y_{<t}, x_t; i) := \frac{\mathbb{P}_\sigma\left[f_t(x_{<t}, \mathbb{1}(y \in \sigma)_{<t}, x_t) = 1 \mid i \in \sigma\right] + \mathbb{P}_\sigma\left[f_t(x_{<t}, \mathbb{1}(y \in \sigma)_{<t}, x_t) = 0 \mid i \notin \sigma\right]}{2}.$$

where we slightly abuse notations and write $\mathbb{1}(y \in \sigma)$ to denote $(\mathbb{1}(y_t \in \sigma))_{t\geq 1}$. Intuitively, $p^t(\mathbb{X}_{<t}, \mathbb{Y}_{<t}, X_t; i)$ gives the proportion of subsets σ for which the hypothesis $f^*(X_t) = i$ would be consistent with the prediction on the model trained to predict $f_\sigma^*(X_t)$. We first note that although the definition of $p^t(x_{<t}, y_{<t}, x_t; i)$ involves computing expectations over the product measure for σ, its computation can be made practical by considering the values of B_j for observed values j, i.e. $j \in \{y_{t'} : t' < t\} := \mathcal{Y}$. Indeed, we can conveniently write $p^t(x_{<t}, y_{<t}, x_t; i)$ as

$$\begin{aligned} p^t(x_{<t}, y_{<t}, x_t; i) = \frac{1}{2^{|\mathcal{Y}|}} \sum_{(b_j)_{j\in\mathcal{Y}} \in \{0,1\}^{|\mathcal{Y}|}} & \mathbb{P}[f_t(x_{<t}, (b_{y_{t'}})_{t'<t}, x_t) = 1]\mathbb{1}(b_i = 1) \\ & + \mathbb{P}[f_t(x_{<t}, (b_{y_{t'}})_{t'<t}, x_t) = 0]\mathbb{1}(b_i = 0), \end{aligned}$$

where the probability is taken on the possible randomness of the learning rule only. As a result, the function $p^t(\cdot,\cdot,\cdot;\cdot)$ can be practically computed and is also measurable.

Note that if the learning rule $f_\cdot$ had no errors we would have a simple discrimination as follows

$$\frac{\mathbb{P}_\sigma\left[f_\sigma^*(X_t) = 1 \mid i \in \sigma\right] + \mathbb{P}_\sigma\left[f_\sigma^*(X_t) = 0 \mid i \notin \sigma\right]}{2} = \begin{cases} 1 & \text{if } f^*(X_t) = i, \\ 1/2 & \text{otherwise.} \end{cases}$$

We are now ready to define a learning rule $\hat{f} := \{\hat{f}_t\}_{t=1}^\infty$ for countable classification as follows

$$\hat{f}_t(x_{<t}, y_{<t}, x_t) := \begin{cases} \min_{i\in\mathbb{N}} \left\{ i : \; p^t(x_{<t}, y_{<t}, x_t; i) > \dfrac{3}{4} \right\} & \text{if } \exists i \in \mathbb{N},\; p^t(x_{<t}, y_{<t}, x_t; i) > \frac{3}{4} \\ 0 & \text{otherwise.} \end{cases}$$

This is a valid measurable learning rule as a result of the measurability of $p^t(\cdot,\cdot,\cdot;\cdot)$ for all $t \geq 1$. We now show that the learning rule $\hat{f}$ is universally consistent. By hypothesis of binary classification universal consistency, for any subset $S \in \mathcal{P}(\mathbb{N})$, we have $\mathbb{P}_{\mathbb{X}}[\mathcal{L}_{\mathbb{X}}(f., f_S^*; T) \underset{T}{\to} 0] = 1$. Because this result is true for any subset S, we get

$$\mathbb{P}_{\mathbb{X},\sigma}\left[\mathcal{L}_{\mathbb{X}}(f., f_\sigma^*; T) \xrightarrow[T\to\infty]{} 0\right] = 1$$

where the randomness is taken on both $\mathbb{X}$ and σ – and potentially the learning process $f.$. Therefore, we have

$$\mathbb{P}_{\sigma}\left[\mathcal{L}_{\mathbb{X}}(f., f_\sigma^*; T) \xrightarrow[T\to\infty]{} 0\right] = 1, \quad \text{a.s. in } \mathbb{X}$$

Denote by $\mathcal{E}$ this event of probability 1. We will show that on this event, the learning rule is consistent. We now fix an input trajectory $\mathbb{X}$ falling in $\mathcal{E}$ which we denote by $x = (x_t)_{t=0}^\infty$ to make clear that there is no randomness on the trajectory anymore – one can think of a deterministic process. We additionally denote $y = (y_t)_{t=0}^\infty := (f^*(x_t))_{t=0}^\infty$ for simplicity.

