Learning Whenever Learning is Possible: Universal Learning under General Stochastic Processes

Steve Hanneke steve.hanneke@gmail.com

Abstract

This work initiates a general study of learning and generalization without the i.i.d. assumption, starting from first principles. While the standard approach to statistical learning theory is based on assumptions chosen largely for their convenience (e.g., i.i.d. or stationary ergodic), in this work we are interested in developing a theory of learning based only on the most fundamental and natural assumptions implicit in the requirements of the learning problem itself. We specifically study universally consistent function learning, where the objective is to obtain low long-run average loss for any target function, when the data follow a given stochastic process. We are then interested in the question of whether there exist learning rules guaranteed to be universally consistent given *only* the assumption that universally consistent learning is *possible* for the given data process. The reasoning that motivates this criterion emanates from a kind of *optimist's decision theory*, and so we refer to such learning rules as being *optimistically universal*. We study this question in three natural learning settings: *inductive*, *self-adaptive*, and *online*. Remarkably, as our strongest positive result, we find that optimistically universal learning rules do indeed exist in the self-adaptive learning setting. Establishing this fact requires us to develop new approaches to the design of learning algorithms. Along the way, we also identify concise characterizations of the family of processes under which universally consistent learning is possible in the inductive and self-adaptive settings. We additionally pose a number of enticing open problems, particularly for the online learning setting.

Keywords: statistical learning theory, universal consistency, nonparametric estimation, stochastic processes, non-stationary processes, generalization, domain adaptation, online learning

1. Introduction

At least since the time of the ancient Pyrrhonists, it has been observed that learning in general is sometimes not possible. Rather than turning to radical skepticism, modern learning theorists have preferred to introduce constraining assumptions, under which learning becomes possible, and have established positive guarantees for various learning strategies under these assumptions. However, one problem is that the assumptions we have focused on in the literature tend to be assumptions of convenience, simplifying the analysis, rather than assumptions rooted in a principled approach. This is typified by the overwhelming reliance on the assumption that training samples are independent and identically distributed, or resembling this (e.g., stationary ergodic). In the present work, we revisit the issue of the assumptions at the foundations of statistical learning theory, starting from first principles, without relying on assumptions of convenience about the data, such as independence or stationarity.

We approach this via a kind of **optimist's decision theory**, reasoning that if we are tasked with achieving a given objective O in some scenario, then already we have implicitly committed to the assumption that achieving objective O is at least *possible* in that scenario. We may therefore *rely* on this assumption in our strategy for achieving the objective. We are then most interested in strategies guaranteed to achieve objective O in *all* scenarios where it is possible to do so: that is, strategies that rely *only* on the assumption that objective O is achievable. Such strategies have the satisfying property that, if ever they fail to achieve the objective, we may rest assured that no other strategy could have succeeded, so that nothing was lost.

Thus, in approaching the problem of learning (suitably formalized), we may restrict focus to those scenarios in which *learning is possible*. This assumption — that learning is possible — essentially represents a most "natural" assumption, since it is *necessary* for a theory of learning. Concretely, in this work, we initiate this line of exploration by focusing on (arguably) the most basic type of learning problem: *universal consistency* in learning a function. Following the optimist's reasoning above, we are interested in determining whether there exist learning strategies that are *optimistically* universal learners, in the sense that they are guaranteed to be universally consistent given only the assumption that universally consistent learning is *possible* under the given data process: that is, they are universally consistent under all data processes that admit the existence of universally consistent learners. We find that, in certain learning protocols, such optimistically universal learners do indeed exist, and we provide a construction of such a learning rule. Interestingly, it turns out that not all learning rules consistent under the i.i.d. assumption satisfy this type of universality, so that this criterion can serve as an informative desideratum in the design of learning methods. Along the way, we are also interested in expressing concise necessary and sufficient conditions for universally consistent learning to be possible for a given data process.

We specifically consider three natural learning settings — *inductive*, *self-adaptive*, and *online* — distinguished by the level of access to the data available to the learner. In all three settings, we suppose there is an unknown *target function* $f^\star$ and a sequence of data $(X_1,Y_1),(X_2,Y_2),\ldots$ with $Y_t = f^\star(X_t)$, of which the learner is permitted to observe the first n samples $(X_1,Y_1),\ldots,(X_n,Y_n)$: the *training data*. Based on these observations, the learner is tasked with producing a predictor f_n. The performance of the learner is determined by how well $f_n(X_t)$ approximates the (unobservable) Y_t value for data (X_t,Y_t) encountered in the *future* (i.e., $t > n$).[1] To quantify this, we suppose there is a *loss function* ℓ, and we are interested in obtaining a small *long-run average* value of $\ell(f_n(X_t),Y_t)$. A learning rule is said to be *universally consistent* under the process $\{X_t\}$ if it achieves this (almost surely, as $n \to \infty$) for all target functions $f^\star$.[2] The three different settings are then formed as natural variants of this high-level description. The first is the basic *inductive* learning setting, in which f_n is fixed after observing the initial n samples, and we are interested in obtaining a small value of $\frac{1}{m}\sum_{t=n+1}^{n+m} \ell(f_n(X_t),Y_t)$ for all large m. This inductive setting is perhaps the most commonly-studied in the prior literature on statistical learning theory (see

1. Of course, in certain real learning scenarios, these future Y_t values might never actually be observable, and therefore should be considered merely as hypothetical values for the purpose of theoretical analysis of performance.
2. Technically, to be consistent with the terminology used in the literature on universal consistency, we should qualify this as "universally consistent for function learning," to indicate that Y_t is a fixed function of X_t. However, since we do not consider noisy Y_t values or drifting target functions in this work, we omit this qualification and simply write "universally consistent" for brevity.

e.g., Devroye, Györfi, and Lugosi, 1996). The second setting is a more-advanced variant, which we call *self-adaptive* learning, in which f_n may be updated after each subsequent prediction $f_n(X_t)$, based on the additional *unlabeled* observations $X_{n+1}, \ldots, X_t$: that is, it continues to learn from its *test data*. In this case, denoting by $f_{n,t}$ the predictor chosen after observing $(X_1, Y_1), \ldots, (X_n, Y_n), X_{n+1}, \ldots, X_t$, we are interested in obtaining a small value of $\frac{1}{m} \sum_{t=n+1}^{n+m} \ell(f_{n,t-1}(X_t), Y_t)$ for all large m. This setting is related to several others studied in the literature, including *semi-supervised* learning (Chapelle, Schölkopf, and Zien, 2010), *transductive* learning (Vapnik, 1982, 1998), and (perhaps most-importantly) the problems of *domain adaptation* and *covariate shift* (Huang, Smola, Gretton, Borgwardt, and Schölkopf, 2007; Cortes, Mohri, Riley, and Rostamizadeh, 2008; Ben-David, Blitzer, Crammer, Kulesza, Pereira, and Vaughan, 2010). Finally, the strongest setting considered in this work is the *online* learning setting, in which, after each prediction $f_n(X_t)$, the learner is permitted to *observe* Y_t and update its predictor f_n. We are then interested in obtaining a small value of $\frac{1}{m} \sum_{n=0}^{m-1} \ell(f_n(X_{n+1}), Y_{n+1})$ for all large m. This is a particularly strong setting, since it requires that the supervisor providing the Y_t responses remains present in perpetuity. Nevertheless, this is sometimes the case to a certain extent (e.g., in forecasting problems), and consequently the online setting has received considerable attention (e.g., Littlestone, 1988; Haussler, Littlestone, and Warmuth, 1994; Cesa-Bianchi and Lugosi, 2006; Ben-David, Pál, and Shalev-Shwartz, 2009; Rakhlin, Sridharan, and Tewari, 2015).

Our strongest result is for the self-adaptive setting, where we propose a new learning rule and prove that it is universally consistent under *every* data process $\{X_t\}$ for which there exist universally consistent self-adaptive learning rules. As mentioned above, we refer to this property as being *optimistically universal*. Interestingly, we also prove that there is *no* optimistically universal *inductive* learning rule, so that the additional ability to learn from the (unlabeled) test data is crucial. For both inductive and self-adaptive learning, we also prove that the family of processes $\{X_t\}$ that admit the existence of universally consistent learning rules is completely characterized by a simple condition on the tail behavior of empirical frequencies. In particular, this also means that these two families of processes are equal. In contrast, we find that the family of processes admitting the existence of universally consistent *online* learning rules forms a strict *superset* of these other two families. However, beyond this, the treatment of the online learning setting in this work remains incomplete, and leaves a number of enticing open problems regarding whether or not there exist optimistically universal online learning rules, and concisely characterizing the family of processes admitting the existence of universally consistent online learners. In addition to results about learning rules, we also argue that there is no consistent *hypothesis test* for whether a given process admits the existence of universally consistent learners (in any of these settings), indicating that the possibility of learning must indeed be considered an *assumption*, rather than merely a verifiable hypothesis. The above results are all established for general bounded losses. We also discuss the case of unbounded losses, a much more demanding setting for universal learners. In that setting, the theory becomes significantly simpler, and we are able to resolve the essential questions of interest for all three learning settings, with the exception of one particular question on the types of processes that admit the existence of universally consistent learning rules.

1.1 Formal Definitions

We begin our formal discussion with a few basic definitions. Let $(\mathcal{X}, \mathcal{B})$ be a measurable space, with $\mathcal{B}$ a Borel σ-algebra generated by a separable metrizable topological space $(\mathcal{X}, \mathcal{T})$, where $\mathcal{X}$ is called the *instance space* and is assumed to be nonempty. Fix a space $\mathcal{Y}$, called the *value space*, and a function $\ell : \mathcal{Y}^2 \to [0, \infty)$, called the *loss function*. We also denote $\bar{\ell} = \sup\limits_{y,y' \in \mathcal{Y}} \ell(y, y')$. Unless otherwise indicated explicitly, we will suppose $\bar{\ell} < \infty$ (i.e., ℓ is *bounded*); the sole exception to this is Section 8, which is devoted to exploring the setting of unbounded ℓ. Furthermore, to focus on nontrivial scenarios, we will suppose $\mathcal{X}$ and $\mathcal{Y}$ are nonempty and $\bar{\ell} > 0$ throughout.

For simplicity, we suppose that ℓ is a *metric*, and that $(\mathcal{Y}, \ell)$ is a *separable* metric space. For instance, this is the case for discrete classification under the 0-1 loss, or real-valued regression under the absolute loss. However, we note that most of the theory developed here easily extends (with only superficial modifications) to any ℓ that is merely *dominated* by a separable metric ℓ_o, in the sense that $\forall y, y' \in \mathcal{Y}, \ell(y, y') \leq \phi(\ell_o(y, y'))$ for some continuous nondecreasing function $\phi : [0, \infty) \to [0, \infty)$ with $\phi(0) = 0$, and which satisfies a non-triviality condition $\sup_{y_0,y_1} \inf_y \max\{\ell(y, y_0), \ell(y, y_1)\} > 0$. This then admits regression under the squared loss, discrete classification with asymmetric misclassification costs, and many other interesting cases. We include a brief discussion of this generalization in Section 9.1.

Below, any reference to a *measurable set* $A \subseteq \mathcal{X}$ should be taken to mean $A \in \mathcal{B}$, unless otherwise specified. Additionally, let $\mathcal{T}_y$ be the topology on $\mathcal{Y}$ induced by ℓ, and let $\mathcal{B}_y = \sigma(\mathcal{T}_y)$ denote the Borel σ-algebra on $\mathcal{Y}$ generated by $\mathcal{T}_y$; references to measurability of subsets $B \subseteq \mathcal{Y}$ below should be taken to indicate $B \in \mathcal{B}_y$. We will be interested in the problem of learning from data described by a discrete-time stochastic process $\mathbb{X} = \{X_t\}_{t=1}^{\infty}$ on $\mathcal{X}$. We do not make any assumptions about the nature of this process. For any $s \in \mathbb{N}$ and $t \in \mathbb{N} \cup \{\infty\}$, and any sequence $\{x_i\}_{i=1}^{\infty}$, define $x_{s:t} = \{x_i\}_{i=s}^{t}$, or $x_{s:t} = \{\}$ if $t < s$, where $\{\}$ or $\emptyset$ denotes the empty sequence (overloading notation, as these may also denote the empty *set*); for convenience, also define $x_{s:0} = \{\}$. For any function f and sequence $\mathbf{x} = \{x_i\}_{i=1}^{\infty}$ in the domain of f, we denote $f(\mathbf{x}) = \{f(x_i)\}_{i=1}^{\infty}$ and $f(x_{s:t}) = \{f(x_i)\}_{i=s}^{t}$. Also, for any set $A \subseteq \mathcal{X}$, we denote by $x_{s:t} \cap A$ or $A \cap x_{s:t}$ the subsequence of all entries of $x_{s:t}$ contained in A, and $|x_{s:t} \cap A|$ denotes the number of indices $i \in \mathbb{N} \cap [s, t]$ with $x_i \in A$.

For any function $g : \mathcal{X} \to \mathbb{R}$, and any sequence $\mathbf{x} = \{x_t\}_{t=1}^{\infty}$ in $\mathcal{X}$, define

$$\hat{\mu}_{\mathbf{x}}(g) = \limsup_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} g(x_t).$$

For any set $A \subseteq \mathcal{X}$ we overload this notation, defining $\hat{\mu}_{\mathbf{x}}(A) = \hat{\mu}_{\mathbf{x}}(\mathbb{1}_A)$, where $\mathbb{1}_A$ is the binary indicator function for the set A. We also use the notation $\mathbb{1}[p]$, for any logical proposition p, to denote a value that is 1 if p holds (evaluates to "True"), and 0 otherwise. We also make use of the standard notation for limits of sequences $\{A_i\}_{i=1}^{\infty}$ of sets (see e.g., Ash and Doléans-Dade, 2000): $\limsup\limits_{i \to \infty} A_i = \bigcap\limits_{k=1}^{\infty} \bigcup\limits_{i=k}^{\infty} A_i$, $\liminf\limits_{i \to \infty} A_i = \bigcup\limits_{k=1}^{\infty} \bigcap\limits_{i=k}^{\infty} A_i$, and $\lim\limits_{i \to \infty} A_i$ exists and equals $\limsup\limits_{i \to \infty} A_i$ if and only if $\limsup\limits_{i \to \infty} A_i = \liminf\limits_{i \to \infty} A_i$. Additionally, for a set A, a function $T : A \to \mathbb{R}$ bounded from below, and a value $\varepsilon > 0$, define $\operatorname*{argmin}^{\varepsilon}_{a \in A} T(a)$ as an arbitrary element $a^* \in A$ with $T(a^*) \leq \inf_{a \in A} T(a) + \varepsilon$; we also allow $\varepsilon = 0$ in this definition in the case $\inf_{a \in A} T(a)$ is realized by some $T(a)$, $a \in A$; to be clear, we suppose $\operatorname{argmin}^{\varepsilon}_{a \in A} T(a)$ evaluates to the *same* a^* every time it appears (for a given function T and set A).

As discussed above, we are interested in three learning settings, defined as follows. An *inductive* learning rule is any sequence of measurable functions $f_n : \mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{X} \to \mathcal{Y}$, $n \in \mathbb{N} \cup \{0\}$. A *self-adaptive* learning rule is any array of measurable functions $f_{n,m} : \mathcal{X}^m \times \mathcal{Y}^n \times \mathcal{X} \to \mathcal{Y}$, $n, m \in \mathbb{N} \cup \{0\}$, $m \geq n$. An *online* learning rule is any sequence of measurable functions $f_n : \mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{X} \to \mathcal{Y}$, $n \in \mathbb{N} \cup \{0\}$. In each case, these functions can potentially be stochastic (that is, we allow f_n itself to be a random variable), though independent from $\mathbb{X}$. For any measurable $f^\star : \mathcal{X} \to \mathcal{Y}$, any inductive learning rule f_n, any self-adaptive learning rule $g_{n,m}$, and any online learning rule h_n, we define

$$\hat{\mathcal{L}}_{\mathbb{X}}(f_n, f^\star; n) = \limsup_{t \to \infty} \frac{1}{t} \sum_{m=n+1}^{n+t} \ell\left(f_n(X_{1:n}, f^\star(X_{1:n}), X_m), f^\star(X_m)\right),$$

$$\hat{\mathcal{L}}_{\mathbb{X}}(g_{n,\cdot}, f^\star; n) = \limsup_{t \to \infty} \frac{1}{t+1} \sum_{m=n}^{n+t} \ell(g_{n,m}(X_{1:m}, f^\star(X_{1:n}), X_{m+1}), f^\star(X_{m+1})),$$

$$\hat{\mathcal{L}}_{\mathbb{X}}(h_\cdot, f^\star; n) = \frac{1}{n} \sum_{t=0}^{n-1} \ell(h_t(X_{1:t}, f^\star(X_{1:t}), X_{t+1}), f^\star(X_{t+1})).$$

In each case, $\hat{\mathcal{L}}_{\mathbb{X}}(\cdot, f^\star; n)$ measures a kind of limiting loss of the learning rule, relative to the source of the target values: $f^\star$. In this context, we refer to $f^\star$ as the *target function*. Note that, in the cases of inductive and self-adaptive learning rules, we are interested in the average *future* losses after some initial number n of "training" observations, for which target values are provided, and after which no further target values are observable. Thus, a small value of the loss $\hat{\mathcal{L}}_{\mathbb{X}}$ in these settings represents a kind of *generalization* to future (possibly previously-unseen) data points. In particular, in the special case of i.i.d. $\mathbb{X}$ with marginal distribution $\mathbb{P}_X$, the strong law of large numbers implies that the loss $\hat{\mathcal{L}}_{\mathbb{X}}(f_n, f^\star; n)$ of an inductive learning rule f_n is equal (almost surely) to the usual notion of the *risk* of $f_n(X_{1:n}, f^\star(X_{1:n}), \cdot)$ — namely, $\int \ell(f_n(X_{1:n}, f^\star(X_{1:n}), x), f^\star(x)) \mathbb{P}_X(\mathrm{d}x)$ — commonly studied in the statistical learning theory literature, so that $\hat{\mathcal{L}}_{\mathbb{X}}(f_n, f^\star; n)$ represents a generalization of the notion of *risk* (for deterministic responses). Note that, in the general case, the average loss $\frac{1}{t} \sum_{m=n+1}^{n+t} \ell(f_n(X_{1:n}, f^\star(X_{1:n}), X_m), f^\star(X_m))$ might not have a well-defined limit as $t \to \infty$, particularly for *non-stationary* processes $\mathbb{X}$, and it is for this reason that we use the limit superior in the definition (and similarly for $\hat{\mathcal{L}}_{\mathbb{X}}(g_{n,\cdot}, f^\star; n)$). We also note that, since the loss function is always finite, we could have included the losses on the training samples in the summation in the inductive $\hat{\mathcal{L}}_{\mathbb{X}}(f_n, f^\star; n)$ definition without affecting its value. This observation yields a convenient simplification of the definition, as it implies the following equality.

$$\hat{\mathcal{L}}_{\mathbb{X}}(f_n, f^\star; n) = \hat{\mu}_{\mathbb{X}}(\ell(f_n(X_{1:n}, f^\star(X_{1:n}), \cdot), f^\star(\cdot))).$$

The distinction between the inductive and self-adaptive settings is merely the fact that the self-adaptive learning rule is able to *update* the function used for prediction after observing each "test" point X_t, $t > n$. Note that the target values are not available for these test points: only the "unlabeled" X_t values. In the special case of an i.i.d. process, the self-adaptive setting is closely related to the *semi-supervised* learning setting studied in the statistical learning theory literature (Chapelle, Schölkopf, and Zien, 2010). In the case of non-stationary processes, it has relations to problems of *domain adaptation* and *covariate shift* (Huang, Smola, Gretton, Borgwardt, and Schölkopf, 2007; Cortes, Mohri, Riley, and Rostamizadeh, 2008; Ben-David, Blitzer, Crammer, Kulesza, Pereira, and Vaughan, 2010).