By construction, for any $\epsilon > 0$ we have

$$\mathbb{P}_\sigma\left[\mathcal{L}_x(f., f_\sigma^*; t) \leq \epsilon, \quad \forall t \geq T\right] \xrightarrow[T\to\infty]{} 1$$

We can then define for any ϵ a time $T_\epsilon \geq 0$ such that

$$\mathbb{P}_\sigma\left[\mathcal{L}_x(f., f_\sigma^*; t) \leq \epsilon, \quad \forall t \geq T_\epsilon\right] \geq \frac{7}{8}.$$

We define the event $\mathcal{A}_\epsilon = \{\mathcal{L}_x(f., f_\sigma^*; t) \leq \epsilon, \quad \forall t \geq T_\epsilon\}$. An important remark is that both T_ϵ and the event $\mathcal{A}_\epsilon$ are dependent on the specific trajectory x: the learning rate of our rule depends on the realization of the input trajectory. We will show that from time T_ϵ, the error rate of $\hat{f}$ is at most 8ϵ. Let $t \geq 0$ and $i_t^* = f^*(x_t)$ be the true (random) value that we want to predict. We have for the true value i_t^*,

$$\begin{aligned}
p^t(x_{<t}, y_{<t}, x_t; i_t^*) &= 1 - \frac{\mathbb{P}_\sigma\left[f_t(x_{<t}, f_\sigma^*(x_{<t}), x_t) = 0 \mid i_t^* \in \sigma\right] + \mathbb{P}_\sigma\left[f_t(x_{<t}, f_\sigma^*(x_{<t}), x_t) = 1 \mid i_t^* \notin \sigma\right]}{2} \\
&\geq 1 - \mathbb{P}_\sigma[\bar{\mathcal{A}}_\epsilon] - \mathbb{E}_\sigma\left[\left(\mathbb{1}_{f_t(x_{<t}, f_\sigma^*(x_{<t}), x_t)=0, i_t^* \in \sigma} + \mathbb{1}_{f_t(x_{<t}, f_\sigma^*(x_{<t}), x_t)=1, i_t^* \notin \sigma}\right)\mathbb{1}_{\mathcal{A}_\epsilon}\right] \\
&\geq 1 - \frac{1}{8} - \mathbb{E}_\sigma\left[\ell(f_t(x_{<t}, f_\sigma^*(x_{<t}), x_t), f_\sigma^*(x_t))\mathbb{1}_{\mathcal{A}_\epsilon}\right]
\end{aligned}$$

However, for any $i \neq i_t^*$,

$$\begin{aligned}
p^t(x_{<t}, y_{<t}, x_t; i) &= \frac{\mathbb{P}_\sigma\left[f_t(x_{<t}, f_\sigma^*(x_{<t}), x_t) = 1 \mid i \in \sigma\right] + \mathbb{P}_\sigma\left[f_t(x_{<t}, f_\sigma^*(x_{<t}), x_t) = 0 \mid i \notin \sigma\right]}{2} \\
&\leq \frac{1}{2} + \mathbb{E}_\sigma\left[\mathbb{1}_{f_t(x_{<t}, f_\sigma^*(x_{<t}), x_t)=1, i \in \sigma, i_t^* \notin \sigma} + \mathbb{1}_{f_t(x_{<t}, f_\sigma^*(x_{<t}), x_t)=0, i \notin \sigma, i_t^* \in \sigma}\right] \\
&\leq \frac{1}{2} + \mathbb{P}_\sigma[\bar{\mathcal{A}}_\epsilon] + \mathbb{E}_\sigma\left[(\mathbb{1}_{f_t(x_{<t}, f_\sigma^*(x_{<t}), x_t)=1, i \in \sigma, i_t^* \notin \sigma} + \mathbb{1}_{f_t(x_{<t}, f_\sigma^*(x_{<t}), x_t)=0, i \notin \sigma, i_t^* \in \sigma})\mathbb{1}_{\mathcal{A}_\epsilon}\right] \\
&\leq \frac{1}{2} + \frac{1}{8} + \mathbb{E}_\sigma\left[\ell(f_t(x_{<t}, f_\sigma^*(x_{<t}), x_t), f_\sigma^*(x_t))\mathbb{1}_{\mathcal{A}_\epsilon}\right]
\end{aligned}$$

Note that the term $e_t := \mathbb{E}_\sigma\left[\ell(f_t(x_{<t}, f_\sigma^*(x_{<t}), x_t), f_\sigma^*(x_t))\mathbb{1}_{\mathcal{A}_\epsilon}\right]$ is a simple scalar. Therefore, by the previous estimates on p^t, whenever $e_t < \frac{1}{8}$, the learning rule classifies the new input point correctly: $\mathbb{1}_{\hat{f}_t(x_{<t},y_{<t},x_t)\neq i_t^*} \leq \mathbb{1}_{e_t\geq\frac{1}{8}}$. We will now show that the bad event $e_t \geq \frac{1}{8}$ only happens with sublinear rate in t. By construction, in $\mathcal{A}_\epsilon$, for any $t \geq T_\epsilon$,

$$\frac{1}{t}\sum_{u=1}^{t}\ell(f_t(x_{<t}, f^*_\sigma(x_{<t}), x_t), f^*_\sigma(x_t)) \leq \epsilon.$$

Therefore, for any $t \geq T_\epsilon$, we have

$$\frac{1}{t}\sum_{u=1}^{T}e_u = \frac{1}{t}\sum_{u=1}^{t}\mathbb{E}_\sigma\left[\ell(f_t(x_{<t}, f^*_\sigma(x_{<t}), x_t), f^*_\sigma(x_t))\mathbb{1}_{\mathcal{A}_\epsilon}\right] \leq \epsilon.$$

The loss of our learning rule on trajectory x now satisfies for all $t \geq T_\epsilon$,

$$\mathcal{L}_x(\hat{f}_\cdot, f^*; t) = \frac{1}{t}\sum_{u=1}^{t}\mathbb{1}_{\hat{f}_u(x_{<u},y_{<u},x_u)\neq i_u^*} \leq \frac{1}{t}\sum_{u=1}^{t}\mathbb{1}_{e_u\geq\frac{1}{8}} \leq \frac{8}{t}\sum_{u=1}^{t}e_u \leq 8\epsilon.$$

Thus, $\mathcal{L}_x(\hat{f}_\cdot, f^*) \leq 8\epsilon$. Taking $\epsilon > 0$ arbitrarily small shows that $\mathcal{L}_x(\hat{f}_\cdot, f^*) = 0$ and hence, the learning rule is consistent on trajectory x. Therefore, $\hat{f}_\cdot$ is consistent on the event $\mathcal{E}$ for the input sequence $\mathbb{X}$, which has probability 1. To summarize, $\mathcal{L}_\mathbb{X}(\hat{f}_\cdot, f^*) = 0$ $(a.s.)$ for any measurable function f^*, showing that $\hat{f}_\cdot$ is universally consistent and thus $\mathbb{X} \in \text{SUOL}_\mathbb{N}$. This ends the proof of the theorem. ■

Together with Theorem 4, this theorem ends the proof of the main result Theorem 1. Theorem 2 is also a direct consequence from the proof of Theorem 6, Theorem 4 and Proposition 3 since the learning rules were all constructed independently from the stochastic process $\mathbb{X}$. The complete proof is given in Appendix C .

3.1. Learning rules preserved by the reduction

Though its definition is little abstruse, the countable classification learning rule that is derived from the proof of Theorem 4 leaves many learning rules unchanged. In particular, the following proposition shows that learning rules based on a representant which depends only on the historical input sequence e.g. nearest neighbor rule, are transported by our construction.

Proposition 7 *Let $\{f_t\}_{t=1}^\infty$ be a learning rule defined by representant function $\phi(t) \in \{1, ..., t-1\}$ which at step t only depends on $(x_1, .., x_t)$ as follows,*

$$f_t(x_{<t}, y_{<t}, x_t) = y_{\phi(t)}.$$

Note that this learning rule can be defined for any output setting $(\mathcal{Y}, \ell)$. If $\{f_t\}_{t=1}^\infty$ is universally consistent on a process $\mathbb{X}$ for binary classification, it is also universally consistent on $\mathbb{X}$ for any separable near-metric setting $(\mathcal{Y}, \ell)$ with bounded loss.

Proof We first show that the learning rule $f_\cdot = \{f_t\}_{t=1}^\infty$ is transported by our construction in Theorem 6 for classification with countable number of classes. In the rest of the proof, we will denote by $\phi(\cdot)$ the representant function of $f_\cdot$. With

$$p^t(x_{<t}, y_{<t}, x_t; i) := \frac{1}{2}\left(\mathbb{P}_\sigma\left[f_t(x_{<t}, \mathbb{1}(y \in \sigma)_{<t}, x_t) = 1 \mid i \in \sigma\right]\right.$$
$$\left. + \mathbb{P}_\sigma\left[f_t(x_{<t}, \mathbb{1}(y \in \sigma)_{<t}, x_t) = 0 \mid i \notin \sigma\right]\right),$$

we define our learning rule $f^{\mathbb{N}}_\cdot := \{f^{\mathbb{N}}_t\}_{t=1}^\infty$ for countably-many classification as in Theorem 6:

$$f^{\mathbb{N}}_t(x_{<t}, y_{<t}, x_t) := \begin{cases} \min_{i\in\mathbb{N}}\left\{i,\ p^t(x_{<t}, y_{<t}, x_t; i) > \frac{3}{4}\right\} & \text{if } \exists i \in \mathbb{N},\ p^t(x_{<t}, y_{<t}, x_t; i) > \frac{3}{4} \\ 0 & \text{otherwise.} \end{cases}$$

We now show that $f^{\mathbb{N}}_\cdot$ is in fact defined with a similar representant function. Indeed,

$$\begin{aligned} p^t(x_{<t}, y_{<t}, x_t; i) &= \frac{\mathbb{P}_\sigma\left[\mathbb{1}(y_{\phi(t)} \in \sigma) = 1 \mid i \in \sigma\right] + \mathbb{P}_\sigma\left[\mathbb{1}(y_{\phi(t)} \in \sigma) = 0 \mid i \notin \sigma\right]}{2} \\ &= \begin{cases} 1 \text{ if } y_{\phi(t)} = i \\ \frac{1}{2} \text{ if } y_{\phi(t)} \neq i \end{cases} \end{aligned}$$

Therefore, we obtain $f^{\mathbb{N}}_t(x_{<t}, y_{<t}, x_t) = y_{\phi(t)}$, which shows that $f^{\mathbb{N}}_\cdot = f_\cdot$ i.e. that the learning rule $f_\cdot$ is transported by the construction.