In the case of online learning, the prediction function is again allowed to update after every test point, but in this case the target value for the test point *is* accessible (after the prediction is made). This online setting, with precisely this same $\hat{\mathcal{L}}_{\mathbb{X}}(h_\cdot, f^\star; n)$ objective function, has been studied in the learning theory literature, both in the case of i.i.d. processes and relaxations thereof (e.g., Haussler, Littlestone, and Warmuth, 1994; Györfi, Kohler, Krzyżak, and Walk, 2002) and in the very-general setting of $\mathbb{X}$ an *arbitrary* process (e.g., Littlestone, 1988; Cesa-Bianchi and Lugosi, 2006; Rakhlin, Sridharan, and Tewari, 2015).

Our interest in the present work is the basic problem of *universal consistency*, wherein the objective is to design a learning rule with the guarantee that the long-run average loss $\hat{\mathcal{L}}_{\mathbb{X}}$ approaches *zero* (almost surely) as the training sample size n grows large, and that this fact holds true for *any* target function $f^\star$. Specifically, we have the following definitions.

Definition 1 *We say an inductive learning rule f_n is strongly universally consistent under $\mathbb{X}$ if, for every measurable $f^\star : \mathcal{X} \to \mathcal{Y}$, $\lim_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(f_n, f^\star; n) = 0$ (a.s.).*
We say a process $\mathbb{X}$ admits strong universal inductive learning if there exists an inductive learning rule f_n that is strongly universally consistent under $\mathbb{X}$.
We denote by SUIL *the set of all processes $\mathbb{X}$ that admit strong universal inductive learning.*

Definition 2 *We say a self-adaptive learning rule $f_{n,m}$ is strongly universally consistent under $\mathbb{X}$ if, for every measurable $f^\star : \mathcal{X} \to \mathcal{Y}$, $\lim_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(f_{n,\cdot}, f^\star; n) = 0$ (a.s.).*
We say a process $\mathbb{X}$ admits strong universal self-adaptive learning if there exists a self-adaptive learning rule $f_{n,m}$ that is strongly universally consistent under $\mathbb{X}$.
We denote by SUAL *the set of all processes $\mathbb{X}$ that admit strong universal self-adaptive learning.*

Definition 3 *We say an online learning rule f_n is strongly universally consistent under $\mathbb{X}$ if, for every measurable $f^\star : \mathcal{X} \to \mathcal{Y}$, $\lim_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(f_\cdot, f^\star; n) = 0$ (a.s.).*
We say a process $\mathbb{X}$ admits strong universal online learning if there exists an online learning rule f_n that is strongly universally consistent under $\mathbb{X}$.
We denote by SUOL *the set of all processes $\mathbb{X}$ that admit strong universal online learning.*

Technically, the above definitions of universal consistency are defined relative to the loss function ℓ. However, we will establish below that SUIL and SUAL are in fact *invariant* to the choice of $(\mathcal{Y}, \ell)$, subject to the basic assumptions stated above (separable, $0 < \bar{\ell} < \infty$). We will also find that this is true of SUOL, subject to the additional constraint that $(\mathcal{Y}, \ell)$ is *totally bounded*. Furthermore, for unbounded losses we find that all three families are invariant to $(\mathcal{Y}, \ell)$, subject to separability and $\bar{\ell} > 0$.

As noted above, much of the prior literature on universal consistency without the i.i.d. assumption has focused on relaxations of the i.i.d. assumption to more-general families of processes, such as stationary mixing, stationary ergodic, or certain limited forms of non-stationarity (see e.g., Steinwart, Hush, and Scovel, 2009, Chapter 27 of Györfi, Kohler, Krzyżak, and Walk, 2002, and references therein). In each case, these relaxations were chosen largely for their *convenience*, as they preserve the essential features of the i.i.d. setting used in the traditional approaches to proving consistency of certain learning rules (particularly, features related to concentration of measure). In contrast, our primary interest in the present work is to study the *natural* assumption *intrinsic to the universal consistency problem itself*: the assumption that universal consistency is *possible*. In other words, we are interested in the following abstract question:

Do there exist learning rules that are strongly universally consistent under *every* process $\mathbb{X}$ that admits strong universal learning?

Each of the three learning settings yields a concrete instantiation of this question. For the reason discussed in the introductory remarks, we refer to any such learning rule as being *optimistically universal*. Thus, we have the following definition.

Definition 4 *An (inductive/self-adaptive/online) learning rule is* optimistically universal *if it is strongly universally consistent under every process* $\mathbb{X}$ *that admits strong universal (inductive/self-adaptive/online) learning.*

1.2 Summary of the Main Results

Here we briefly summarize the main results of this work. Their proofs, along with several other results, will be developed throughout the rest of this article.

The main positive result in this work is the following theorem, which establishes that optimistically universal self-adaptive learning is indeed possible. In fact, in proving this result, we develop a specific construction of one such self-adaptive learning rule.

Theorem 5 *There exists an optimistically universal self-adaptive learning rule.*

Interestingly, it turns out that the additional capabilities of self-adaptive learning, compared to inductive learning, are actually *necessary* for optimistically universal learning. This is reflected in the following result.

Theorem 6 *There* does not exist *an optimistically universal inductive learning rule, if* $(\mathcal{X}, \mathcal{T})$ *is an uncountable Polish space.*

Taken together, these two results are interesting indeed, as they indicate there can be strong advantages to designing learning methods to be self-adaptive. This seems particularly interesting when we note that very few learning methods in common use are designed to exploit this capability: that is, to adjust their trained predictor based on the (unlabeled) test samples they encounter. In light of these results, it therefore seems worthwhile to revisit the definitions of these methods with a view toward designing self-adaptive variants.

As for the online learning setting, the present work makes only partial progress toward resolving the question of the existence of optimistically universal online learning rules (in Section 6). In particular, the following question remains open at this time.

Open Problem 1 *Does there exist an optimistically universal online learning rule?*

To be clear, as we discuss in Section 6, one can convert the optimistically universal self-adaptive learning rule from Theorem 5 into an online learning rule that is strongly universally consistent for any process $\mathbb{X}$ that admits strong universal *self-adaptive* learning. However, as we prove below, the set of processes $\mathbb{X}$ that admit strong universal *online* learning is a strict superset of these, and so optimistically universal online learning represents a much stronger requirement for the learner.

In the process of studying the above, we also investigate the problem of concisely *characterizing* the family of processes that admit strong universal learning, of each of the three types: that is, SUIL, SUAL, and SUOL. In particular, consider the following simple condition on the tail behavior of a given process $\mathbb{X}$.

Condition 1 *For every monotone sequence $\{A_k\}_{k=1}^{\infty}$ of sets in $\mathcal{B}$ with $A_k \downarrow \emptyset$,*

$$\lim_{k\to\infty} \mathbb{E}[\hat{\mu}_{\mathbb{X}}(A_k)] = 0.$$

Denote by $\mathcal{C}_1$ the set of all processes $\mathbb{X}$ satisfying Condition 1. In Section 2 below, we discuss this condition in detail, and also provide several equivalent forms of the condition. One interesting instance of this is Theorem 12, which notes that Condition 1 is equivalent to the condition that the set function $\mathbb{E}[\hat{\mu}_{\mathbb{X}}(\cdot)]$ is a *continuous submeasure* (Definition 10 below). For our present interest, the most important fact about Condition 1 is that it precisely identifies which processes $\mathbb{X}$ admit strong universal inductive or self-adaptive learning, as the following theorem states.

Theorem 7 *The following statements are equivalent for any process $\mathbb{X}$.*

- *$\mathbb{X}$ satisfies Condition 1.*
- *$\mathbb{X}$ admits strong universal inductive learning.*
- *$\mathbb{X}$ admits strong universal self-adaptive learning.*

Equivalently, $\text{SUIL} = \text{SUAL} = \mathcal{C}_1$.

Certainly any i.i.d. process satisfies Condition 1 (by the strong law of large numbers). Indeed, we argue in Section 3.1 that any process satisfying the law of large numbers — or more generally, having pointwise convergent relative frequencies — satisfies Condition 1, and hence by Theorem 7 admits strong universal learning (in both settings). For instance, this implies that *all stationary processes* admit strong universal inductive and self-adaptive learning. However, as we also demonstrate in Section 3.1, there are many other types of processes, which do not have convergent relative frequencies, but which do satisfy Condition 1, and hence admit universal learning, so that Condition 1 represents a strictly more-general condition.

Other than the fact that Condition 1 precisely characterizes the families of processes that admit strong universal inductive or self-adaptive learning, another interesting fact established by Theorem 7 is that these two families are actually *equivalent*: that is, SUIL =

SUAL. Interestingly, as alluded to above, we find that this equivalence does *not* extend to *online* learning. Specifically, in Section 6 we find that SUAL $\subseteq$ SUOL, with *strict* inclusion iff $\mathcal{X}$ is infinite.

As for the problem of concisely characterizing the family of processes that admit strong universal *online* learning, again the present work only makes partial progress. Specifically, in Section 6, we formulate a concise *necessary* condition for a process $\mathbb{X}$ to admit strong universal online learning (Condition 2 below), but we leave open the important question of whether this condition is also *sufficient*, or more-broadly of identifying a concise condition on $\mathbb{X}$ equivalent to the condition that $\mathbb{X}$ admits strong universal online learning.

In addition to the questions of optimistically universal learning and concisely characterizing the family of processes admitting universal learning, another interesting question is whether it is possible to empirically *test* whether a given process admits universal learning (of any of the three types). However, in Section 7 we find that in all three settings this is *not* the case. Specifically, in Theorem 43 we prove that (when $\mathcal{X}$ is infinite) there does not exist a consistent hypothesis test for whether a given $\mathbb{X}$ admits strong universal (inductive/self-adaptive/online) learning. Hence, the assumption that learning is possible truly is an *assumption*, rather than a testable hypothesis.

While all of the above results are established for *bounded* losses, Section 8 is devoted to the study of these same issues in the case of *unbounded* losses. In that case, the theory becomes significantly simplified, as universal consistency is much more difficult to achieve, and hence the family of processes that admit universal learning is severely restricted. We specifically find that, when the loss is unbounded, there exists an optimistically universal learning rule of *all three* types. We also identify a concise condition (Condition 3 below) that is necessary and sufficient for a process to admit strong universal learning in any/all of the three settings.

We discuss extensions of this theory in Section 9, discussing more-general loss functions, as well as relaxation of the requirement of *strong* consistency to mere *weak* consistency. Finally, we conclude the article in Section 10 by summarizing several interesting open questions that arise from the theory developed below.

2. Equivalent Expressions of Condition 1

Before getting into the analysis of learning, we first discuss basic properties of the $\hat{\mu}_{\mathbf{x}}$ functional. In particular, we find that there are several equivalent ways to state Condition 1, which will be useful in various parts of the proofs below, and which may themselves be of independent interest in some cases.

2.1 Basic Lemmas

We begin by stating some basic properties of the $\hat{\mu}_{\mathbf{x}}$ functional that will be indispensable in the proofs below.

Lemma 8 *For any sequence $\mathbf{x} = \{x_t\}_{t=1}^{\infty}$ in $\mathcal{X}$, and any functions $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{X} \to \mathbb{R}$, if $\hat{\mu}_{\mathbf{x}}(f)$ and $\hat{\mu}_{\mathbf{x}}(g)$ are not both infinite and of opposite signs, then the following* *properties hold.*

1. *(monotonicity)* *if* $f \leq g$, *then* $\hat{\mu}_{\mathbf{x}}(f) \leq \hat{\mu}_{\mathbf{x}}(g)$,
2. *(homogeneity)* $\forall c \in (0, \infty), \hat{\mu}_{\mathbf{x}}(cf) = c\hat{\mu}_{\mathbf{x}}(f)$,
3. *(subadditivity)* $\hat{\mu}_{\mathbf{x}}(f + g) \leq \hat{\mu}_{\mathbf{x}}(f) + \hat{\mu}_{\mathbf{x}}(g)$.

Proof Properties 1 and 2 follow directly from the definition of $\hat{\mu}_{\mathbf{x}}$, and monotonicity and homogeneity (for positive constants) of lim sup. Property 3 is established by noting

$$\begin{aligned}\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n}(f(x_t) + g(x_t)) &\leq \lim_{k\to\infty}\left(\sup_{n\geq k}\frac{1}{n}\sum_{t=1}^{n}f(x_t)\right) + \left(\sup_{n\geq k}\frac{1}{n}\sum_{t=1}^{n}g(x_t)\right) \\ &= \left(\limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}f(x_t)\right) + \left(\limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}g(x_t)\right).\end{aligned}$$

∎

These properties immediately imply related properties for the *set function* $\hat{\mu}_{\mathbf{x}}$.

Lemma 9 *For any sequence* $\mathbf{x} = \{x_t\}_{t=1}^{\infty}$ *in* $\mathcal{X}$, *and any sets* $A, B \subseteq \mathcal{X}$,

1. *(nonnegativity)* $0 \leq \hat{\mu}_{\mathbf{x}}(A)$,
2. *(monotonicity)* $\hat{\mu}_{\mathbf{x}}(A \cap B) \leq \hat{\mu}_{\mathbf{x}}(A)$,
3. *(subadditivity)* $\hat{\mu}_{\mathbf{x}}(A \cup B) \leq \hat{\mu}_{\mathbf{x}}(A) + \hat{\mu}_{\mathbf{x}}(B)$.

Proof These follow directly from the properties listed in Lemma 8, since $0 \leq \mathbb{1}_A$, $\mathbb{1}_{A\cap B} \leq \mathbb{1}_A$, and $\mathbb{1}_{A\cup B} \leq \mathbb{1}_A + \mathbb{1}_B$. ∎

2.2 An Equivalent Expression in Terms of Continuous Submeasures

Next, we note a connection to a much-studied definition from the measure theory literature: namely, the notion of a *continuous submeasure*. This notion appears in the measure theory literature, most commonly under the name *Maharam submeasure* (see e.g., Maharam, 1947; Talagrand, 2008; Bogachev, 2007), but is also referred to as a *subadditive Dobrakov submeasure* (see e.g., Dobrakov, 1974, 1984), and related notions arise in discussions of *Choquet capacities* (see e.g., Choquet, 1954; O'Brien and Vervaat, 1994).

Definition 10 *A* submeasure *on* $\mathcal{B}$ *is a function* $\nu : \mathcal{B} \to [0, \infty]$ *satisfying the following properties.*

1. $\nu(\emptyset) = 0$.
2. $\forall A, B \in \mathcal{B},\ A \subseteq B \Rightarrow \nu(A) \leq \nu(B)$.
3. $\forall A, B \in \mathcal{B},\ \nu(A \cup B) \leq \nu(A) + \nu(B)$.

A submeasure is called continuous *if it additionally satisfies the condition*

4. *For every monotone sequence* $\{A_k\}_{k=1}^{\infty}$ *in* $\mathcal{B}$ *with* $A_k \downarrow \emptyset$, $\lim_{k\to\infty} \nu(A_k) = 0$.

The relevance of this definition to our present discussion is via the set function $\mathbb{E}[\hat{\mu}_{\mathbb{X}}(\cdot)]$, which is always a submeasure, as stated in the following lemma.

Lemma 11 *For any process* $\mathbb{X}$, $\mathbb{E}[\hat{\mu}_{\mathbb{X}}(\cdot)]$ *is a submeasure.*

Proof Since $\hat{\mu}_{\mathbb{X}}(\emptyset) = 0$ follows directly from the definition of $\hat{\mu}_{\mathbb{X}}$, we have $\mathbb{E}[\hat{\mu}_{\mathbb{X}}(\emptyset)] = \mathbb{E}[0] = 0$ as well (property 1 of Definition 10). Furthermore, monotonicity of $\hat{\mu}_{\mathbb{X}}$ (Lemma 8) and monotonicity of the expectation imply monotonicity of $\mathbb{E}[\hat{\mu}_{\mathbb{X}}(\cdot)]$ (property 2 of Definition 10). Likewise, finite subadditivity of $\hat{\mu}_{\mathbb{X}}$ (Lemma 9) implies that for $A, B \in \mathcal{B}$, $\hat{\mu}_{\mathbb{X}}(A \cup B) \leq \hat{\mu}_{\mathbb{X}}(A) + \hat{\mu}_{\mathbb{X}}(B)$, so that monotonicity and linearity of the expectation imply $\mathbb{E}[\hat{\mu}_{\mathbb{X}}(A \cup B)] \leq \mathbb{E}[\hat{\mu}_{\mathbb{X}}(A) + \hat{\mu}_{\mathbb{X}}(B)] = \mathbb{E}[\hat{\mu}_{\mathbb{X}}(A)] + \mathbb{E}[\hat{\mu}_{\mathbb{X}}(B)]$ (property 3 of Definition 10). ■

Together with the definition of Condition 1, this immediately implies the following theorem, which states that Condition 1 is *equivalent* to $\mathbb{E}[\hat{\mu}_{\mathbb{X}}(\cdot)]$ being a continuous submeasure.

Theorem 12 *A process* $\mathbb{X}$ *satisfies Condition 1 if and only if* $\mathbb{E}[\hat{\mu}_{\mathbb{X}}(\cdot)]$ *is a continuous submeasure.*

2.3 Other Equivalent Expressions of Condition 1

We next state several other results expressing equivalent formulations of Condition 1, and other related properties. These equivalent forms will be useful in later proofs below.

Lemma 13 *The following conditions are all equivalent to Condition 1.*

- *For every monotone sequence* $\{A_k\}_{k=1}^{\infty}$ *of sets in* $\mathcal{B}$ *with* $A_k \downarrow \emptyset$,

$$\lim_{k\to\infty} \hat{\mu}_{\mathbb{X}}(A_k) = 0 \text{ (a.s.).}$$

- *For every sequence* $\{A_k\}_{k=1}^{\infty}$ *of sets in* $\mathcal{B}$,

$$\lim_{i\to\infty} \hat{\mu}_{\mathbb{X}}\left(\bigcup_{k\geq i} A_k\right) = \hat{\mu}_{\mathbb{X}}\left(\limsup_{k\to\infty} A_k\right) \text{ (a.s.).}$$

- *For every disjoint sequence* $\{A_k\}_{k=1}^{\infty}$ *of sets in* $\mathcal{B}$,

$$\lim_{i\to\infty} \hat{\mu}_{\mathbb{X}}\left(\bigcup_{k\geq i} A_k\right) = 0 \text{ (a.s.).}$$

Proof First, suppose $\mathbb{X}$ satisfies Condition 1, and let $\{A_k\}_{k=1}^{\infty}$ be any monotone sequence in $\mathcal{B}$ with $A_k \downarrow \emptyset$. By monotonicity and nonnegativity of the set function $\hat{\mu}_{\mathbb{X}}$ (Lemma 9), $\lim\limits_{k\to\infty} \hat{\mu}_{\mathbb{X}}(A_k)$ always exists and is nonnegative. Therefore, since the set function $\hat{\mu}_{\mathbb{X}}$ is bounded in $[0,1]$, the dominated convergence theorem implies $\mathbb{E}\left[\lim\limits_{k\to\infty} \hat{\mu}_{\mathbb{X}}(A_k)\right] = \lim\limits_{k\to\infty} \mathbb{E}\left[\hat{\mu}_{\mathbb{X}}(A_k)\right] = 0$, where the last equality is due to Condition 1. Combined with the fact that $\lim\limits_{k\to\infty} \hat{\mu}_{\mathbb{X}}(A_k) \geq 0$, it follows that $\lim\limits_{k\to\infty} \hat{\mu}_{\mathbb{X}}(A_k) = 0$ (a.s.) (e.g., Ash and Doléans-Dade, 2000, Theorem 1.6.6). Thus, Condition 1 implies the first condition in the lemma.

Next, let $\mathbb{X}$ be any process satisfying the first condition in the lemma, and let $\{A_k\}_{k=1}^{\infty}$ be any sequence in $\mathcal{B}$. For each $k \in \mathbb{N}$, let $B_k = A_k \setminus \bigcup\limits_{j>k} A_j$. Note that $\{B_k\}_{k=1}^{\infty}$ is a sequence of disjoint measurable sets. In particular, this implies $\bigcup\limits_{k\geq i} B_k \downarrow \emptyset$, so that (since $\mathbb{X}$ satisfies the first condition) $\lim\limits_{i\to\infty} \hat{\mu}_{\mathbb{X}}\left(\bigcup\limits_{k\geq i} B_k\right) = 0$ (a.s.). Furthermore, for any $i \in \mathbb{N}$, we have $\bigcup\limits_{k\geq i} A_k = \left(\limsup\limits_{j\to\infty} A_j\right) \cup \bigcup\limits_{k\geq i} B_k$. Therefore, by finite subadditivity of $\hat{\mu}_{\mathbb{X}}$ (Lemma 9),

$$\lim_{i\to\infty} \hat{\mu}_{\mathbb{X}}\left(\bigcup_{k\geq i} A_k\right) = \lim_{i\to\infty} \hat{\mu}_{\mathbb{X}}\left(\left(\limsup_{j\to\infty} A_j\right) \cup \bigcup_{k\geq i} B_k\right)$$
$$\leq \hat{\mu}_{\mathbb{X}}\left(\limsup_{j\to\infty} A_j\right) + \lim_{i\to\infty} \hat{\mu}_{\mathbb{X}}\left(\bigcup_{k\geq i} B_k\right) = \hat{\mu}_{\mathbb{X}}\left(\limsup_{j\to\infty} A_j\right) \text{ (a.s.)}.$$

Furthermore, since $\limsup\limits_{j\to\infty} A_j \subseteq \bigcup\limits_{k\geq i} A_k$ for every $i \in \mathbb{N}$, monotonicity of $\hat{\mu}_{\mathbb{X}}$ (Lemma 8) implies $\hat{\mu}_{\mathbb{X}}\left(\bigcup\limits_{k\geq i} A_k\right) \geq \hat{\mu}_{\mathbb{X}}\left(\limsup\limits_{j\to\infty} A_j\right)$, which implies $\lim\limits_{i\to\infty} \hat{\mu}_{\mathbb{X}}\left(\bigcup\limits_{k\geq i} A_k\right) \geq \hat{\mu}_{\mathbb{X}}\left(\limsup\limits_{j\to\infty} A_j\right)$. Together, we have that the first condition implies the second condition in this lemma. Furthermore, the second condition in this lemma trivially implies the third condition, since any *disjoint* sequence $\{A_k\}_{k=1}^{\infty}$ in $\mathcal{B}$ has $\limsup\limits_{k\to\infty} A_k = \emptyset$, and $\hat{\mu}_{\mathbb{X}}(\emptyset) = 0$ is immediate from the definition of $\hat{\mu}_{\mathbb{X}}$.

Finally, suppose the third condition in this lemma holds, and let $\{A_k\}_{k=1}^{\infty}$ be a monotone sequence in $\mathcal{B}$ with $A_k \downarrow \emptyset$. For each $k \in \mathbb{N}$, let $B_k = A_k \setminus \bigcup\limits_{j>k} A_j$. Note that $\{B_k\}_{k=1}^{\infty}$ is a sequence of disjoint sets in $\mathcal{B}$, and that monotonicity of $\{A_k\}_{k=1}^{\infty}$ implies $\forall k \in \mathbb{N}$, $A_k = \left(\limsup\limits_{j\to\infty} A_j\right) \cup \bigcup\limits_{i\geq k} B_i$; furthermore, $A_k \downarrow \emptyset$ implies $\limsup\limits_{j\to\infty} A_j = \emptyset$, so that $A_k = \bigcup\limits_{i\geq k} B_i$. Therefore, the third condition in the lemma implies

$$\lim_{k\to\infty} \hat{\mu}_{\mathbb{X}}(A_k) = \lim_{k\to\infty} \hat{\mu}_{\mathbb{X}}\left(\bigcup_{i\geq k} B_i\right) = 0 \text{ (a.s.)}.$$

Since the set function $\hat{\mu}_{\mathbb{X}}$ is bounded in $[0,1]$, combining this with the dominated convergence theorem implies $\lim_{k\to\infty} \mathbb{E}[\hat{\mu}_{\mathbb{X}}(A_k)] = \mathbb{E}\left[\lim_{k\to\infty} \hat{\mu}_{\mathbb{X}}(A_k)\right] = 0$. Since this applies to any such sequence $\{A_k\}_{k=1}^{\infty}$, we have that Condition 1 holds. This completes the proof of the lemma. ■

In combination with Lemma 13, the following lemma allows us to extend Condition 1 to other useful equivalent forms. In particular, the form expressed in (2) will be a key component in the proof below (in Lemma 19) that Condition 1 is a *necessary* condition for a process $\mathbb{X}$ to admit strong universal self-adaptive learning.

Lemma 14 *For any sequence $\mathbf{x} = \{x_t\}_{t=1}^{\infty}$ of elements of $\mathcal{X}$, and any sequence $\{A_i\}_{i=1}^{\infty}$ of disjoint subsets of $\mathcal{X}$, the following conditions are all equivalent.*

$$\lim_{k\to\infty} \hat{\mu}_{\mathbf{x}}\left(\bigcup_{i\geq k} A_i\right) = 0. \tag{1}$$

$$\lim_{n\to\infty} \hat{\mu}_{\mathbf{x}}\Big(\bigcup\{A_i : x_{1:n} \cap A_i = \emptyset\}\Big) = 0. \tag{2}$$

$$\lim_{m\to\infty} \lim_{n\to\infty} \hat{\mu}_{\mathbf{x}}\Big(\bigcup\{A_i : |x_{1:n} \cap A_i| < m\}\Big) = 0. \tag{3}$$

Proof Fix $\mathbf{x}$ and $\{A_i\}_{i=1}^{\infty}$ as described. For each $x \in \bigcup_{i=1}^{\infty} A_i$, let $i(x)$ denote the index $i \in \mathbb{N}$ with $x \in A_i$; for each $x \in \mathcal{X} \setminus \bigcup_{i=1}^{\infty} A_i$, let $i(x) = 0$.

First, suppose (2) is satisfied. For any $k \in \mathbb{N}$, let

$$n_k = \max\left\{ n \in \mathbb{N} \cup \{0, \infty\} : x_{1:n} \cap \bigcup_{i\geq k} A_i = \emptyset \right\}.$$

By definition of n_k, we have $\bigcup_{i\geq k} A_i \subseteq \bigcup\{A_i : x_{1:n_k} \cap A_i = \emptyset\}$, so that monotonicity of $\hat{\mu}_{\mathbf{x}}$ (Lemma 9) implies

$$\lim_{k\to\infty} \hat{\mu}_{\mathbf{x}}\left(\bigcup_{i\geq k} A_i\right) \leq \lim_{k\to\infty} \hat{\mu}_{\mathbf{x}}\Big(\bigcup\{A_i : x_{1:n_k} \cap A_i = \emptyset\}\Big). \tag{4}$$

Next note that monotonicity of $\bigcup_{i\geq k} A_i$ implies n_k is nondecreasing in k. In particular, this implies that if any $k \in \mathbb{N}$ has $n_k = \infty$, then

$$\lim_{k\to\infty} \hat{\mu}_{\mathbf{x}}\Big(\bigcup\{A_i : x_{1:n_k} \cap A_i = \emptyset\}\Big) = \hat{\mu}_{\mathbf{x}}\Big(\bigcup\{A_i : \mathbf{x} \cap A_i = \emptyset\}\Big) = 0$$

by definition of $\hat{\mu}_{\mathbf{x}}$, which establishes (1) when combined with (4). Otherwise, suppose $n_k < \infty$ for all $k \in \mathbb{N}$. In this case, note that $\forall k \in \mathbb{N}$, by maximality of n_k, we have $x_{n_k+1} \in \bigcup_{i\geq k} A_i$, so that $i(x_{n_k+1}) \geq k$. Together with the definition of n_k this also implies

$x_{1:n_k} \cap \bigcup_{i \geq i(x_{n_k+1})+1} A_i = \emptyset$, and by the definition of $i(x_{n_k+1})$ we know $x_{n_{k+1}} \notin \bigcup_{i \geq i(x_{n_k+1})+1} A_i$, so that in fact $x_{1:(n_k+1)} \cap \bigcup_{i \geq i(x_{n_k+1})+1} A_i = \emptyset$. This implies $n_{i(x_{n_k+1})+1} \geq n_k+1$. Thus, $\exists k' > k$ s.t. $n_{k'} \geq n_k + 1$. Together with monotonicity of n_k, this implies $n_k \uparrow \infty$. Combined with (2), this implies that

$$\lim_{k \to \infty} \hat{\mu}_{\mathbf{x}}\Big(\bigcup \{A_i : x_{1:n_k} \cap A_i = \emptyset\}\Big) = \lim_{n \to \infty} \hat{\mu}_{\mathbf{x}}\Big(\bigcup \{A_i : x_{1:n} \cap A_i = \emptyset\}\Big) = 0,$$

which establishes (1) when combined with (4) and nonnegativity of $\hat{\mu}_{\mathbf{x}}$ (Lemma 9).

Next, suppose (1) is satisfied, and fix any $m \in \mathbb{N}$. By inductively applying the finite subadditivity property of $\hat{\mu}_{\mathbf{x}}$ (Lemma 9), for any $n, k \in \mathbb{N}$,

$$\hat{\mu}_{\mathbf{x}}\Big(\bigcup \{A_i : |x_{1:n} \cap A_i| < m\}\Big) \leq \hat{\mu}_{\mathbf{x}}\Big(\bigcup \{A_i : |x_{1:n} \cap A_i| < m, i \geq k\}\Big) + \sum_{\substack{i \in \{1,\ldots,k-1\}: \\ |x_{1:n} \cap A_i| < m}} \hat{\mu}_{\mathbf{x}}(A_i). \tag{5}$$

Note that, for any $i \in \mathbb{N}$ with $\hat{\mu}_{\mathbf{x}}(A_i) > 0$, there must be an infinite subsequence of $\mathbf{x}$ contained in A_i; in particular, this implies $\exists n'_i \in \mathbb{N}$ with $|x_{1:n'_i} \cap A_i| = m$. Also define $n'_i = 0$ for every $i \in \mathbb{N}$ with $\hat{\mu}_{\mathbf{x}}(A_i) = 0$. Therefore, defining

$$k_n = \min\left(\left\{i \in \mathbb{N} : n'_i > n\right\} \cup \{\infty\}\right)$$

for every $n \in \mathbb{N}$, we have that every $i < k_n$ has either $|x_{1:n} \cap A_i| \geq m$ or $\hat{\mu}_{\mathbf{x}}(A_i) = 0$. Therefore,

$$\sum_{\substack{i \in \{1,\ldots,k_n-1\}: \\ |x_{1:n} \cap A_i| < m}} \hat{\mu}_{\mathbf{x}}(A_i) = 0. \tag{6}$$

Furthermore, by definition, k_n is nondecreasing, and if $k_n < \infty$, then any $n' \geq n'_{k_n}$ has $n' > n$ (since $n'_{k_n} > n$ by the definition of k_n), and hence $n'_i \leq n'$ for every $i \leq k_n$ (since minimality of k_n implies $n'_i \leq n < n'$ for every $i < k_n$, and by assumption $n'_{k_n} \leq n'$), which implies $k_{n'} \geq k_n + 1$. Therefore, we have $k_n \to \infty$. Thus, combined with (5) and (6), and monotonicity of $\hat{\mu}_{\mathbf{x}}$ (Lemma 9), we have

$$\begin{aligned}
&\lim_{n \to \infty} \hat{\mu}_{\mathbf{x}}\Big(\bigcup \{A_i : |x_{1:n} \cap A_i| < m\}\Big) \\
&\leq \lim_{n \to \infty} \hat{\mu}_{\mathbf{x}}\Big(\bigcup \{A_i : |x_{1:n} \cap A_i| < m, i \geq k_n\}\Big) + \sum_{\substack{i \in \{1,\ldots,k_n-1\}: \\ |x_{1:n} \cap A_i| < m}} \hat{\mu}_{\mathbf{x}}(A_i) \\
&= \lim_{n \to \infty} \hat{\mu}_{\mathbf{x}}\Big(\bigcup \{A_i : |x_{1:n} \cap A_i| < m, i \geq k_n\}\Big) \leq \lim_{k \to \infty} \hat{\mu}_{\mathbf{x}}\left(\bigcup_{i \geq k} A_i\right).
\end{aligned}$$

If (1) is satisfied, this last expression is 0. Thus,

$$\lim_{n \to \infty} \hat{\mu}_{\mathbf{x}}\Big(\bigcup \{A_i : |x_{1:n} \cap A_i| < m\}\Big) = 0$$

for all $m \in \mathbb{N}$. Taking the limit of both sides as $m \to \infty$ establishes (3).

Finally, note that for any $n \in \mathbb{N}$,

$$\hat{\mu}_{\mathbf{x}}\Big(\bigcup\{A_i : x_{1:n} \cap A_i = \emptyset\}\Big) = \hat{\mu}_{\mathbf{x}}\Big(\bigcup\{A_i : |x_{1:n} \cap A_i| < 1\}\Big),$$

and monotonicity of $\hat{\mu}_{\mathbf{x}}$ (Lemma 9) implies that for any $m \in \mathbb{N}$,

$$\hat{\mu}_{\mathbf{x}}\Big(\bigcup\{A_i : |x_{1:n} \cap A_i| < 1\}\Big) \leq \hat{\mu}_{\mathbf{x}}\Big(\bigcup\{A_i : |x_{1:n} \cap A_i| < m\}\Big).$$

Taking limits of both sides, we have

$$\lim_{n\to\infty} \hat{\mu}_{\mathbf{x}}\Big(\bigcup\{A_i : |x_{1:n} \cap A_i| < 1\}\Big) \leq \lim_{m\to\infty} \lim_{n\to\infty} \hat{\mu}_{\mathbf{x}}\Big(\bigcup\{A_i : |x_{1:n} \cap A_i| < m\}\Big).$$

Thus, if (3) is satisfied, then (2) must also hold. ■

One interesting property of processes $\mathbb{X}$ satisfying Condition 1 is that $\hat{\mu}_{\mathbb{X}}$ is *countably subadditive* (almost surely), as implied by the following two lemmas. Note that this is not necessarily true of processes $\mathbb{X}$ failing to satisfy Condition 1 (e.g., the process $X_i = i$ on $\mathbb{N}$ does not have countably subadditive $\hat{\mu}_{\mathbb{X}}$). However, we note that this kind of countable subadditivity is not actually *equivalent* to Condition 1, as not every process satisfying this countable subadditivity condition also satisfies Condition 1 (e.g., any process $\mathbb{X}$ on $\mathbb{N}$ with $\forall i \in \mathbb{N}, \hat{\mu}_{\mathbb{X}}(\{i\}) = 1$).