We now fix a separable near-metric space $(\mathcal{Y}, \ell)$ and a process $\mathbb{X}$ such that $f_\cdot$ is universally consistent for binary classification. By the above arguments, Theorem 6 shows that $f_\cdot$ is also universally consistent for countable classification. We now aim to show that $f_\cdot$ on $(\mathcal{Y}, \ell)$ is universally consistent on $\mathbb{X}$. Let f^* be a measurable target function and $\epsilon > 0$. We take a sequence $(y^i)_{i\geq 1}$ dense on $\mathcal{Y}$ with respect to ℓ and construct the function $h(y) = \inf\{i \geq 1, \ell(y^i, y) < \epsilon\}$. Then, $y^{f_t(x_{<t}, h_k(y_{<t}), x_t)} = y^{h(y_{\phi(t)})}$. Hence, if $f_t(x_{<t}, h(y_{<t}), x_t) = h(y_t)$ we obtain $y^{h(y_{\phi(t)})} = y^{h(y_t)}$. Therefore, we can write

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

This yields $\mathcal{L}_{\mathbb{X}}(f_\cdot, f^*; T) \leq \bar{\ell}\mathcal{L}_{\mathbb{X}}(f_\cdot, h \circ f^*; T) + 2c_\ell\epsilon$. Because $f_\cdot$ is universally consistent for the setting $(\mathbb{N}, \ell_{01})$, it is in particular consistent for target function $h \circ f^* : \mathcal{X} \to \mathbb{N}$. Therefore, $\limsup_T \mathcal{L}_{\mathbb{X}}(f_\cdot, f^*; T) \leq 2c_\ell\epsilon,\ (a.s.)$. This is valid for $\epsilon_k = 2^{-k}$ for all $k \geq 1$. Therefore, by union bound, $\mathcal{L}_{\mathbb{X}}(f_\cdot, f^*; T) \to 0,\ (a.s.)$, which ends the proof that $f_\cdot$ is universally consistent on $\mathbb{X}$ for the setting $(\mathcal{Y}, \ell)$. ∎

4. Properties of the 1-Nearest Neighbour learning rule

In this section, we will study some interesting properties of the simple nearest neighbour learning rule in the context of strong universal online learning. Formally, we can define $\mathrm{NN} = \{\mathrm{NN}_t\}_{t=1}^\infty$ as follows: for $t > 1$,

$$\mathrm{NN}_t((x_i)_{i<t}, (y_i)_{i<t}, x_t) = y_{\phi(t)} \quad \text{where } \phi(t) = \arg\min_{1\leq i<t} \rho(x_t, x_i).$$

Ties can be broken arbitrarily, for simplicity we split ties in favor of the most ancient closest input point. We will refer to $x_{\phi(t)}$ as the representant of x_t for the nearest neighbor rule. Proposition 3.1 shows that if nearest neighbor is universally consistent for process $\mathbb{X}$ in the binary classification setting, it is also universally consistent for any bounded separable near-metric setting $(\mathcal{Y}, \ell)$. As an immediate consequence of this result, all processes that are i.i.d admit nearest neighbour as an universally consistent learning rule for any near-metric setting. This comes from the universal consistency of nearest neighbour on such processes for binary classification Devroye et al. (2013).

This reduction motivates the analysis of the consistency of nearest neighbour for binary classification. In the rest of this section, we will focus on the specific case $\mathcal{X} = \mathbb{R}$ with classical Euclidian distance ρ as metric. The results presented here can be extended to the d-dimensional euclidean space $\mathbb{R}^d$. We will show that if the input stream $\mathbb{X}$ is in SUOL, the nearest neighbour learning rule is at least able to learn functions that represent a finite union of intervals which are in some sense "simple" functions.

Theorem 8 *For any process $\mathbb{X} \in$ SUOL, the nearest neighbour learning rule is consistent for any finite union of intervals $A = \bigcup_{k=1}^n I_k$ for any arbitrary $n \geq 1$, i.e., for $f^* = \mathbb{1}(\cdot \in A)$ we have $\mathcal{L}_{\mathbb{X}}(NN, f^*, T) \to 0$ $(a.s.)$.*

The proof of this result comes in two steps. First we show that the collections of set that are consistent with the nearest neighbour learning rule is closed by complement and finite union. Second, we show that this collection contains the intervals. Note that in order to prove universal consistency of the nearest neighbour learning rule, we would need to prove that this collection is closed under *countable* union. This is unfortunately beyond the results of this paper.

To build some intuition on the significance of the result, we provide a simple process (deterministic, yet not in SUOL) for which nearest neighbor fails on an interval. Let $\mathcal{X} = [-1, 1]$, $f^* = \mathbb{1}_{[0,1]}$ and $X_t = (-\frac{1}{3})^t$. Then, the nearest neighbor of X_t is X_{t-1} for all $t \geq 1$, inducing an error at each step. On the other hand, SUOL processes do not have this behavior.

Proof We fix a stochastic process $\mathbb{X} \in$ SUOL. Recall that as a consequence $\mathbb{X}$ satisfies condition SMV (Condition 2 in Hanneke (2021a)). This condition states that $\mathbb{X}$ can only makes a sublinear number of visits of different regions of any measurable partition of $\mathcal{X}$. The condition is formally stated as follows.