Lemma 15 *For any sequence $\mathbf{x} = \{x_t\}_{t=1}^{\infty}$ of elements of $\mathcal{X}$, and any sequence $\{A_i\}_{i=1}^{\infty}$ of disjoint subsets of $\mathcal{X}$, if (1) is satisfied, then*

$$\hat{\mu}_{\mathbf{x}}\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \hat{\mu}_{\mathbf{x}}(A_i).$$

Proof By finite subadditivity of $\hat{\mu}_{\mathbf{x}}$ (Lemma 9 and induction), we have that for any $k \in \mathbb{N}$,

$$\hat{\mu}_{\mathbf{x}}\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \hat{\mu}_{\mathbf{x}}\left(\bigcup_{i\geq k} A_i\right) + \sum_{i=1}^{k-1} \hat{\mu}_{\mathbf{x}}(A_i). \tag{7}$$

If (1) is satisfied, then $\lim_{k\to\infty} \hat{\mu}_{\mathbf{x}}\left(\bigcup_{i\geq k} A_i\right) = 0$, so that taking the limit as $k \to \infty$ in (7) yields the claimed inequality, completing the proof. ■

Lemma 16 *If $\mathbb{X}$ satisfies Condition 1, then for any sequence $\{A_i\}_{i=1}^{\infty}$ in $\mathcal{B}$,*

$$\hat{\mu}_{\mathbb{X}}\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \hat{\mu}_{\mathbb{X}}(A_i) \qquad \textit{(a.s.)}.$$

Proof Let $B_1 = A_1$, and for each $i \in \mathbb{N} \setminus \{1\}$, let $B_i = A_i \setminus \bigcup_{j=1}^{i-1} A_j$. Then $\{B_i\}_{i=1}^{\infty}$ is a disjoint sequence in $\mathcal{B}$. If $\mathbb{X}$ satisfies Condition 1, then Lemma 13 implies $\lim_{k\to\infty} \hat{\mu}_{\mathbb{X}}\left(\bigcup_{j\geq k} B_j\right) = 0$ (a.s.). Combined with Lemma 15, this implies that $\hat{\mu}_{\mathbb{X}}\left(\bigcup_{i=1}^{\infty} B_i\right) \leq \sum_{i=1}^{\infty} \hat{\mu}_{\mathbb{X}}(B_i)$ (a.s.). Noting that $\bigcup_{i=1}^{\infty} B_i = \bigcup_{i=1}^{\infty} A_i$, we have $\hat{\mu}_{\mathbb{X}}\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \hat{\mu}_{\mathbb{X}}(B_i)$ (a.s.). Finally, since $B_i \subseteq A_i$ for every $i \in \mathbb{N}$, monotonicity of $\hat{\mu}_{\mathbb{X}}$ (Lemma 9) implies $\hat{\mu}_{\mathbb{X}}(B_i) \leq \hat{\mu}_{\mathbb{X}}(A_i)$, so that $\hat{\mu}_{\mathbb{X}}\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \hat{\mu}_{\mathbb{X}}(A_i)$ (a.s.). ∎

3. Relation to the Condition of Convergent Relative Frequencies

Before proceeding with the general analysis, we first discuss the relation between Condition 1 and the commonly-studied condition of *convergent relative frequencies*. In particular, we show that Condition 1 is a *strictly more-general* condition. This is interesting in the context of learning, as the vast majority of the prior literature on statistical learning theory without the i.i.d. assumption studies learning rules designed for and analyzed under assumptions that imply convergent relative frequencies. These results therefore indicate that we should not expect such learning rules to be optimistically universal, and hence that we will need to seek more general strategies in designing optimistically universal learning rules.

Formally, define CRF as the set of processes $\mathbb{X}$ such that, $\forall A \in \mathcal{B}$,

$$\lim_{m\to\infty} \frac{1}{m} \sum_{t=1}^{m} \mathbb{1}_A(X_t) \text{ exists (a.s.).} \tag{8}$$

These processes are said to have *convergent relative frequencies*. Equivalently, this is the family of processes with *ergodic properties* with respect to the class of measurements $\{\mathbb{1}_{A\times\mathcal{X}^\infty} : A \in \mathcal{B}\}$ (Gray, 2009). Certainly CRF contains every i.i.d. process, by the *strong law of large numbers*. More generally, it is known that any *stationary* process $\mathbb{X}$ is contained in CRF (by Birkhoff's ergodic theorem), and in fact, it suffices for the process to be *asymptotically mean stationary* (see Gray, 2009, Theorem 8.1).

3.1 Processes with Convergent Relative Frequencies Satisfy Condition 1

The following theorem establishes that every $\mathbb{X}$ with convergent relative frequencies satisfies Condition 1, and that the inclusion is *strict* in all nontrivial cases.

Theorem 17 CRF $\subseteq \mathcal{C}_1$, *and the inclusion is* strict *iff* $|\mathcal{X}| \geq 2$.

Proof Fix any $\mathbb{X} \in$ CRF. For each $A \in \mathcal{B}$, define $\pi_m(A) = \frac{1}{m}\sum_{t=1}^{m} \mathbb{P}(X_t \in A)$. One can easily verify that π_m is a probability measure. The definition of CRF implies that, $\forall A \in \mathcal{B}$, there exists an event E_A of probability one, on which $\lim_{m\to\infty} \frac{1}{m}\sum_{t=1}^{m} \mathbb{1}_A(X_t)$ exists;

in particular, this implies $\hat{\mu}_{\mathbb{X}}(A) = \lim_{m\to\infty} \frac{1}{m}\sum_{t=1}^{m} \mathbb{1}_A(X_t)\mathbb{1}_{E_A}$ almost surely. Together with the dominated convergence theorem and linearity of expectations, this implies

$$\begin{aligned}\mathbb{E}[\hat{\mu}_{\mathbb{X}}(A)] &= \mathbb{E}\left[\lim_{m\to\infty}\frac{1}{m}\sum_{t=1}^{m}\mathbb{1}_A(X_t)\mathbb{1}_{E_A}\right] = \lim_{m\to\infty}\mathbb{E}\left[\frac{1}{m}\sum_{t=1}^{m}\mathbb{1}_A(X_t)\mathbb{1}_{E_A}\right]\\ &= \lim_{m\to\infty}\frac{1}{m}\sum_{t=1}^{m}\mathbb{P}(X_t\in A) = \lim_{m\to\infty}\pi_m(A).\end{aligned}$$

In particular, this establishes that the limit in the rightmost expression exists. The Vitali-Hahn-Saks theorem then implies that $\lim_{m\to\infty}\pi_m(\cdot)$ is also a probability measure (see Gray, 2009, Lemma 7.4). Thus, we have established that $A \mapsto \mathbb{E}[\hat{\mu}_{\mathbb{X}}(A)]$ is a probability measure, and hence is a continuous submeasure (see e.g., Schervish, 1995, Theorem A.19). That $\mathrm{CRF} \subseteq \mathcal{C}_1$ now follows from Theorem 12.

For the claim about strict inclusion, first note that if $|\mathcal{X}| = 1$ then there is effectively only one possible process (infinitely repeating the sole element of $\mathcal{X}$), and it is trivially in CRF, so that $\mathrm{CRF} = \mathcal{C}_1$. On the other hand, suppose $|\mathcal{X}| \geq 2$, let x_0, x_1 be distinct elements of $\mathcal{X}$, and define a deterministic process $\mathbb{X}$ such that, for every $i \in \mathbb{N}$ and every $t \in \{3^{i-1}, \ldots, 3^i - 1\}$, $X_t = x_{i-2\lfloor i/2\rfloor}$: that is, $X_t = x_0$ if i is even and $X_t = x_1$ if i is odd. Since any monotone sequence $\{A_k\}_{k=1}^{\infty}$ in $\mathcal{B}$ with $A_k \downarrow \emptyset$ necessarily has some $k_0 \in \mathbb{N}$ with $\{x_0, x_1\} \cap A_k = \emptyset$ for all $k \geq k_0$, we have $\mathbb{E}[\hat{\mu}_{\mathbb{X}}(A_k)] = 0$ for all $k \geq k_0$, so that $\mathbb{X} \in \mathcal{C}_1$. However, for any odd i, $\frac{1}{3^i-1}\sum_{t=1}^{3^i-1}\mathbb{1}_{\{x_1\}}(X_t) \geq \frac{2}{3}$, so that $\limsup_{m\to\infty}\frac{1}{m}\sum_{t=1}^{m}\mathbb{1}_{\{x_1\}}(X_t) \geq \frac{2}{3}$, while for any even i, $\frac{1}{3^i-1}\sum_{t=1}^{3^i-1}\mathbb{1}_{\{x_1\}}(X_t) \leq \frac{1}{3}$, so that $\liminf_{m\to\infty}\frac{1}{m}\sum_{t=1}^{m}\mathbb{1}_{\{x_1\}}(X_t) \leq \frac{1}{3}$. Therefore $\frac{1}{m}\sum_{t=1}^{m}\mathbb{1}_{\{x_1\}}(X_t)$ does not have a limit as $m \to \infty$, and hence $\mathbb{X} \notin \mathrm{CRF}$. ■

3.2 Inconsistency of the Nearest Neighbor Rule

The separation $\mathcal{C}_1 \setminus \mathrm{CRF} \neq \emptyset$ established above indicates that, in approaching the design of consistent inductive or self-adaptive learning rules under processes in $\mathcal{C}_1$, we should not rely on the property of having convergent relative frequencies, as it is not generally guaranteed to hold. Since most learning rules in the prior literature rely heavily on this property for their performance guarantees, we should not generally expect them to be consistent under processes in $\mathcal{C}_1$. To give a concrete example illustrating this, consider $\mathcal{X} = [0, 1]$ (with the standard topology), and let f_n be the well-known *nearest neighbor* learning rule: an inductive learning rule defined by the property that $f_n(x_{1:n}, y_{1:n}, x) = y_{i_n}$, where $i_n = \underset{i\in\{1,\ldots,n\}}{\operatorname{argmin}}\, |x - x_i|$ (breaking ties arbitrarily). This learning rule is known to be strongly universally consistent (in the sense of Definition 1) under every i.i.d. process (e.g., Devroye, Györfi, and Lugosi, 1996).

We exhibit a process $\mathbb{X} \in \mathcal{C}_1$, under which the nearest neighbor inductive learning rule is *not* universally consistent.[3] This also provides a second proof that $\mathcal{C}_1 \setminus \mathrm{CRF} \neq \emptyset$ for this space, as this process will not have convergent relative frequencies. Specifically, let $\{W_i\}_{i=1}^{\infty}$ be independent Uniform$(0, 1/2)$ random variables. Let $n_1 = 1$, and for each $k \in \mathbb{N}$ with $k \geq 2$, inductively define $n_k = n_{k-1} + k \cdot n_{k-1}^2$. Now for each $k \in \mathbb{N}$, let $a_k = k - 2\lfloor k/2 \rfloor$ (i.e., $a_k = 1$ if k is odd, and otherwise $a_k = 0$), and let $b_k = 1 - a_k$. Define $X_1 = 0$, and for each $k \in \mathbb{N}$ with $k \geq 2$, and each $i \in \{1, \ldots, n_{k-1}^2\}$, define $X_{n_{k-1}+(i-1)k+1} = \frac{b_k}{2} + \frac{i-1}{2n_{k-1}^2}$, and for each $j \in \{2, \ldots, k\}$, define $X_{n_{k-1}+(i-1)k+j} = \frac{a_k}{2} + W_{n_{k-1}+(i-1)k+j}$.

The intention in constructing this process is that there are segments of the sequence in which $[0, 1/2)$ is relatively sparse compared to $[1/2, 1]$, and other segments of the sequence in which $[1/2, 1]$ is relatively sparse compared to $[0, 1/2)$. Furthermore, at certain time points (namely, the n_k times), the vast majority of the points on the sparse side are determined *a priori*, in contrast to the points on the dense side, which are uniform random. This is designed to frustrate most learning rules designed under the CRF assumption, many of which would base their predictions on the sparse side on these deterministic points, rather than the relatively very-sparse random points in the same region left over from the previous epoch (i.e., when that region was relatively dense, and the majority of points in that region were uniform random). It is easy to verify that, because of this switching of which side is dense and which side sparse, which occurs infinitely many times, this process $\mathbb{X}$ does *not* have convergent relative frequencies.

We first argue that $\mathbb{X}$ satisfies Condition 1. Let $I = \{1\} \cup \{n_{k-1} + (i-1)k + 1 : k \in \mathbb{N} \setminus \{1\}, i \in \{1, \ldots, n_{k-1}^2\}\}$. Note that every $t \in \mathbb{N} \setminus I$ has $X_t \in \left\{W_t, \frac{1}{2} + W_t\right\}$. Now note that, for any $k \in \mathbb{N} \setminus \{1, 2\}$ and any $m \in \{n_{k-1} + 1, \ldots, n_k\}$,

$$|\{t \in I : t \leq m\}| \leq n_{k-2} + \frac{n_{k-1} - n_{k-2}}{k-1} + \left\lceil \frac{m - n_{k-1}}{k} \right\rceil \leq 1 + n_{k-2} + \frac{n_{k-1}}{k-1} + \frac{m - n_{k-1}}{k-1}$$
$$= 1 + n_{k-2} + \frac{m}{k-1} \leq 1 + \sqrt{\frac{n_{k-1}}{k-1}} + \frac{m}{k-1} \leq 1 + \sqrt{\frac{m}{k-1}} + \frac{m}{k-1}.$$

Thus, letting $k_m = \min\{k \in \mathbb{N} : m \leq n_k\}$ for each $m \in \mathbb{N}$, and noting that $k_m \to \infty$ (since each n_k is finite), we have that

$$\lim_{m\to\infty} \frac{|\{t \in I : t \leq m\}|}{m} \leq \lim_{m\to\infty} \frac{1}{m} + \sqrt{\frac{1}{m(k_m - 1)}} + \frac{1}{k_m - 1} = 0.$$

We therefore have that, for any set $A \in \mathcal{B}$,

$$\hat{\mu}_{\mathbb{X}}(A) \leq \limsup_{m\to\infty} \frac{1}{m} \sum_{t=1}^{m} \mathbb{1}_{\mathbb{N}\setminus I}(t) \mathbb{1}_A(X_t) + \lim_{m\to\infty} \frac{|\{t \in I : t \leq m\}|}{m} = \limsup_{m\to\infty} \frac{1}{m} \sum_{t=1}^{m} \mathbb{1}_{\mathbb{N}\setminus I}(t) \mathbb{1}_A(X_t).$$

Furthermore, the rightmost expression above is at most

$$\limsup_{m\to\infty} \frac{1}{m} \sum_{t=1}^{m} \left(\mathbb{1}_A(W_t) + \mathbb{1}_A\left(\frac{1}{2} + W_t\right) \right) \leq \limsup_{m\to\infty} \frac{1}{m} \sum_{t=1}^{m} \mathbb{1}_A(W_t) + \limsup_{m\to\infty} \frac{1}{m} \sum_{t=1}^{m} \mathbb{1}_A\left(\frac{1}{2} + W_t\right)$$

3. Of course, Theorem 6 indicates that *any* inductive learning rule has processes in $\mathcal{C}_1$ for which it is not universally consistent. However, the construction here is more direct, and illustrates a common failing of many learning rules designed for i.i.d. data, so it is worth presenting this specialized argument as well.

and the strong law of large numbers and the union bound imply that, with probability one, the expression on the right hand side equals $2\lambda(A \cap (0,1/2)) + 2\lambda(A \cap (1/2,1)) = 2\lambda(A)$, where λ is the Lebesgue measure. In particular, this implies $\mathbb{E}[\hat{\mu}_{\mathbb{X}}(A)] \leq 2\lambda(A)$ for every $A \in \mathcal{B}$. Therefore, for any monotone sequence $\{A_k\}_{k=1}^{\infty}$ in $\mathcal{B}$ with $A_k \downarrow \emptyset$, $\lim\limits_{k\to\infty} \mathbb{E}[\hat{\mu}_{\mathbb{X}}(A_k)] \leq \lim\limits_{k\to\infty} 2\lambda(A_k) = 0$ since $2\lambda(\cdot)$ is a finite measure (because $\mathcal{X}$ is bounded) and therefore is continuous (see e.g., Schervish, 1995, Theorem A.19). Thus, $\mathbb{X}$ satisfies Condition 1.

Now to see that the nearest neighbor rule is not universally consistent under this process $\mathbb{X}$, let $y_0, y_1 \in \mathcal{Y}$ be such that $\ell(y_0, y_1) > 0$. Define

$$V = \left\{ \frac{b_k}{2} + \frac{i-1}{2n_{k-1}^2} : k \in \mathbb{N} \setminus \{1\}, i \in \{1, \ldots, n_{k-1}^2\} \right\},$$

and take $f^\star(x) = y_1$ for $x \in [0,1] \setminus V$, and $f^\star(x) = y_0$ for $x \in V$, and note that this is a measurable function since V is measurable. Note that we have defined $f^\star$ so that, with probability one, every $t \in I$ has $f^\star(X_t) = y_0$, and every $t \in \mathbb{N} \setminus I$ has $f^\star(X_t) = y_1$. Then note that, for any $k \in \mathbb{N} \setminus \{1,2\}$ with $a_k = 1$, the points $\{X_i : 1 \leq i \leq n_k, f^\star(X_i) = y_0\}$ form a $\frac{1}{2n_{k-1}^2}$ cover of $[0,1/2)$. Furthermore, the set $\{X_i : 1 \leq i \leq n_k, f^\star(X_i) = y_1\} \cap (0,1/2)$ contains at most n_{k-1} points. Together, these facts imply that the set $N_k = \{x \in [0,1] : f_{n_k}(X_{1:n_k}, f^\star(X_{1:n_k}), x) = y_0\}$ has $\lambda(N_k \cap (0,1/2)) \geq \frac{1}{2} - \frac{n_{k-1}}{2n_{k-1}^2} = \frac{1}{2}\left(1 - \frac{1}{n_{k-1}}\right)$. In particular, this implies that a Uniform$(0,1/2)$ random sample (independent from f_{n_k} and $X_{1:n_k}$) has probability at least $1 - \frac{1}{n_{k-1}}$ of being in N_k. However, for every $k' \in \mathbb{N} \setminus \{1\}$ with $2k' > k$, we have $a_{2k'} = 0$, so that the set $\{X_i : n_{2k'-1} < i \leq n_{2k'}\} \cap (0,1/2)$ consists of $(2k'-1)n_{2k'-1}^2 = \frac{2k'-1}{2k'}(n_{2k'} - n_{2k'-1})$ independent Uniform$(0,1/2)$ samples (also independent from f_{n_k} and $X_{1:n_k}$). Since V is countable, with probability one every one of these samples has $f^\star(X_i) = y_1$. Furthermore, a Chernoff bound (under the conditional distribution given f_{n_k} and $X_{1:n_k}$) and the law of total probability imply that, with probability at least $1 - \exp\left\{-\frac{1}{2(2k'-1)^2}\left(1 - \frac{1}{n_{k-1}}\right)(2k'-1)n_{2k'-1}^2\right\} \geq 1 - \exp\{-(2k'-1)/4\}$,

$$|N_k \cap \{X_i : n_{2k'-1} < i \leq n_{2k'}\} \cap (0,1/2)| \geq \left(1 - \frac{1}{2k'-1}\right)\left(1 - \frac{1}{n_{k-1}}\right)\frac{2k'-1}{2k'}(n_{2k'} - n_{2k'-1}).$$

Since $\sum\limits_{k'=1}^{\infty} \exp\{-(2k'-1)/4\} < \infty$, the Borel-Cantelli lemma implies that with probability one this occurs for all sufficiently large k'. Thus, by the union bound, we have that with probability one,

$$\begin{aligned}
&\hat{\mu}_{\mathbb{X}}(\ell(f_{n_k}(X_{1:n_k}, f^\star(X_{1:n_k}), \cdot), f^\star(\cdot))) \\
&\geq \limsup_{k'\to\infty} \frac{1}{n_{2k'}} \sum_{t=1}^{n_{2k'}} \ell(f_{n_k}(X_{1:n_k}, f^\star(X_{1:n_k}), X_t), f^\star(X_t)) \\
&\geq \limsup_{k'\to\infty} \frac{|N_k \cap \{X_i : n_{2k'-1} < i \leq n_{2k'}\} \cap (0,1/2)|}{n_{2k'}} \ell(y_0, y_1) \\
&\geq \ell(y_0, y_1) \limsup_{k'\to\infty} \left(1 - \frac{1}{2k'-1}\right)\left(1 - \frac{1}{n_{k-1}}\right)\frac{2k'-1}{2k'}\left(1 - \frac{n_{2k'-1}}{n_{2k'}}\right) = \ell(y_0, y_1)\left(1 - \frac{1}{n_{k-1}}\right).
\end{aligned}$$

By the union bound, with probability one, this holds for every odd value of $k \in \mathbb{N} \setminus \{1, 2\}$. Thus, with probability one,

$$\limsup_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(f_n, f^\star; n) \geq \limsup_{k\to\infty} \hat{\mu}_{\mathbb{X}}\left(\ell\left(f_{n_{2k+1}}(X_{1:n_{2k+1}}, f^\star(X_{1:n_{2k+1}}), \cdot), f^\star(\cdot)\right)\right)$$
$$\geq \limsup_{k\to\infty} \ell(y_0, y_1)\left(1 - \frac{1}{n_{2k}}\right) = \ell(y_0, y_1).$$

In particular, this implies f_n is not strongly universally consistent under $\mathbb{X}$. Similar arguments can be constructed for most learning methods in common use (e.g., kernel rules, the k-nearest neighbors rule, support vector machines with radial basis kernel).

It is clear from this example that obtaining consistency under general $\mathbb{X}$ satisfying Condition 1 will require a new approach to the design of learning rules. We develop such an approach in the sections below. The essential innovation is to base the predictions not only on performance on points that seem typical relative to the present data set $X_{1:n}$, but also on the *prefixes* $X_{1:n'}$ of the data set (for a well-chosen range of values $n' \leq n$).

4. Condition 1 is Necessary and Sufficient for Universal Inductive and Self-Adaptive Learning

This section presents the proof of Theorem 7 from Section 1.2, establishing equivalence of the set of processes admitting strong universal inductive learning, the set of processes admitting strong universal self-adaptive learning, and the set of processes satisfying Condition 1. For convenience, we restate that result here (in simplified form) as follows.

Theorem 7 (restated) SUIL = SUAL = $\mathcal{C}_1$.

The proof is by way of three lemmas: Lemma 19, representing necessity of Condition 1 for strong universal self-adaptive learning, Lemma 25, representing sufficiency of Condition 1 for strong universal inductive learning, and Lemma 18, which indicates that any process admitting strong universal inductive learning necessarily admits strong universal self-adaptive learning. We begin with the last (and simplest) of these.

Lemma 18 SUIL $\subseteq$ SUAL.

Proof Let $\mathbb{X} \in$ SUIL, and let f_n be an inductive learning rule such that, for every measurable $f^\star : \mathcal{X} \to \mathcal{Y}$, $\lim\limits_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(f_n, f^\star; n) = 0$ (a.s.). Then define a self-adaptive learning rule $g_{n,m}$ as follows. For every $n, m \in \mathbb{N}$, and every $\{x_i\}_{i=1}^m \in \mathcal{X}^m$, $\{y_i\}_{i=1}^n \in \mathcal{Y}^n$, and $z \in \mathcal{X}$, if $n \leq m$, define $g_{n,m}(x_{1:m}, y_{1:n}, z) = f_n(x_{1:n}, y_{1:n}, z)$. With this definition, we have that for every measurable $f^\star : \mathcal{X} \to \mathcal{Y}$, for every $n \in \mathbb{N}$,

$$\hat{\mathcal{L}}_{\mathbb{X}}(g_{n,\cdot}, f^\star; n) = \limsup_{t\to\infty} \frac{1}{t+1} \sum_{m=n}^{n+t} \ell(g_{n,m}(X_{1:m}, f^\star(X_{1:n}), X_{m+1}), f^\star(X_{m+1}))$$
$$= \limsup_{t\to\infty} \frac{1}{t+1} \sum_{m=n}^{n+t} \ell(f_n(X_{1:n}, f^\star(X_{1:n}), X_{m+1}), f^\star(X_{m+1})) = \hat{\mathcal{L}}_{\mathbb{X}}(f_n, f^\star; n),$$

so that $\lim_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(g_{n,\cdot}, f^\star; n) = \lim_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(f_n, f^\star; n) = 0$ (a.s.). ■

Next, we prove necessity of Condition 1 for strong universal self-adaptive learning.

Lemma 19 SUAL $\subseteq \mathcal{C}_1$.

Proof We prove this result in the contrapositive. Suppose $\mathbb{X} \notin \mathcal{C}_1$. By Lemma 13, there exists a disjoint sequence $\{A_k\}_{k=1}^\infty$ in $\mathcal{B}$ such that $\lim_{i\to\infty} \hat{\mu}_{\mathbb{X}}\left(\bigcup_{k\geq i} A_k\right) > 0$ with probability greater than 0. Furthermore, since this property involves only the limit as $i \to \infty$, we may take this sequence $\{A_k\}_{k=1}^\infty$ to have $A_1 = \mathcal{X} \setminus \bigcup_{i=2}^\infty A_i$, so that $\bigcup_{i=1}^\infty A_i = \mathcal{X}$. Lemma 14 then implies that, for this sequence $\{A_k\}_{k=1}^\infty$, with probability greater than 0, $\lim_{n\to\infty} \hat{\mu}_{\mathbb{X}}(\bigcup\{A_i : X_{1:n} \cap A_i = \emptyset\}) > 0$.

For any $n \in \mathbb{N}$, let $\bar{\mathcal{A}}(X_{1:n}) = \bigcup\{A_i : X_{1:n} \cap A_i = \emptyset\}$. Now take any two distinct values $y_0, y_1 \in \mathcal{Y}$, and construct a set of target functions $\{f_\kappa^\star : \kappa \in [0,1)\}$ as follows. For any $\kappa \in [0,1)$ and $i \in \mathbb{N}$, let $\kappa_i = \lfloor 2^i \kappa \rfloor - 2\lfloor 2^{i-1}\kappa \rfloor$: the i^{th} bit of the binary representation of κ. For each $i \in \mathbb{N}$ and each $x \in A_i$, define $f_\kappa^\star(x) = y_{\kappa_i}$. Note that for any $\kappa \in [0,1)$, $f_\kappa^\star$ is a measurable function (as it is constant within each A_i, and the A_i sets are measurable).

For any $t \in \mathbb{N}$, let i_t denote the value of $i \in \mathbb{N}$ for which $X_t \in A_i$. Now fix any self-adaptive learning rule $g_{n,m}$, and for brevity define a function $f_{n,m}^\kappa : \mathcal{X} \to \mathcal{Y}$ as $f_{n,m}^\kappa(\cdot) = g_{n,m}(X_{1:m}, f_\kappa^\star(X_{1:n}), \cdot)$. Then

$$\sup_{\kappa\in[0,1)} \mathbb{E}\left[\limsup_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(g_{n,\cdot}, f_\kappa^\star; n)\right] \geq \int_0^1 \mathbb{E}\left[\limsup_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(g_{n,\cdot}, f_\kappa^\star; n)\right] \mathrm{d}\kappa$$
$$\geq \int_0^1 \mathbb{E}\left[\limsup_{n\to\infty} \limsup_{t\to\infty} \frac{1}{t+1} \sum_{m=n}^{n+t} \ell\big(f_{n,m}^\kappa(X_{m+1}), f_\kappa^\star(X_{m+1})\big) \mathbb{1}_{\bar{\mathcal{A}}(X_{1:n})}(X_{m+1})\right] \mathrm{d}\kappa.$$

By Fubini's theorem, this is equal

$$\mathbb{E}\left[\int_0^1 \limsup_{n\to\infty} \limsup_{t\to\infty} \frac{1}{t+1} \sum_{m=n}^{n+t} \ell\big(f_{n,m}^\kappa(X_{m+1}), f_\kappa^\star(X_{m+1})\big) \mathbb{1}_{\bar{\mathcal{A}}(X_{1:n})}(X_{m+1}) \mathrm{d}\kappa\right].$$

Since ℓ is bounded, Fatou's lemma implies this is at least as large as

$$\mathbb{E}\left[\limsup_{n\to\infty} \limsup_{t\to\infty} \int_0^1 \frac{1}{t+1} \sum_{m=n}^{n+t} \ell\big(f_{n,m}^\kappa(X_{m+1}), f_\kappa^\star(X_{m+1})\big) \mathbb{1}_{\bar{\mathcal{A}}(X_{1:n})}(X_{m+1}) \mathrm{d}\kappa\right],$$

and linearity of integration implies this equals

$$\mathbb{E}\left[\limsup_{n\to\infty} \limsup_{t\to\infty} \frac{1}{t+1} \sum_{m=n}^{n+t} \mathbb{1}_{\bar{\mathcal{A}}(X_{1:n})}(X_{m+1}) \int_0^1 \ell\big(f_{n,m}^\kappa(X_{m+1}), f_\kappa^\star(X_{m+1})\big) \mathrm{d}\kappa\right]. \quad (9)$$

Note that, for any m, the value of $f_{n,m}^\kappa(X_{m+1})$ is a function of $\mathbb{X}$ and the values $\kappa_{i_1}, \ldots, \kappa_{i_n}$. Therefore, for any m with $X_{m+1} \in \bar{\mathcal{A}}(X_{1:n})$, the value of $f_{n,m}^\kappa(X_{m+1})$ is functionally independent of $\kappa_{i_{m+1}}$. Thus, letting $K \sim \text{Uniform}([0,1))$ be independent of $\mathbb{X}$ and $g_{n,m}$, for any such m we have

$$\int_0^1 \ell\big(f_{n,m}^{\kappa}(X_{m+1}), f_{\kappa}^{\star}(X_{m+1})\big)\,\mathrm{d}\kappa = \mathbb{E}\Big[\ell\big(f_{n,m}^{K}(X_{m+1}), f_{K}^{\star}(X_{m+1})\big)\,\Big|\mathbb{X}, g_{n,m}\Big]$$
$$= \mathbb{E}\Big[\mathbb{E}\Big[\ell\Big(g_{n,m}(X_{1:m}, \{y_{K_{i_j}}\}_{j=1}^{n}, X_{m+1}), y_{K_{m+1}}\Big)\,\Big|\mathbb{X}, g_{n,m}, K_{i_1}, \ldots, K_{i_n}\Big]\,\Big|\mathbb{X}, g_{n,m}\Big]$$
$$= \mathbb{E}\left[\sum_{b\in\{0,1\}} \frac{1}{2}\ell\Big(g_{n,m}(X_{1:m}, \{y_{K_{i_j}}\}_{j=1}^{n}, X_{m+1}), y_b\Big)\,\Big|\mathbb{X}, g_{n,m}\right].$$

By the triangle inequality, this is no smaller than $\mathbb{E}\left[\frac{1}{2}\ell(y_0,y_1)\middle|\mathbb{X}, g_{n,m}\right] = \frac{1}{2}\ell(y_0,y_1)$, so that (9) is at least as large as

$$\mathbb{E}\left[\limsup_{n\to\infty}\limsup_{t\to\infty}\frac{1}{t+1}\sum_{m=n}^{n+t}\mathbb{1}_{\bar{\mathcal{A}}(X_{1:n})}(X_{m+1})\frac{1}{2}\ell(y_0,y_1)\right]$$
$$= \frac{1}{2}\ell(y_0,y_1)\mathbb{E}\left[\limsup_{n\to\infty}\hat{\mu}_{\mathbb{X}}\Big(\bigcup\{A_i : X_{1:n}\cap A_i=\emptyset\}\Big)\right].$$

Since any nonnegative random variable with mean 0 necessarily equals 0 almost surely (e.g., Ash and Doléans-Dade, 2000, Theorem 1.6.6), and since $\lim_{n\to\infty}\hat{\mu}_{\mathbb{X}}\left(\bigcup\{A_i : X_{1:n}\cap A_i=\emptyset\}\right) > 0$ with probability strictly greater than 0, and the left hand side is nonnegative, we have that $\mathbb{E}\left[\limsup_{n\to\infty}\hat{\mu}_{\mathbb{X}}(\bigcup\{A_i : X_{1:n}\cap A_i=\emptyset\})\right] > 0$. Furthermore, since ℓ is a metric, we also have $\ell(y_0,y_1) > 0$. Altogether we have that

$$\sup_{\kappa\in[0,1)}\mathbb{E}\left[\limsup_{n\to\infty}\hat{\mathcal{L}}_{\mathbb{X}}(g_{n,\cdot}, f_{\kappa}^{\star}; n)\right] \geq \frac{1}{2}\ell(y_0,y_1)\mathbb{E}\left[\limsup_{n\to\infty}\hat{\mu}_{\mathbb{X}}\Big(\bigcup\{A_i : X_{1:n}\cap A_i=\emptyset\}\Big)\right] > 0.$$

In particular, this implies $\exists\kappa\in[0,1)$ such that $\mathbb{E}\left[\limsup_{n\to\infty}\hat{\mathcal{L}}_{\mathbb{X}}(g_{n,\cdot}, f_{\kappa}^{\star}; n)\right] > 0$. Since any random variable equal 0 (a.s.) necessarily has expected value 0, and since $\limsup_{n\to\infty}\hat{\mathcal{L}}_{\mathbb{X}}(g_{n,\cdot}, f_{\kappa}^{\star}; n)$ is nonnegative, we must have that, with probability greater than 0, $\limsup_{n\to\infty}\hat{\mathcal{L}}_{\mathbb{X}}(g_{n,\cdot}, f_{\kappa}^{\star}; n) > 0$, so that $g_{n,m}$ is not strongly universally consistent. Since $g_{n,m}$ was an arbitrary self-adaptive learning rule, we conclude that there does not exist a self-adaptive learning rule that is strongly universally consistent under $\mathbb{X}$: that is, $\mathbb{X}\notin$ SUAL. Since this argument holds for any $\mathbb{X}\notin\mathcal{C}_1$, the lemma follows. ■

Finally, to complete the proof of Theorem 7, we prove that Condition 1 is sufficient for $\mathbb{X}$ to admit strong universal inductive learning. We prove this via a more general strategy: namely, a kind of constrained maximum empirical risk minimization. Though the lemmas below are in fact somewhat stronger than needed to prove Theorem 7, some of them are useful later for establishing Theorem 5, and some should also be of independent interest. We propose to study an inductive learning rule $\hat{f}_n$ such that, for any $n\in\mathbb{N}$, $x_{1:n}\in\mathcal{X}^n$, and $y_{1:n}\in\mathcal{Y}^n$, the function $\hat{f}_n(x_{1:n}, y_{1:n}, \cdot)$ is defined as

$$\operatorname*{argmin}_{f\in\mathcal{F}_n}^{\varepsilon_n}\max_{\hat{m}_n\leq m\leq n}\frac{1}{m}\sum_{t=1}^{m}\ell(f(x_t), y_t), \tag{10}$$

where $\mathcal{F}_n$ is a well-chosen class of functions $\mathcal{X} \to \mathcal{Y}$, $\hat{m}_n$ is a well-chosen integer, and ε_n is an arbitrary sequence in $[0,\infty)$ with $\varepsilon_n \to 0$. For our purposes, $\hat{f}_0(\{\},\{\},\cdot)$ can be defined as an arbitrary measurable function $\mathcal{X} \to \mathcal{Y}$. The class $\mathcal{F}_n$ and integer $\hat{m}_n$, and the guarantees they provide, originate in the following several lemmas. In particular, the sets $\mathcal{F}_n$ will be chosen as *finite* sets, and as such one can easily verify that the selection in the $\operatorname{argmin}^{\varepsilon_n}$ in (10) can be chosen in a way that makes $\hat{f}_n$ a measurable function.