Condition SMV *The stochastic process $\mathbb{X}$ satisfies condition SMV i.if for every disjoint sequence $\{A_k\}_{k=1}^{\infty}$ in $\mathcal{B}$ with $\bigcup_{k=1}^{\infty} A_k = \mathcal{X}$ (i.e., every countable measurable partition),*

$$\#\{k \in \mathbb{N} : A_k \cap \mathbb{X}_{<T} \neq \emptyset\} = o(T) \quad (a.s).$$

We define $\mathcal{F}_{\mathbb{X}}$ the collection of measurable sets $A \in \mathcal{B}$ for which the nearest neighbour learning rule is consistent on the associated indicator function $\mathbb{1}(\cdot \in A)$. Formally,

$$\mathcal{F}_{\mathbb{X}} = \{A \in \mathcal{B} \mid \mathcal{L}_{\mathbb{X}}(\text{NN}, \mathbb{1}(\cdot \in A); T) \to 0,\ (a.s.)\}.$$

Note that $\mathcal{F}_{\mathbb{X}}$ is stable by complement because the choice of representant in the nearest neighbor rule is independent of the target function: for any measurable set A we have $\mathcal{L}_{\mathbb{X}}(\text{NN}, \mathbb{1}_{\cdot \in A}; T) = \mathcal{L}_{\mathbb{X}}(\text{NN}, 1 - \mathbb{1}_{\cdot \in A}; T) = \mathcal{L}_{\mathbb{X}}(\text{NN}, \mathbb{1}_{\cdot \in A^c}; T)$.

We now show that $\mathcal{F}_{\mathbb{X}}$ is stable by finite union. Let $A_i \in \mathcal{F}_{\mathbb{X}}$ for $i = 1, \ldots, k$. For simplicity we denote $A := \bigcup_{i=1}^k A_i$. Again, because the choice of representant is independent from the target function, if nearest neighbor makes a mistake at time t for target function A, it makes at also a mistake for at least one of the functions A_i, $1 \leq i \leq k$.

$$\begin{aligned}\mathcal{L}_{\mathbb{X}}(\text{NN}, \mathbb{1}(\cdot \in A); T) &= \sum_{t=1}^{T} \mathbb{1}_{x_t \in A} \mathbb{1}_{x_{\phi(t)} \notin A} + \mathbb{1}_{x_t \notin A} \mathbb{1}_{x_{\phi(t)} \in A} \\ &\leq \sum_{t=1}^{T} \sum_{i=1}^{k} 1_{x_t \in A_i} \mathbb{1}_{x_{\phi(t)} \notin A} + \mathbb{1}_{x_t \notin A} \mathbb{1}_{x_{\phi(t)} \in A_i} \\ &= \sum_{i=1}^{k} \mathcal{L}_{\mathbb{X}}(\text{NN}, \mathbb{1}(\cdot \in A_i); T).\end{aligned}$$

Since $A_i \in \mathcal{F}_{\mathbb{X}}$ we have $\mathcal{L}_{\mathbb{X}}(\text{NN}, \mathbb{1}(\cdot \in A_i); T) \to 0, (a.s)$. Therefore, we obtain directly $\mathcal{L}_{\mathbb{X}}(\text{NN}, \mathbb{1}(\cdot \in A); T) \to 0, (a.s)$ i.e. $A \in \mathcal{F}_{\mathbb{X}}$.

We now show that $\mathcal{F}_{\mathbb{X}}$ contains all intervals of the form $(-\infty, a)$ and $(-\infty, a]$ for $a \in \mathbb{R}$. Let $f^* = \mathbb{1}_{\cdot \in (-\infty, a)}$ or $f^* = \mathbb{1}_{\cdot \in (-\infty, a]}$ and consider the following countable partition

$$\mathcal{P} : \quad \{a\} \cup \bigcup_{i \in \mathbb{Z}} [a + 2^i, a + 2^{i+1}) \cup \bigcup_{i \in \mathbb{Z}} (a - 2^{i+1}, a - 2^i].$$

For any $t \geq 1$, let $P_t \in \mathcal{P}$ the set of the partition in which x_t falls. Observe that by construction, if there exists $u < t$ such that $x_u \in P_t$, then nearest neighbor classifies x_t correctly. Indeed, assuming that $x_t > a$, we can write and $P_t = (a + 2^i, a + 2^{i+1}]$. Then we have $|x_{\phi(t)} - x_t| \leq |x_u - x_t|$. Therefore, $x_{\phi(t)} \geq x_t - |x_u - x_t| > a + 2^i - 2^i = a$. Therefore, $y_{\phi(t)} = f^*(x_t)$. The case $x_t < a$ is symmetric, and the case $x_t = a$ is immediate. Thus, if nearest neighbor makes a mistake at time t, the input x_t visited a new set of the partition:

$$\mathcal{L}_{\mathbb{X}}(\text{NN}, f^*; T) \leq \frac{1}{T} |\{k \in \mathbb{N} : A_k \cap \mathbb{X}_{<T} \neq \emptyset\}|.$$

Because $\mathbb{X} \in$ SMV, we can apply the property to the countable partition $\mathcal{P}$ and which yields $\mathcal{L}_{\mathbb{X}}(\text{NN}, f^*; T) \to 0, \ (a.s.)$. This ends the proof of the theorem. ■

5. Conclusion

We resolve an open problem of Hanneke (2021a). We present a novel reduction from a general (separable near-metric) setting to the binary classification setting in the context of universal online learning. This reduction shows that the stochastic processes admitting strong universal consistency for regression are exactly those admitting strong universal consistency for binary classification. Our proof technique is probabilistic but enjoys the property of transporting many natural learning rules such as nearest neighbour. We analyze this particular learning rule in the context of classification for finite union of intervals.