Lemma 20 *For any finite set $\mathcal{G}$ of bounded measurable functions $\mathcal{X} \to \mathbb{R}$, for any process $\mathbb{X}$, there exists a (nonrandom) nondecreasing sequence $\{m_n\}_{n=1}^{\infty}$ in $\mathbb{N}$ with $m_n \to \infty$ s.t.*

$$\lim_{n\to\infty} \mathbb{E}\left[\sup_{n' \geq n} \max_{g \in \mathcal{G}} \left| \hat{\mu}_{\mathbb{X}}(g) \;-\; \max_{m_n \leq m \leq n'} \frac{1}{m} \sum_{t=1}^{m} g(X_t) \right| \right] = 0.$$

Proof Fix any sequence $\mathbf{x} = \{x_t\}_{t=1}^{\infty}$ in $\mathcal{X}$ and any bounded function $g : \mathcal{X} \to \mathbb{R}$. By definition,

$$\hat{\mu}_{\mathbf{x}}(g) = \lim_{s\to\infty} \lim_{n\to\infty} \max_{s \leq m \leq n} \frac{1}{m} \sum_{t=1}^{m} g(x_t).$$

In particular, for each $s \in \mathbb{N}$, since $\max_{s\leq m\leq n} \frac{1}{m} \sum_{t=1}^{m} g(x_t)$ is nondecreasing in n, and g is bounded, $\lim_{n\to\infty} \max_{s\leq m\leq n} \frac{1}{m} \sum_{t=1}^{m} g(x_t)$ exists and is finite. This implies that, for each $s \in \mathbb{N}$, $\exists n_s^g(\mathbf{x}) \in \mathbb{N}$ s.t. $n_s^g(\mathbf{x}) \geq s$ and every $n \geq n_s^g(\mathbf{x})$ has

$$\max_{s\leq m\leq n} \frac{1}{m} \sum_{t=1}^{m} g(x_t) \leq \sup_{s \leq m < \infty} \frac{1}{m} \sum_{t=1}^{m} g(x_t) \leq 2^{-s} + \max_{s\leq m\leq n} \frac{1}{m} \sum_{t=1}^{m} g(x_t). \tag{11}$$

In particular, let us define $n_s^g(\mathbf{x})$ to be the minimal value in $\mathbb{N}$ with this property. We first argue that $n_s^g(\mathbf{x})$ is nondecreasing in s. To see this, first note that the left inequality in (11) is trivially satisfied for every $s, n \in \mathbb{N}$ with $n \geq s$. Moreover, for any $n, s \in \mathbb{N}$ with $s \geq 2$ and $n \geq n_s^g(\mathbf{x})$, either $\sup_{s-1\leq m<\infty} \frac{1}{m} \sum_{t=1}^{m} g(x_t)$ is achieved at $m = s-1$, in which case it is clearly less than $2^{-(s-1)} + \max_{s-1\leq m\leq n} \frac{1}{m} \sum_{t=1}^{m} g(x_t)$, or else $\sup_{s-1\leq m<\infty} \frac{1}{m} \sum_{t=1}^{m} g(x_t) = \sup_{s\leq m<\infty} \frac{1}{m} \sum_{t=1}^{m} g(x_t)$, in which case (since $n \geq n_s^g(\mathbf{x})$) it is at most $2^{-s} + \max_{s\leq m\leq n} \frac{1}{m} \sum_{t=1}^{m} g(x_t) \leq 2^{-(s-1)} + \max_{s-1\leq m\leq n} \frac{1}{m} \sum_{t=1}^{m} g(x_t)$. Furthermore, we have $n_s^g(\mathbf{x}) \geq s \geq s-1$. Altogether, we have $n_{s-1}^g(\mathbf{x}) \leq n_s^g(\mathbf{x})$, so that $n_s^g(\mathbf{x})$ is indeed nondecreasing in s.

For each $n \in \mathbb{N}$ with $n \geq n_1^g(\mathbf{x})$, let $s_n^g(\mathbf{x}) = \max\{s \in \{1,\ldots,n\} : n \geq n_s^g(\mathbf{x})\}$; for completeness, let $s_n^g(\mathbf{x}) = 0$ for $n < n_1^g(\mathbf{x})$. Then, for any finite set $\mathcal{G}$ of bounded functions $\mathcal{X} \to \mathbb{R}$, define $s_n^{\mathcal{G}}(\mathbf{x}) = \min_{g\in\mathcal{G}} s_n^g(\mathbf{x}) = \max\left\{ s \in \{1,\ldots,n\} : n \geq \max_{g\in\mathcal{G}} n_s^g(\mathbf{x}) \right\} \cup \{0\}$. Since $n_s^g(\mathbf{x})$ is nondecreasing in s, for any $n, n' \in \mathbb{N}$ with $n' \geq n$, for $1 \leq s \leq s_n^{\mathcal{G}}(\mathbf{x})$, every $g \in \mathcal{G}$ has $n' \geq n_s^g(\mathbf{x})$, so that (11) is satisfied for every $g \in \mathcal{G}$, which implies

$$\max_{g\in\mathcal{G}} \left| \sup_{s\leq m<\infty} \frac{1}{m} \sum_{t=1}^{m} g(x_t) - \max_{s\leq m\leq n'} \frac{1}{m} \sum_{t=1}^{m} g(x_t) \right| \leq 2^{-s}.$$

Therefore, for any sequence $s_n \to \infty$ such that $\exists n_0 \in \mathbb{N}$ with $1 \leq s_n \leq s_n^{\mathcal{G}}(\mathbf{x})$ for all $n \geq n_0$, we have

$$\lim_{n\to\infty} \sup_{n' \geq n} \max_{g \in \mathcal{G}} \left| \sup_{s_n \leq m < \infty} \frac{1}{m} \sum_{t=1}^{m} g(x_t) \; - \max_{s_n \leq m \leq n'} \frac{1}{m} \sum_{t=1}^{m} g(x_t) \right| \leq \lim_{n\to\infty} 2^{-s_n} = 0.$$

Furthermore, since each $s \in \mathbb{N}$ and $g \in \mathcal{G}$ have $n_s^g(\mathbf{x}) < \infty$, and $\mathcal{G}$ is a finite set, we have $s_n^{\mathcal{G}}(\mathbf{x}) \to \infty$, so that such sequences s_n do indeed exist. Furthermore, for any such sequence s_n, for every $g \in \mathcal{G}$,

$$\lim_{n\to\infty} \sup_{s_n \leq m < \infty} \frac{1}{m} \sum_{t=1}^{m} g(x_t) = \hat{\mu}_{\mathbf{x}}(g),$$

and since $\mathcal{G}$ has finite cardinality, this implies

$$\lim_{n\to\infty} \max_{g \in \mathcal{G}} \left| \hat{\mu}_{\mathbf{x}}(g) \; - \sup_{s_n \leq m < \infty} \frac{1}{m} \sum_{t=1}^{m} g(x_t) \right| = \max_{g \in \mathcal{G}} \lim_{n\to\infty} \left| \hat{\mu}_{\mathbf{x}}(g) \; - \sup_{s_n \leq m < \infty} \frac{1}{m} \sum_{t=1}^{m} g(x_t) \right| = 0.$$

Altogether, the triangle inequality implies

$$\begin{aligned} &\lim_{n\to\infty} \sup_{n' \geq n} \max_{g \in \mathcal{G}} \left| \hat{\mu}_{\mathbf{x}}(g) \; - \max_{s_n \leq m \leq n'} \frac{1}{m} \sum_{t=1}^{m} g(x_t) \right| \\ &\leq \lim_{n\to\infty} \sup_{n' \geq n} \max_{g \in \mathcal{G}} \left| \sup_{s_n \leq m < \infty} \frac{1}{m} \sum_{t=1}^{m} g(x_t) \; - \max_{s_n \leq m \leq n'} \frac{1}{m} \sum_{t=1}^{m} g(x_t) \right| \\ &\qquad + \lim_{n\to\infty} \max_{g \in \mathcal{G}} \left| \hat{\mu}_{\mathbf{x}}(g) \; - \sup_{s_n \leq m < \infty} \frac{1}{m} \sum_{t=1}^{m} g(x_t) \right| = 0. \end{aligned} \tag{12}$$

Next, suppose the bounded functions in the set $\mathcal{G}$ are measurable. Note that this implies that, for any $g \in \mathcal{G}$, the set of sequences $\mathbf{x}$ satisfying (11) for a given $s, n \in \mathbb{N}$ is a measurable subset of $\mathcal{X}^\infty$, so that for each $s, n' \in \mathbb{N}$ the set of sequences $\mathbf{x}$ with $n_s^g(\mathbf{x}) = n'$ is also a measurable set, so that n_s^g is a measurable function. Since the value of s_n^g is obtained from the values n_s^g via operations that preserve measurability, we also have that s_n^g is a measurable function. Since the minimum of a finite set of measurable functions is also measurable, we also have that $s_n^{\mathcal{G}}$ is a measurable function.

Now fix any process $\mathbb{X}$, and for any $n \in \mathbb{N}$ and $\delta \in (0,1)$ let

$$s_n^{\mathcal{G}}(\delta) = \max\left\{ s \in \{0, 1, \ldots, n\} : \mathbb{P}(s_n^{\mathcal{G}}(\mathbb{X}) \geq s) \geq 1 - \delta \right\}.$$

Since $s_n^{\mathcal{G}}(\mathbf{x})$ is nondecreasing for each sequence $\mathbf{x}$, we must also have that $s_n^{\mathcal{G}}(\delta)$ is nondecreasing in n. Furthermore, since each $\mathbf{x} \in \mathcal{X}^\infty$ has $s_n^{\mathcal{G}}(\mathbf{x}) \to \infty$, by continuity of probability measures (e.g., Schervish, 1995, Theorem A.19), $\forall s \in \mathbb{N}$, $\lim_{n\to\infty} \mathbb{P}(s_n^{\mathcal{G}}(\mathbb{X}) < s) = 0$. We therefore have $s_n^{\mathcal{G}}(\delta) \to \infty$ for any $\delta \in (0,1)$. In particular, letting

$$s_n = \max\left(\left\{ s \in \mathbb{N} : s_n^{\mathcal{G}}(2^{-s}) \geq s \right\} \cup \{0\} \right)$$

for each $n \in \mathbb{N}$, we have that s_n is nondecreasing, and $s_n \to \infty$. Furthermore, by definition, we have $\mathbb{P}(s_n^{\mathcal{G}}(\mathbb{X}) \geq s_n) \geq 1 - 2^{-s_n}$. Let $n_1 = 1$, and let $n_2, n_3, \ldots$ denote the increasing subsequence of all values $n \in \mathbb{N} \setminus \{1\}$ for which $s_n > s_{n-1}$; since $s_n \to \infty$ while each n has $s_n < \infty$, there are indeed an infinite number of such n_k values. Note that, since s_n is nondecreasing, and hence these s_{n_k} are each distinct values in $\mathbb{N} \cup \{0\}$, we have

$$\sum_{k=1}^{\infty} \mathbb{P}(s_{n_k}^{\mathcal{G}}(\mathbb{X}) < s_{n_k}) \leq \sum_{k=1}^{\infty} 2^{-s_{n_k}} \leq \sum_{i=0}^{\infty} 2^{-i} = 2 < \infty.$$

Therefore, the Borel-Cantelli Lemma implies that, with probability one, for all sufficiently large k, $s_{n_k}^{\mathcal{G}}(\mathbb{X}) \geq s_{n_k}$. Furthermore, since $s_n^{\mathcal{G}}(\mathbb{X})$ is nondecreasing in n, and $s_n = s_{n_k}$ for all $n \in \{n_k, \ldots, n_{k+1} - 1\}$ (due to s_n nondecreasing), if $s_{n_k}^{\mathcal{G}}(\mathbb{X}) \geq s_{n_k}$ for a given $k \in \mathbb{N}$, then $s_n^{\mathcal{G}}(\mathbb{X}) \geq s_n$ for every $n \in \{n_k, \ldots, n_{k+1} - 1\}$. Thus, we may conclude that, with probability one, for all sufficiently large $n \in \mathbb{N}$, $s_n^{\mathcal{G}}(\mathbb{X}) \geq s_n \geq 1$. Therefore, (12) implies

$$\lim_{n\to\infty} \sup_{n' \geq n} \max_{g \in \mathcal{G}} \left| \hat{\mu}_{\mathbb{X}}(g) - \max_{s_n \leq m \leq n'} \frac{1}{m} \sum_{t=1}^{m} g(X_t) \right| = 0 \text{ (a.s.)}. \tag{13}$$

Finally, since the functions in $\mathcal{G}$ are bounded and $\mathcal{G}$ has finite cardinality,

$$\left\{ \sup_{n' \geq n} \max_{g \in \mathcal{G}} \left| \hat{\mu}_{\mathbb{X}}(g) - \max_{s_n \leq m \leq n'} \frac{1}{m} \sum_{t=1}^{m} g(X_t) \right| \right\}_{n=1}^{\infty}$$

is a uniformly bounded sequence of random variables, so that combining (13) with the dominated convergence theorem implies

$$\lim_{n\to\infty} \mathbb{E}\left[\sup_{n' \geq n} \max_{g \in \mathcal{G}} \left| \hat{\mu}_{\mathbb{X}}(g) - \max_{s_n \leq m \leq n'} \frac{1}{m} \sum_{t=1}^{m} g(X_t) \right| \right] = 0.$$

The result now follows by taking $m_n = s_n$ for all $n \in \mathbb{N}$. ■

Lemma 21 *Suppose $\{\mathcal{G}_i\}_{i=1}^{\infty}$ is a sequence of nonempty finite sets of bounded measurable functions $\mathcal{X} \to \mathbb{R}$, with $\mathcal{G}_1 \subseteq \mathcal{G}_2 \subseteq \cdots$, and $\{\gamma_i\}_{i=1}^{\infty}$ is a sequence in $(0, \infty)$ with $\gamma_1 \geq \max_{g \in \mathcal{G}_1} \left(\sup_{x \in \mathcal{X}} g(x) - \inf_{x \in \mathcal{X}} g(x) \right)$. Then for any process $\mathbb{X}$, there exist (nonrandom) nondecreasing sequences $\{m_i\}_{i=1}^{\infty}$ and $\{i_n\}_{n=1}^{\infty}$ in $\mathbb{N}$ with $m_i \to \infty$ and $i_n \to \infty$ such that $\forall n \in \mathbb{N}$,*

$$\mathbb{E}\left[\max_{g \in \mathcal{G}_{i_n}} \left| \hat{\mu}_{\mathbb{X}}(g) - \max_{m_{i_n} \leq m \leq n} \frac{1}{m} \sum_{t=1}^{m} g(X_t) \right| \right] \leq \gamma_{i_n}.$$

Proof For each $i \in \mathbb{N}$, let $\{m_{i,n}\}_{n=1}^{\infty}$ denote a nondecreasing sequence in $\mathbb{N}$ with $\lim_{n\to\infty} m_{i,n} = \infty$ and

$$\lim_{n\to\infty} \mathbb{E}\left[\sup_{n' \geq n} \max_{g \in \mathcal{G}_i} \left| \hat{\mu}_{\mathbb{X}}(g) - \max_{m_{i,n} \leq m \leq n'} \frac{1}{m} \sum_{t=1}^{m} g(X_t) \right| \right] = 0. \tag{14}$$

Such a sequence is guaranteed to exist by Lemma 20. Now for each $n \in \mathbb{N}$, define

$$j_n = \max\left\{ i \in \{1, \ldots, n\} : \forall i' \leq i, \sup_{n'' \geq n} \mathbb{E}\left[\sup_{n' \geq n''} \max_{g \in \mathcal{G}_{i'}} \left| \hat{\mu}_{\mathbb{X}}(g) - \max_{m_{i',n''} \leq m \leq n'} \frac{1}{m} \sum_{t=1}^{m} g(X_t) \right| \right] \leq \gamma_{i'} \right\}.$$

First note that the set on the right hand side is nonempty, since every $n'' \in \mathbb{N}$ has

$$\mathbb{E}\left[\sup_{n' \geq n''} \max_{g \in \mathcal{G}_1} \left| \hat{\mu}_{\mathbb{X}}(g) \ - \max_{m_{1,n''} \leq m \leq n'} \frac{1}{m} \sum_{t=1}^{m} g(X_t) \right| \right] \leq \max_{g \in \mathcal{G}_1} \left(\sup_{x \in \mathcal{X}} g(x) - \inf_{x \in \mathcal{X}} g(x) \right) \leq \gamma_1.$$

Thus, j_n is well-defined for every $n \in \mathbb{N}$. In particular, by this definition, we have $\forall n \in \mathbb{N}$, $\forall i \in \{1, \ldots, j_n\}$,

$$\mathbb{E}\left[\sup_{n' \geq n} \max_{g \in \mathcal{G}_i} \left| \hat{\mu}_{\mathbb{X}}(g) \ - \max_{m_{i,n} \leq m \leq n'} \frac{1}{m} \sum_{t=1}^{m} g(X_t) \right| \right] \leq \gamma_i. \tag{15}$$

Furthermore, since

$$\sup_{n'' \geq n} \mathbb{E}\left[\sup_{n' \geq n''} \max_{g \in \mathcal{G}_i} \left| \hat{\mu}_{\mathbb{X}}(g) \ - \max_{m_{i,n''} \leq m \leq n'} \frac{1}{m} \sum_{t=1}^{m} g(X_t) \right| \right]$$

is nonincreasing in n for every $i \in \mathbb{N}$, we have that j_n is nondecreasing. Also note that, for any $i \in \mathbb{N}$, since $\gamma_i > 0$, (14) implies that $\exists n_i' \in \mathbb{N}$ such that, $\forall n \geq n_i'$,

$$\sup_{n'' \geq n} \mathbb{E}\left[\sup_{n' \geq n''} \max_{g \in \mathcal{G}_i} \left| \hat{\mu}_{\mathbb{X}}(g) \ - \max_{m_{i,n''} \leq m \leq n'} \frac{1}{m} \sum_{t=1}^{m} g(X_t) \right| \right] \leq \gamma_i.$$

Therefore $j_n \geq i$ for every $n \geq \max\left\{ i, \max_{1 \leq i' \leq i} n_{i'}' \right\}$. Since this is true of every $i \in \mathbb{N}$, we have that $j_n \to \infty$.

Next, let $n_1 = 1$, and for each $i \in \mathbb{N} \setminus \{1\}$, inductively define

$$n_i = \min\left\{ n \in \mathbb{N} : j_n \geq i, \min_{1 \leq j \leq i} m_{j,n} > m_{i-1,n_{i-1}} \right\}.$$

Note that, given the value $n_{i-1} \in \mathbb{N}$, the value n_i is well-defined since $j_n \to \infty$ and $\lim_{n \to \infty} \min_{1 \leq j \leq i} m_{j,n} = \min_{1 \leq j \leq i} \lim_{n \to \infty} m_{j,n} = \infty$. Thus, by induction, n_i is a well-defined value in $\mathbb{N}$ for all $i \in \mathbb{N}$. For each $i \in \mathbb{N}$, define $m_i = m_{i,n_i}$. In particular, by definition of n_i, for all $i \in \mathbb{N}$ we have $m_{i+1} \geq \min_{1 \leq j \leq i+1} m_{j,n_{i+1}} > m_{i,n_i} = m_i$, so that m_i is increasing, with $m_i \to \infty$. Finally, for each $n \in \mathbb{N}$, define $i_n = \max\{i \in \{1, \ldots, n\} : n_i \leq n\}$. Since $n_1 = 1$, $\{i \in \{1, \ldots, n\} : n_i \leq n\} \neq \emptyset$ for all $n \in \mathbb{N}$, so that i_n is a well-defined value in $\mathbb{N}$ for all $n \in \mathbb{N}$. Also, any $i \in \{1, \ldots, n\}$ with $n_i \leq n$ also has $n_i \leq n+1$, so that i_n is nondecreasing in n. Furthermore, since $n_i < \infty$ for every $i \in \mathbb{N}$, we have $i_n \to \infty$. Also note that, $\forall n \in \mathbb{N}$, $n \geq n_{i_n}$. Thus, for every $n \in \mathbb{N}$,

$$\mathbb{E}\left[\max_{g \in \mathcal{G}_{i_n}} \left| \hat{\mu}_{\mathbb{X}}(g) - \max_{m_{i_n} \leq m \leq n} \frac{1}{m} \sum_{t=1}^{m} g(X_t) \right| \right] \leq \mathbb{E}\left[\sup_{n' \geq n_{i_n}} \max_{g \in \mathcal{G}_{i_n}} \left| \hat{\mu}_{\mathbb{X}}(g) - \max_{m_{i_n,n_{i_n}} \leq m \leq n'} \frac{1}{m} \sum_{t=1}^{m} g(X_t) \right| \right].$$

By definition of n_{i_n}, we have $j_{n_{i_n}} \geq i_n$ (this is immediate from the n_i definition if $i_n \geq 2$, and is also trivially true for $i_n = 1$ since $j_1 \geq 1$), so that (15) implies the rightmost expression above is at most γ_{i_n}, which completes the proof. ■

The following lemma represents the first use of Condition 1 in the proof of sufficiency of Condition 1 for strong universal inductive learning.

Lemma 22 *There exists a countable set* $\mathcal{T}_1 \subseteq \mathcal{B}$ *such that,* $\forall \mathbb{X} \in \mathcal{C}_1$, $\forall A \in \mathcal{B}$,

$$\inf_{G \in \mathcal{T}_1} \mathbb{E}[\hat{\mu}_{\mathbb{X}}(G \triangle A)] = 0.$$

Proof By assumption, $\mathcal{B}$ is generated by a separable metrizable topology $\mathcal{T}$, and since every separable metrizable topological space is second countable (see Srivastava, 1998, Proposition 2.1.9), we have that there exists a *countable* set $\mathcal{T}_0 \subseteq \mathcal{T}$ such that, $\forall A \in \mathcal{T}$, $\exists \mathcal{A} \subseteq \mathcal{T}_0$ s.t. $A = \bigcup \mathcal{A}$. Now from this, there is an immediate proof of the lemma if we were to take $\mathcal{T}_1$ as the algebra generated by $\mathcal{T}_0$ (which is a countable set) via the monotone class theorem (Ash and Doléans-Dade, 2000, Theorem 1.3.9), using Condition 1 to argue that the sets A satisfying the claim in the lemma form a monotone class. However, here we will instead establish the lemma with a *smaller* choice of the set $\mathcal{T}_1$, which thereby simplifies the problem of implementing the resulting learning rule. Specifically, we take $\mathcal{T}_1 = \{\bigcup \mathcal{A} : \mathcal{A} \subseteq \mathcal{T}_0, |\mathcal{A}| < \infty\}$: all finite unions of sets in $\mathcal{T}_0$. Note that, given an indexing of $\mathcal{T}_0$ by $\mathbb{N}$, each $A \in \mathcal{T}_1$ can be indexed by a finite subset of $\mathbb{N}$ (the indices of elements of the corresponding $\mathcal{A}$), of which there are countably many, so that $\mathcal{T}_1$ is countable. Now fix any $\mathbb{X} \in \mathcal{C}_1$ and let

$$\Lambda = \left\{ A \in \mathcal{B} : \inf_{G \in \mathcal{T}_1} \mathbb{E}[\hat{\mu}_{\mathbb{X}}(G \triangle A)] = 0 \right\}.$$

We will prove that $\Lambda = \mathcal{B}$ by establishing that $\mathcal{T} \subseteq \Lambda$ and that Λ is a σ-algebra.

First consider any $A \in \mathcal{T}$. As mentioned above, $\exists \{B_i\}_{i=1}^{\infty}$ in $\mathcal{T}_0$ such that $A = \bigcup_{i=1}^{\infty} B_i$. But then letting $A_k = \bigcup_{i=1}^{k} B_i$ for each $k \in \mathbb{N}$, we have $A_k \triangle A = A \setminus A_k \downarrow \emptyset$, and $A_k \in \mathcal{T}_1$ for each $k \in \mathbb{N}$. Therefore, $\inf_{G \in \mathcal{T}_1} \mathbb{E}\left[\hat{\mu}_{\mathbb{X}}(G \triangle A)\right] \leq \lim_{k \to \infty} \mathbb{E}\left[\hat{\mu}_{\mathbb{X}}(A_k \triangle A)\right]$, and the right hand side equals 0 by Condition 1. Together with nonnegativity of $\inf_{G \in \mathcal{T}_1} \mathbb{E}\left[\hat{\mu}_{\mathbb{X}}(G \triangle A)\right]$ (Lemma 9), this implies $A \in \Lambda$. Since this holds for any $A \in \mathcal{T}$, we have $\mathcal{T} \subseteq \Lambda$.

Next, we argue that Λ is a σ-algebra. We begin by showing it is closed under complements. Toward this end, consider any $A \in \Lambda$, and for any $k \in \mathbb{N}$ denote by G_k an element of $\mathcal{T}_1$ with $\mathbb{E}\left[\hat{\mu}_{\mathbb{X}}(G_k \triangle A)\right] < 1/k$ (guaranteed to exist by the definition of Λ). Since $G_k \in \mathcal{T}_1 \subseteq \mathcal{T}$, it follows that $\mathcal{X} \setminus G_k$ is a closed set. Therefore, since $(\mathcal{X}, \mathcal{T})$ is metrizable, $\exists \{B_{ki}\}_{i=1}^{\infty}$ in $\mathcal{T}$ such that $\mathcal{X} \setminus G_k = \bigcap_{i=1}^{\infty} B_{ki}$ (Kechris, 1995, Proposition 3.7). Denoting $C_{kj} = \bigcap_{i=1}^{j} B_{ki}$ for each $j \in \mathbb{N}$, we have that $C_{kj} \triangle (\mathcal{X} \setminus G_k) = C_{kj} \setminus (\mathcal{X} \setminus G_k) \downarrow \emptyset$ as $j \to \infty$, and $C_{kj} \in \mathcal{T}$ for each $j \in \mathbb{N}$. In particular, by Condition 1, $\exists j_k \in \mathbb{N}$ such that $\mathbb{E}[\hat{\mu}_{\mathbb{X}}(C_{kj_k} \triangle (\mathcal{X} \setminus G_k))] < 1/k$. Also, since $C_{kj_k} \in \mathcal{T}$, and we proved above that $\mathcal{T} \subseteq \Lambda$, $\exists D_k \in \mathcal{T}_1$ such that $\mathbb{E}\left[\hat{\mu}_{\mathbb{X}}(D_k \triangle C_{kj_k})\right] < 1/k$. Together with the facts that $D_k \triangle (\mathcal{X} \setminus A) \subseteq (D_k \triangle C_{kj_k}) \cup (C_{kj_k} \triangle (\mathcal{X} \setminus G_k)) \cup ((\mathcal{X} \setminus G_k) \triangle (\mathcal{X} \setminus A))$ and $(\mathcal{X} \setminus G_k) \triangle (\mathcal{X} \setminus A) = G_k \triangle A$, we have that

$$\mathbb{E}[\hat{\mu}_{\mathbb{X}}(D_k \triangle (\mathcal{X} \setminus A))] \leq \mathbb{E}[\hat{\mu}_{\mathbb{X}}(D_k \triangle C_{kj_k})] + \mathbb{E}[\hat{\mu}_{\mathbb{X}}(C_{kj_k} \triangle (\mathcal{X} \setminus G_k))] + \mathbb{E}[\hat{\mu}_{\mathbb{X}}(G_k \triangle A)] < 3/k,$$

where the first inequality is due to Lemma 11. Since $D_k \in \mathcal{T}_1$, and this argument holds for any $k \in \mathbb{N}$, we have

$$\inf_{G \in \mathcal{T}_1} \mathbb{E}[\hat{\mu}_{\mathbb{X}}(G \triangle (\mathcal{X} \setminus A))] \leq \inf_{k \in \mathbb{N}} 3/k = 0.$$

Together with nonnegativity of the left hand side (Lemma 9), this implies $\mathcal{X} \setminus A \in \Lambda$. Thus, Λ is closed under complements.

Next, we argue that Λ is closed under countable unions. Let $\{A_i\}_{i=1}^{\infty}$ be a sequence in Λ, denote $A = \bigcup_{i=1}^{\infty} A_i$, and fix any $\varepsilon > 0$. Denoting $B_k = \bigcup_{i=1}^{k} A_i$ for each $k \in \mathbb{N}$, we have $B_k \triangle A = A \setminus B_k \downarrow \emptyset$. Therefore, Condition 1 implies $\exists k_\varepsilon \in \mathbb{N}$ such that $\mathbb{E}[\hat{\mu}_{\mathbb{X}}(B_{k_\varepsilon} \triangle A)] < \varepsilon$. Next, for each $i \in \mathbb{N}$, let G_i be an element of $\mathcal{T}_1$ with $\mathbb{E}[\hat{\mu}_{\mathbb{X}}(G_i \triangle A_i)] < \varepsilon / k_\varepsilon$ (guaranteed to exist, since $A_i \in \Lambda$). Denote by $C_{k_\varepsilon} = \bigcup_{i=1}^{k_\varepsilon} G_i$. Noting that it follows immediately from its definition that $\mathcal{T}_1$ is closed under finite unions, we have that $C_{k_\varepsilon} \in \mathcal{T}_1$. Then noting that

$$C_{k_\varepsilon} \triangle A \subseteq (B_{k_\varepsilon} \triangle A) \cup (C_{k_\varepsilon} \triangle B_{k_\varepsilon}) \subseteq (B_{k_\varepsilon} \triangle A) \cup \bigcup_{i=1}^{k_\varepsilon} (G_i \triangle A_i),$$

altogether we have that

$$\inf_{G \in \mathcal{T}_1} \mathbb{E}[\hat{\mu}_{\mathbb{X}}(G \triangle A)] \leq \mathbb{E}[\hat{\mu}_{\mathbb{X}}(C_{k_\varepsilon} \triangle A)] \leq \mathbb{E}[\hat{\mu}_{\mathbb{X}}(B_{k_\varepsilon} \triangle A)] + \sum_{i=1}^{k_\varepsilon} \mathbb{E}[\hat{\mu}_{\mathbb{X}}(G_i \triangle A_i)] < \varepsilon + \sum_{i=1}^{k_\varepsilon} \frac{\varepsilon}{k_\varepsilon} = 2\varepsilon,$$

where the second inequality is due to Lemma 11. Since this argument holds for any $\varepsilon > 0$, taking the limit as $\varepsilon \to 0$ reveals that $\inf_{G \in \mathcal{T}_1} \mathbb{E}[\hat{\mu}_{\mathbb{X}}(G \triangle A)] \leq 0$. Together with nonnegativity of the left hand side (Lemma 9), this implies $A \in \Lambda$. Thus, Λ is closed under countable unions.

Finally, recalling that $\mathcal{T}$ is a topology, by definition we have $\mathcal{X} \in \mathcal{T}$, and since $\mathcal{T} \subseteq \Lambda$, this implies $\mathcal{X} \in \Lambda$. Altogether, we have established that Λ is a σ-algebra. Therefore, since $\mathcal{B}$ is the σ-algebra generated by $\mathcal{T}$, and $\mathcal{T} \subseteq \Lambda$, it immediately follows that $\mathcal{B} \subseteq \Lambda$ (which also implies $\Lambda = \mathcal{B}$). Since this argument holds for any choice of $\mathbb{X} \in \mathcal{C}_1$, the lemma immediately follows. ■

For example, in the special case of $\mathcal{X} = \mathbb{R}^p$ ($p \in \mathbb{N}$) with the Euclidean topology, the above proof implies it suffices to take the set $\mathcal{T}_1$ as the finite unions of rational-centered rational-radius open balls. Now, continuing with the general case, the next lemma extends Lemma 22 from set approximation to function approximation, again using Condition 1.

Lemma 23 *There exists a sequence $\{\mathcal{F}_i\}_{i=1}^{\infty}$ of nonempty finite sets of measurable functions $\mathcal{X} \to \mathcal{Y}$ with $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots$ such that, for every $\mathbb{X} \in \mathcal{C}_1$, for every measurable $f : \mathcal{X} \to \mathcal{Y}$,*

$$\lim_{i \to \infty} \min_{f_i \in \mathcal{F}_i} \mathbb{E}[\hat{\mu}_{\mathbb{X}}(\ell(f_i(\cdot), f(\cdot)))] = 0.$$

Proof We will first prove that there exists a countable set $\tilde{\mathcal{F}}$ of measurable functions $\mathcal{X} \to \mathcal{Y}$ such that, for every $\mathbb{X} \in \mathcal{C}_1$, $\forall \varepsilon > 0$, for every measurable $f : \mathcal{X} \to \mathcal{Y}$, $\exists \tilde{f} \in \tilde{\mathcal{F}}$ s.t. $\mathbb{E}\Big[\hat{\mu}_{\mathbb{X}}(\ell(\tilde{f}(\cdot), f(\cdot)))\Big] < 3\varepsilon$. Let $\mathcal{T}_1$ be as in Lemma 22, and let $\tilde{\mathcal{Y}} \subseteq \mathcal{Y}$ be a countable set with $\sup_{y \in \mathcal{Y}} \inf_{\tilde{y} \in \tilde{\mathcal{Y}}} \ell(y, \tilde{y}) = 0$; this must exist, by the assumption that $(\mathcal{Y}, \ell)$ is separable. Fix some arbitrary value $y_0 \in \mathcal{Y}$, and let $A_0 = \mathcal{X}$. For any $k \in \mathbb{N}$, values $y_1, \ldots, y_k \in \mathcal{Y}$, and sets $A_1, \ldots, A_k \in \mathcal{B}$, for any $x \in \mathcal{X}$, define $\tilde{f}(x; \{y_i\}_{i=1}^k, \{A_i\}_{i=1}^k) = y_{\max\{j \in \{0,\ldots,k\} : x \in A_j\}}$; one can easily verify that $\tilde{f}(\cdot; \{y_i\}_{i=1}^k, \{A_i\}_{i=1}^k)$ is a measurable function (indeed, it is a *simple* function). Define

$$\tilde{\mathcal{F}} = \left\{ \tilde{f}(\cdot; \{y_i\}_{i=1}^k, \{A_i\}_{i=1}^k) : k \in \mathbb{N}, \forall i \leq k, y_i \in \tilde{\mathcal{Y}}, A_i \in \mathcal{T}_1 \right\},$$

and note that, given an indexing of $\tilde{\mathcal{Y}}$ and $\mathcal{T}_1$ by $\mathbb{N}$, we can index $\tilde{\mathcal{F}}$ by finite tuples of integers (the indices of the corresponding y_i and A_i values), of which there are countably many, so that $\tilde{\mathcal{F}}$ is countable.

Enumerate the elements of $\tilde{\mathcal{Y}}$ as $\tilde{y}_1, \tilde{y}_2, \ldots$ (for simplicity of notation, we suppose this sequence is infinite; otherwise, we can simply repeat the elements to get an infinite sequence). For each $\varepsilon > 0$, let $B_{\varepsilon,1} = \{y \in \mathcal{Y} : \ell(y, \tilde{y}_1) \leq \varepsilon\}$, and for each integer $i \geq 2$, inductively define $B_{\varepsilon,i} = \{y \in \mathcal{Y} : \ell(y, \tilde{y}_i) \leq \varepsilon\} \setminus \bigcup_{j=1}^{i-1} B_{\varepsilon,j}$. Note that the sets $B_{\varepsilon,i}$ are measurable and disjoint over $i \in \mathbb{N}$, and that $\bigcup_{i=1}^{\infty} B_{\varepsilon,i} = \mathcal{Y}$.