Though the nearest neighbour learning rule has already been extensively studied, there remain interesting questions related to its consistency. For a process $\mathbb{X}$ in SUOL, what is the class of functions f^* for which nearest neighbour achieves strong consistency? In this paper, we showed in the context of binary classification that this class must contain finite unions of intervals, but the general class is possibly much larger. Reciprocally, can we characterize the set of processes for which nearest neighbour is a strong universal online learning rule?

On another note, this paper highlights the importance of the open problems formulated in (Hanneke, 2021b) for the binary classification setting — the existence of an optimistically universal learning rule and the characterization of SUOL. The present paper shows that any solution to these problems would transport from the binary classification setting to the general setting. The authors note that subsequently to this paper, the reduction presented in this work was applied by Blanchard (2022) to obtain optimistically universal learning rules for general metric value spaces.

Acknowledgments

Moïse Blanchard is being partly funded by ONR grant N00014-18-1-2122.

References

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

Moïse Blanchard. Universal online learning: an optimistically universal learning rule. *arXiv preprint arXiv:2201.05947*, 2022.

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.

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.

Dan Tsir Cohen and Aryeh Kontorovich. Metric-valued regression. *arXiv preprint arXiv:2202.03045*, 2022.

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.

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

Louis Gordon and Richard A Olshen. Asymptotically efficient solutions to the classification problem. *The Annals of Statistics*, pages 515–533, 1978.

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

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

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.

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

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, 2021b.

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

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

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

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

H-G Müller. Weak and universal consistency of moving weighted averages. *Periodica Mathematica Hungarica*, 18(3):241–250, 1987.

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

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

Charles J Stone. Consistent nonparametric regression. *The annals of statistics*, pages 595–620, 1977.

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. Proof of Proposition 3

Proof Let $y^0, y^1 \in \mathcal{Y}$ such that $\ell(y^0, y^1) := \delta > 0$. It suffices to observe that measurable functions $\mathcal{X} \to \{0,1\}$ can be mapped to the measurable functions $\mathcal{X} \to \{y^0, y^1\}$ by composing with the simple mapping ϕ such that $\phi(i) = y^i$ for $i \in \{0,1\}$. Consider a sequence $\mathbb{X} \in \text{SUOL}_{(\mathcal{Y},\ell)}$ and let $f.$ be a universal learner for $\mathbb{X}$, we will show that $\mathbb{X} \in \text{SUOL}_{(\{0,1\},\ell_{01})}$ by using this learner to perform binary classification. We define the learning rule $\hat{f}. = (\hat{f}_t)_{t\geq 1}$ as follows, for any $x_{\leq t} \in \mathcal{X}^t$ and $y_{<t} \in \{0,1\}^{t-1}$,

$$\hat{f}_t(x_{<t}, y_{<t}, x_t) := \begin{cases} 0 & \text{if } \ell(f_t(x_{<t}, \phi(y)_{<t}, x_t), y^0) \leq \ell(f_t(x_{<t}, \phi(y)_{<t}, x_t), y^1) \\ 1 & \text{otherwise.} \end{cases}$$

where we used the notation $\phi(y) := (\phi(y_t))_{t\geq 1}$. Note that by relaxed triangle inequality,

$$\begin{aligned}\ell_{01}(\hat{f}_t(x_{<t}, y_{<t}, x_t), y_t) &\leq \mathbb{1}[\ell(f_t(x_{<t}, \phi(y)_{<t}, x_t), \phi(y_t)) \geq \ell(f_t(x_{<t}, \phi(y)_{<t}, x_t), \phi(1-y_t))] \\ &\leq \mathbb{1}[\ell(f_t(x_{<t}, \phi(y)_{<t}, x_t), \phi(y_t)) \geq \frac{c_\ell}{2}\delta] \\ &\leq \frac{2}{c_\ell \delta}\ell(f_t(x_{<t}, \phi(y)_{<t}, x_t), \phi(y_t)).\end{aligned}$$

Then, for any measurable function $f^* : \mathcal{X} \to \{0,1\}$ we have $\mathcal{L}_{\mathbb{X}}^{(\{0,1\},\ell_{01})}(\hat{f}_\cdot, f^*) \leq \frac{2}{c_\ell\delta}\mathcal{L}_{\mathbb{X}}^{(\mathcal{Y},\ell)}(f_\cdot, \phi \circ f^*)$, which by universal consistency of $f_\cdot$ shows that $\mathcal{L}_{\mathbb{X}}^{(\{0,1\},\ell_{01})}(\hat{f}_\cdot, f^*) = 0$ almost surely. Hence, $\hat{f}_\cdot$ is a universal learner for the process $\mathbb{X}$ for the setting $(\{0,1\}, \ell_{01})$ i.e. $\mathbb{X} \in \text{SUOL}_{(\{0,1\},\ell_{01})}$. ■

Appendix B. Proof of Lemma 5

Proof By Proposition 3, it suffices to prove that any process $\mathbb{X} \in \text{SUOL}_{(\{0,1\},\ell_{01})}$ admits universal learning in the setup $([k], \ell_{01})$. To learn an unknown function $f^* : \mathcal{X} \to [k]$, it suffices to learn the k individual binary functions which predict each class: $f^{*,i}(\cdot) := \mathbb{1}(f^*(\cdot) = i)$ where $i \in [k]$. Given a universal learner $f_\cdot$ for $\mathbb{X}$ for binary classification, We can therefore consider a universal learner for $k-$multiclass classification $\hat{f}_\cdot$ which follows the prediction of $f_\cdot$ for all functions f^i as follows: for any $x_{\leq t} \in \mathcal{X}^t$ and $y_{<t} \in [k]^{t-1}$ we pose $\hat{f}_t(x_{<t}, y_{<t}, x_t) := \arg\max_{1\leq i\leq k} f_t(x_{<t}, \mathbb{1}(y=i)_{<t}, x_t)$ where $\mathbb{1}(y = i)_{<t}$ denotes the sequence $\mathbb{1}(y_{t'} = i)_{t'\leq t}$. We can note that this learner makes a mistake only if $f_\cdot$ made a mistake in the prediction of at least one of the functions $f^{*,i}$ for $1 \leq i \leq k$. Thus,

$$\mathcal{L}_{\mathbb{X}}^{([k],\ell_{01})}(\hat{f}_\cdot, f^*) \leq \sum_{i=1}^{k} \mathcal{L}_{\mathbb{X}}^{(\{0,1\},\ell_{01})}(f_\cdot, f^{*,i}).$$

Then, $\mathcal{L}_{\mathbb{X}}^{([k],\ell_{01})}(\hat{f}_\cdot, f^*) = 0$ almost surely by universal consistence of $f_\cdot$ which shows that $\hat{f}_\cdot$ is optimistically universal for $\mathbb{X}$ and $k-$multiclass classification. ■

Appendix C. Proof of Theorem 2

Proof of Theorem 2 We start by supposing that there exists an optimistically universal learning rule $f_\cdot^{\{0,1\}}$ for the binary classification setting, and now construct an optimistically universal learning rule for a general setting $(\mathcal{Y}, \ell)$ satisfying $0 < \bar{\ell} < \infty$. This results from the fact that the construction in the proofs of both Theorem 6 and Theorem 4 are invariant to $\mathbb{X}$. Precisely, we first construct an optimistically universal learning rule for countably-many classification as given in the proof of Theorem 4. With

$$p^t(x_{<t}, y_{<t}, x_t; i) := \frac{1}{2}\left(\mathbb{P}_\sigma\left[f_t^{\{0,1\}}(x_{<t}, \mathbb{1}(y \in \sigma)_{<t}, x_t) = 1 \mid i \in \sigma\right]\right.$$
$$\left.+ \mathbb{P}_\sigma\left[f_t^{\{0,1\}}(x_{<t}, \mathbb{1}(y \in \sigma)_{<t}, x_t) = 0 \mid i \notin \sigma\right]\right),$$

we define

$$f_t^{\mathbb{N}}(x_{<t}, y_{<t}, x_t) := \begin{cases} \min_{i \in \mathbb{N}} \left\{ i,\ p^t(x_{<t}, y_{<t}, x_t; i) > \frac{3}{4} \right\} & \text{if } \exists i \in \mathbb{N},\ p^t(x_{<t}, y_{<t}, x_t; i) > \frac{3}{4} \\ 0 & \text{otherwise.} \end{cases}$$

By construction, and Theorem 6, $f_\cdot^{\mathbb{N}}$ is an optimistically universal learning rule for $(\mathbb{N}, \ell_{01})$. We now use the construction given by Theorem 4 to get an optimistically universal learning rule $f_\cdot^{(\mathcal{Y},\ell)}$ for $(\mathcal{Y}, \ell)$. Define a sequence $(y^i)_{i \geq 1}$ dense in $\mathcal{Y}$ with respect to ℓ. For $k \geq 1$ and $\epsilon_k = 2^{-k}$, we define the functions $h_k(y) = \inf\{i \geq 1, l(y^i, y) < \epsilon_k\}$ and construct the learning rules $f_\cdot^k$ by

$$f_t^k(x_{<t}, y_{<t}, x_t) = y^{f_t^{\mathbb{N}}(x_{<t}, h_k(y_{<t}), x_t)}.$$

Denoting by $B_\ell(y, \epsilon) = \{y' \in \mathcal{Y}, \ell(y, y') < \epsilon\}$ and $\mathcal{B}_i^k := B_\ell(y_i, \epsilon_k) \setminus \bigcup_{1 \leq j < i} B_\ell(y_i, \epsilon_k)$, we now define our final learning rule

$$f_t^{(\mathcal{Y},\ell)}(x_{\leq t}, y_{<t}, x_t) = f_t^{\hat{p}}(x_{\leq t}, y_{<t}, x_t) \text{ for } \hat{p} = \max \left\{ 1 \leq p \leq t, \bigcap_{1 \leq k \leq p} \mathcal{B}^k_{f_t^k(x_{\leq t}, y_{<t}, x_t)} \neq \emptyset \right\},$$

which is invariant to the process $\mathbb{X}$, hence optimistically universal by the proof of Theorem 4.

We now show the converse. Suppose there exists some setting $(\mathcal{Y}, \ell)$ with $0 < \bar{\ell} < \infty$ admitting an optimistically universal learner $f_\cdot^{(\mathcal{Y},\ell)}$. We will construct an optimistically universal learning rule for binary classification using the proof of Proposition 3. Let $y^0, y^1 \in \mathcal{Y}$ such that $\ell(y^0, y^1) > 0$ and consider the function defined by $\phi(i) = y^i$ for $i \in \{0, 1\}$. We now construct a learning rule $f_\cdot^{\{0,1\}}$ for binary classification as follows

$$f_t^{\{0,1\}}(x_{<t}, y_{<t}, x_t) := \begin{cases} 0 & \text{if } \ell(f_t^{(\mathcal{Y},\ell)}(x_{<t}, \phi(y)_{<t}, x_t), y^0) \leq \ell(f_t^{(\mathcal{Y},\ell)}(x_{<t}, \phi(y)_{<t}, x_t), y^1) \\ 1 & \text{otherwise.} \end{cases}$$

This learning rule is invariant to $\mathbb{X}$, hence optimistically universal by the proof of Proposition 3. This ends the proof of the theorem. ■

Appendix D. Additional background

In the core of the paper, we presented the two inclusions $\text{SUOL}_{(\mathbb{N},\ell_{01})} \subset \text{SUOL}_{(\mathcal{Y},\ell)} \subset \text{SUOL}_{(\{0,1\},\ell_{01})}$ shown in Hanneke (2021a) for general bounded loss settings $(\mathcal{Y}, \ell)$ (Prop 3 and Theorem 4). The results of Hanneke (2021a) offer more details which are not useful for this paper but give perspective on previous state of the art as well as useful intuitions. Specifically, the set $\text{SUOL}_{(\mathcal{Y},\ell)}$ only depends on whether the value space $(\mathcal{Y}, \ell)$ is *totally bounded*. We say that $(\mathcal{Y}, \ell)$ is totally bounded if it can be covered by a finite number of $\epsilon-$balls, i.e. $\forall \epsilon > 0, \exists \mathcal{Y}_\epsilon \subset \mathcal{Y}$ s.t. $\#\mathcal{Y}_\epsilon < \infty$ and $\sup_{y \in \mathcal{Y}} \inf_{y \in \mathcal{Y}_\epsilon} \ell(y_\epsilon, y) \leq \epsilon$. Note that $(\{0,1\}, \ell_{01})$ is totally bounded whereas $(\mathbb{N}, \ell_{01})$ is not. Hanneke (2021a) proved that any setup could be reduced to these two cases.

Theorem 9 (Hanneke (2021a)) *For any separable near-metric space $(\mathcal{Y}, \ell)$ with $0 < \bar{\ell} < \infty$,*

- *If $\mathcal{Y}$ is totally bounded,* $\text{SUOL}_{(\mathcal{Y},\ell)} = \text{SUOL}_{(\{0,1\},\ell_{01})}$,
- *If $\mathcal{Y}$ is not totally bounded,* $\text{SUOL}_{(\mathcal{Y},\ell)} = \text{SUOL}_{(\mathbb{N},\ell_{01})}$.

We will now give some intuition on the first point, which reduces the totally bounded setting to $k-$multiclass classification for $k \geq 2$. Finite multiclass classification can then be reduced to binary classification through Lemma 5. It will be useful to keep in mind the proof technique of this reduction for our main result, though it will reveal insufficient to reduce $(\mathbb{N}, \ell_{01})$ to binary classification.

Sketch of proof of Theorem 9. By Theorem 4, we know that for any general setting, $(\mathcal{Y}, \ell)$ we have $\text{SUOL}_{(\mathbb{N},\ell_{01})} \subset \text{SUOL}_{(\mathcal{Y},\ell)}$. The question is now, in which cases can we further reduce the setting to binary classification? Assume that in the construction of the proof of Theorem 4, the partition $(\mathcal{B}_i^\epsilon)_{i\geq 1}$ of $\mathcal{Y}$ into balls of size at most $\epsilon > 0$ can always be made finite. Then, we are able to construct an universally consistent learning rule from universally consistent rules for finitely-many classification, which is equivalent to universal consistence for binary classification by Lemma 5. Thus, we obtain the alternative $\text{SUOL}_{(\mathcal{Y},\ell)} = \text{SUOL}_{(\{0,1\},\ell_{01})}$.

If this is not the case, there exists $\epsilon > 0$ and an infinite—countable— sequence $(y^k)_{k\geq 1}$ in $\mathcal{Y}$ which is $\epsilon-$separated i.e. such that $\ell(y^i, y^j) \geq \epsilon$ for any $i \neq j$. Using the mapping $\phi : \mathbb{N} \to \mathcal{Y}$ defined by $\phi(i) = y^k$ for all $k \geq 1$ similarly to the construction in the proof of Proposition 3, from a universal learner $f.$ for $(\mathcal{Y}, \ell)$ we construct a learning rule $\hat{f}.$ for $(\mathbb{N}, \ell_{01})$, such that for any measurable function $f^* : \mathcal{X} \to \mathbb{N}$,

$$\mathcal{L}_{\mathbb{X}}^{(\mathbb{N},\ell_{01})}(\hat{f}., f^*) \leq \frac{2}{c_\ell \epsilon} \mathcal{L}_{\mathbb{X}}^{(\mathcal{Y},\ell)}(f., \phi \circ f^*),$$

which shows that almost surely, $\mathcal{L}_{\mathbb{X}}^{(\mathbb{N},\ell_{01})}(\hat{f}., f^*) = 0$. Therefore, any sequence which admits universal learning for $(\mathcal{Y}, \ell)$ must admit universal learning for $(\mathbb{N}, \ell_{01})$ i.e. $\text{SUOL}_{(\mathcal{Y},\ell)} \subset \text{SUOL}_{(\mathbb{N},\ell_{01})}$. This ends the alternative of the theorem.