Now fix any $\mathbb{X} \in \mathcal{C}_1$, any measurable $f : \mathcal{X} \to \mathcal{Y}$, and any $\varepsilon > 0$. For each $i \in \mathbb{N}$, define $C_{\varepsilon,i} = f^{-1}(B_{\varepsilon,i})$, which is an element of $\mathcal{B}$ by measurability of f and $B_{\varepsilon,i}$. Note that $\bigcup_{i=1}^{\infty} C_{\varepsilon,i} = f^{-1}\left(\bigcup_{i=1}^{\infty} B_{\varepsilon,i}\right) = f^{-1}(\mathcal{Y}) = \mathcal{X}$, and furthermore that (since the $B_{\varepsilon,i}$ sets are disjoint) the sets $C_{\varepsilon,i}$ are disjoint over $i \in \mathbb{N}$. It follows that $\lim_{k\to\infty} \bigcup_{i=k}^{\infty} C_{\varepsilon,i} = \emptyset$, with $\bigcup_{i=k}^{\infty} C_{\varepsilon,i}$ nonincreasing in k, so that Condition 1 entails $\lim_{k\to\infty} \mathbb{E}\left[\hat{\mu}_{\mathbb{X}}\left(\bigcup_{i=k}^{\infty} C_{\varepsilon,i}\right)\right] = 0$. In particular, this implies $\exists k_\varepsilon \in \mathbb{N}$ such that $\mathbb{E}\left[\hat{\mu}_{\mathbb{X}}\left(\bigcup_{i=k_\varepsilon+1}^{\infty} C_{\varepsilon,i}\right)\right] < \varepsilon/\bar{\ell}$.

For each $i \in \{1, \ldots, k_\varepsilon\}$, let $A_{\varepsilon,i} \in \mathcal{T}_1$ be a set with $\mathbb{E}[\hat{\mu}_{\mathbb{X}}(A_{\varepsilon,i} \triangle C_{\varepsilon,i})] < \varepsilon/(k_\varepsilon \bar{\ell})$, which exists by the defining property of $\mathcal{T}_1$ from Lemma 22. Finally, let

$$\tilde{f}_\varepsilon(\cdot) = \tilde{f}\left(\cdot; \{\tilde{y}_i\}_{i=1}^{k_\varepsilon}, \{A_{\varepsilon,i}\}_{i=1}^{k_\varepsilon}\right),$$

and note that $\tilde{f}_\varepsilon \in \tilde{\mathcal{F}}$. Furthermore, for any $x \in \mathcal{X} = \bigcup_{i=1}^{\infty} C_{\varepsilon,i}$,

$$\begin{aligned}
\ell(f(x), \tilde{f}_\varepsilon(x)) &\leq \bar{\ell}\mathbb{1}_{\bigcup_{i=k_\varepsilon+1}^{\infty} C_{\varepsilon,i}}(x) + \sum_{i=1}^{k_\varepsilon} \ell(f(x), \tilde{f}_\varepsilon(x))\mathbb{1}_{C_{\varepsilon,i}}(x) \\
&\leq \bar{\ell}\mathbb{1}_{\bigcup_{i=k_\varepsilon+1}^{\infty} C_{\varepsilon,i}}(x) + \sum_{i=1}^{k_\varepsilon} \Big(\ell(f(x), \tilde{y}_i)\mathbb{1}_{C_{\varepsilon,i}}(x) + \ell(\tilde{y}_i, \tilde{f}_\varepsilon(x))\mathbb{1}_{C_{\varepsilon,i}}(x)\Big) \\
&\leq \bar{\ell}\mathbb{1}_{\bigcup_{i=k_\varepsilon+1}^{\infty} C_{\varepsilon,i}}(x) + \varepsilon + \sum_{i=1}^{k_\varepsilon} \ell(\tilde{y}_i, \tilde{f}_\varepsilon(x))\mathbb{1}_{C_{\varepsilon,i}}(x). \qquad (16)
\end{aligned}$$

Denote $[k_\varepsilon] = \{1, \ldots, k_\varepsilon\}$. Since $\ell(\tilde{y}_i, \tilde{f}_\varepsilon(x)) = 0$ if $x \in A_{\varepsilon,i} \setminus \bigcup_{j \in [k_\varepsilon] \setminus \{i\}} A_{\varepsilon,j}$,

$$\sum_{i=1}^{k_\varepsilon} \ell(\tilde{y}_i, \tilde{f}_\varepsilon(x)) \mathbb{1}_{C_{\varepsilon,i}}(x) \leq \sum_{i=1}^{k_\varepsilon} \ell(\tilde{y}_i, \tilde{f}_\varepsilon(x)) \mathbb{1}_{(C_{\varepsilon,i} \setminus A_{\varepsilon,i}) \cup (C_{\varepsilon,i} \cap \bigcup_{j \in [k_\varepsilon] \setminus \{i\}} A_{\varepsilon,j})}(x)$$
$$\leq \sum_{i=1}^{k_\varepsilon} \ell(\tilde{y}_i, \tilde{f}_\varepsilon(x)) \left(\mathbb{1}_{C_{\varepsilon,i} \setminus A_{\varepsilon,i}}(x) + \sum_{j \in [k_\varepsilon] \setminus \{i\}} \mathbb{1}_{C_{\varepsilon,i} \cap A_{\varepsilon,j}}(x) \right).$$

Since $C_{\varepsilon,i} \cap C_{\varepsilon,j} = \emptyset$ for $j \neq i$, $C_{\varepsilon,i} \cap A_{\varepsilon,j} = C_{\varepsilon,i} \cap (A_{\varepsilon,j} \setminus C_{\varepsilon,j})$, so that the above equals

$$\sum_{i=1}^{k_\varepsilon} \ell(\tilde{y}_i, \tilde{f}_\varepsilon(x)) \left(\mathbb{1}_{C_{\varepsilon,i} \setminus A_{\varepsilon,i}}(x) + \sum_{j \in [k_\varepsilon] \setminus \{i\}} \mathbb{1}_{C_{\varepsilon,i} \cap (A_{\varepsilon,j} \setminus C_{\varepsilon,j})}(x) \right)$$
$$\leq \bar{\ell} \sum_{i=1}^{k_\varepsilon} \left(\mathbb{1}_{C_{\varepsilon,i} \setminus A_{\varepsilon,i}}(x) + \sum_{j \in [k_\varepsilon] \setminus \{i\}} \mathbb{1}_{C_{\varepsilon,i} \cap (A_{\varepsilon,j} \setminus C_{\varepsilon,j})}(x) \right). \tag{17}$$

Since

$$\sum_{i=1}^{k_\varepsilon} \sum_{j \in [k_\varepsilon] \setminus \{i\}} \mathbb{1}_{C_{\varepsilon,i} \cap (A_{\varepsilon,j} \setminus C_{\varepsilon,j})}(x) = \sum_{j=1}^{k_\varepsilon} \sum_{i \in [k_\varepsilon] \setminus \{j\}} \mathbb{1}_{C_{\varepsilon,i} \cap (A_{\varepsilon,j} \setminus C_{\varepsilon,j})}(x)$$
$$\leq \sum_{j=1}^{k_\varepsilon} \sum_{i=1}^{\infty} \mathbb{1}_{C_{\varepsilon,i} \cap (A_{\varepsilon,j} \setminus C_{\varepsilon,j})}(x) = \sum_{j=1}^{k_\varepsilon} \mathbb{1}_{A_{\varepsilon,j} \setminus C_{\varepsilon,j}}(x),$$

the expression on the right hand side in (17) is at most

$$\bar{\ell} \left(\sum_{i=1}^{k_\varepsilon} \mathbb{1}_{C_{\varepsilon,i} \setminus A_{\varepsilon,i}}(x) + \sum_{j=1}^{k_\varepsilon} \mathbb{1}_{A_{\varepsilon,j} \setminus C_{\varepsilon,j}}(x) \right) = \bar{\ell} \sum_{i=1}^{k_\varepsilon} \mathbb{1}_{C_{\varepsilon,i} \triangle A_{\varepsilon,i}}(x).$$

Plugging this into (16) yields that

$$\ell(f(x), \tilde{f}_\varepsilon(x)) \leq \varepsilon + \bar{\ell} \mathbb{1}_{\bigcup_{i=k_\varepsilon+1}^{\infty} C_{\varepsilon,i}}(x) + \bar{\ell} \sum_{i=1}^{k_\varepsilon} \mathbb{1}_{C_{\varepsilon,i} \triangle A_{\varepsilon,i}}(x).$$

Therefore, by linearity of the expectation, together with monotonicity, homogeneity, and finite subadditivity of $\hat{\mu}_{\mathbb{X}}$ (Lemma 8),

$$\mathbb{E}\left[\hat{\mu}_{\mathbb{X}}\left(\ell\left(\tilde{f}_\varepsilon(\cdot), f(\cdot)\right)\right)\right] \leq \varepsilon + \bar{\ell} \mathbb{E}\left[\hat{\mu}_{\mathbb{X}}\left(\bigcup_{i=k_\varepsilon+1}^{\infty} C_{\varepsilon,i}\right)\right] + \bar{\ell} \sum_{i=1}^{k_\varepsilon} \mathbb{E}[\hat{\mu}_{\mathbb{X}}(C_{\varepsilon,i} \triangle A_{\varepsilon,i})] < 3\varepsilon.$$

To complete the proof, we enumerate the elements of $\tilde{\mathcal{F}} = \{\tilde{g}_1, \tilde{g}_2, \ldots\}$ and define $\mathcal{F}_i = \{\tilde{g}_1, \ldots, \tilde{g}_i\}$ for each $i \in \mathbb{N}$. Fix any measurable $f : \mathcal{X} \to \mathcal{Y}$, and for each $k \in \mathbb{N}$, let i_k

denote the index $i \in \mathbb{N}$ with $\tilde{g}_i = \tilde{f}_{(1/k)}$, for $\tilde{f}_{(1/k)} \in \tilde{\mathcal{F}}$ defined as above for this f. Then for any $i \geq i_k$,

$$\min_{f_i \in \mathcal{F}_i} \mathbb{E}[\hat{\mu}_{\mathbb{X}}(\ell(f_i(\cdot), f(\cdot)))] \leq \mathbb{E}\Big[\hat{\mu}_{\mathbb{X}}\Big(\ell\Big(\tilde{f}_{(1/k)}(\cdot), f(\cdot)\Big)\Big)\Big] < \frac{3}{k},$$

so that

$$\lim_{i\to\infty} \min_{f_i \in \mathcal{F}_i} \mathbb{E}[\hat{\mu}_{\mathbb{X}}(\ell(f_i(\cdot), f(\cdot)))] \leq \lim_{i\to\infty} \min_{k: i_k \leq i} \frac{3}{k} = 0.$$

The result now follows from nonnegativity of the left hand side (by nonnegativity of ℓ, and monotonicity of the expectation and $\hat{\mu}_{\mathbb{X}}$ from Lemma 8). ■

Additionally, we have the following property for the f-approximating sequences of sets $\mathcal{F}_i$ implied by Lemma 23.

Lemma 24 *Fix any process $\mathbb{X}$ on $\mathcal{X}$, any measurable function $f : \mathcal{X} \to \mathcal{Y}$, any non-decreasing sequence $\{u_i\}_{i=1}^{\infty}$ in $\mathbb{N}$ with $u_i \to \infty$, and any sequence $\{\mathcal{F}_i\}_{i=1}^{\infty}$ of finite sets of measurable functions $\mathcal{X} \to \mathcal{Y}$ with $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots$ such that $\lim_{i\to\infty} \min_{g \in \mathcal{F}_i} \mathbb{E}[\hat{\mu}_{\mathbb{X}}(\ell(g(\cdot), f(\cdot)))] = 0$. There exists a (nonrandom) sequence $\{f_i\}_{i=1}^{\infty}$, with $f_i \in \mathcal{F}_i$ for each $i \in \mathbb{N}$, and a (nonrandom) sequence $\{\alpha_i\}_{i=1}^{\infty}$ in $(0,\infty)$ with $\alpha_i \to 0$, such that, on an event K of probability one, $\exists \iota_0 \in \mathbb{N}$ such that $\forall i \geq \iota_0$,*

$$\sup_{u_i \leq m < \infty} \frac{1}{m} \sum_{t=1}^{m} \ell(f_i(X_t), f(X_t)) \leq \alpha_i.$$

Proof Let $\{g_i\}_{i=1}^{\infty}$ be a sequence with $g_i \in \mathcal{F}_i$ for each $i \in \mathbb{N}$, s.t. $\lim_{i\to\infty} \mathbb{E}[\hat{\mu}_{\mathbb{X}}(\ell(g_i(\cdot), f(\cdot)))] = 0$. Then $\forall k \in \mathbb{N}$, $\exists j_k \in \mathbb{N}$ such that $\mathbb{E}[\hat{\mu}_{\mathbb{X}}(\ell(g_{j_k}(\cdot), f(\cdot)))] < 4^{-k}\bar{\ell}$. Let us fix any sequence $\{j_k\}_{k=1}^{\infty}$ in $\mathbb{N}$ such that j_k has this property for every k. For completeness, also define $j_0 = 1$. Furthermore, since $u_i \to \infty$, the dominated convergence theorem implies that $\forall j \in \mathbb{N}$,

$$\begin{aligned}\lim_{i\to\infty} \mathbb{E}\left[\sup_{u_i \leq m < \infty} \frac{1}{m} \sum_{t=1}^{m} \ell(g_j(X_t), f(X_t))\right] &= \mathbb{E}\left[\lim_{i\to\infty} \sup_{u_i \leq m < \infty} \frac{1}{m} \sum_{t=1}^{m} \ell(g_j(X_t), f(X_t))\right] \\ &= \mathbb{E}\left[\limsup_{m\to\infty} \frac{1}{m} \sum_{t=1}^{m} \ell(g_j(X_t), f(X_t))\right] = \mathbb{E}[\hat{\mu}_{\mathbb{X}}(\ell(g_j(\cdot), f(\cdot)))] .\end{aligned}$$

In particular, this implies that $\forall k \in \mathbb{N}$, $\exists i_k \in \mathbb{N}$ such that

$$\mathbb{E}\left[\sup_{u_{i_k} \leq m < \infty} \frac{1}{m} \sum_{t=1}^{m} \ell(g_{j_k}(X_t), f(X_t))\right] \leq \mathbb{E}[\hat{\mu}_{\mathbb{X}}(\ell(g_{j_k}(\cdot), f(\cdot)))] + 4^{-k}\bar{\ell} < 2 \cdot 4^{-k}\bar{\ell}. \tag{18}$$

Also note that, since the leftmost expression in (18) is nonincreasing in i_k, we may choose $i_k > i_{k-1}$ if $k \geq 2$ (or $i_k > 1$ for $k = 1$). Thus, letting $i_0 = 1$, there exists a strictly increasing sequence $\{i_k\}_{k=0}^{\infty}$ in $\mathbb{N}$ such that i_k has the property (18) for every $k \in \mathbb{N}$. We may then note that, by Markov's inequality,

$$\begin{aligned}
&\sum_{k=0}^{\infty} \mathbb{P}\left(\sup_{u_{i_k} \leq m < \infty} \frac{1}{m} \sum_{t=1}^{m} \ell(g_{j_k}(X_t), f(X_t)) > 2^{(1/2)-k}\sqrt{\bar{\ell}} \right) \\
&\leq \sum_{k=0}^{\infty} \frac{1}{2^{(1/2)-k}\sqrt{\bar{\ell}}} \mathbb{E}\left[\sup_{u_{i_k} \leq m < \infty} \frac{1}{m} \sum_{t=1}^{m} \ell(g_{j_k}(X_t), f(X_t)) \right] \\
&\leq \sum_{k=0}^{\infty} \frac{1}{2^{(1/2)-k}\sqrt{\bar{\ell}}} 2 \cdot 4^{-k}\bar{\ell} = \sum_{k=0}^{\infty} 2^{(1/2)-k}\sqrt{\bar{\ell}} = 2^{3/2}\sqrt{\bar{\ell}} < \infty.
\end{aligned}$$

Therefore, by the Borel-Cantelli Lemma, there exists an event K of probability one, on which $\exists \kappa_0 \in \mathbb{N}$ such that, $\forall k \geq \kappa_0$,

$$\sup_{u_{i_k} \leq m < \infty} \frac{1}{m} \sum_{t=1}^{m} \ell(g_{j_k}(X_t), f(X_t)) \leq 2^{(1/2)-k}\sqrt{\bar{\ell}}. \tag{19}$$

Now, $\forall i \in \mathbb{N}$, define

$$k_i = \max\left\{ k \in \mathbb{N} \cup \{0\} : \max\{i_k, j_k\} \leq i \right\},$$

and let $\alpha_i = 2^{(1/2)-k_i}\sqrt{\bar{\ell}}$. To see that the value k_i is well-defined for every $i \in \mathbb{N}$, note that $\max\{i_0, j_0\} = 1 \leq i$, so that the set on the right hand side is nonempty, and furthermore, since $\{i_k\}_{k=0}^{\infty}$ is strictly increasing, every $k \geq i$ has $\max\{i_k, j_k\} > i$, so that the set is finite, and hence has a maximal element. Also, since i_k and j_k are finite for every k, we have that $\lim\limits_{i \to \infty} k_i = \infty$. In particular, this implies that, on the event K, $\exists \iota_0 \in \mathbb{N}$ such that $\forall i \geq \iota_0$, $k_i \geq \kappa_0$, so that (19) implies

$$\sup_{u_{i_{k_i}} \leq m < \infty} \frac{1}{m} \sum_{t=1}^{m} \ell(g_{j_{k_i}}(X_t), f(X_t)) \leq \alpha_i. \tag{20}$$

Now define $f_i = g_{j_{k_i}}$ for every $i \in \mathbb{N}$. Note that, since $j_{k_i} \leq i$ (by definition of k_i) and $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots$, we have $\mathcal{F}_{j_{k_i}} \subseteq \mathcal{F}_i$. In particular, since $f_i = g_{j_{k_i}} \in \mathcal{F}_{j_{k_i}}$ (by definition), this implies $f_i \in \mathcal{F}_i$ for every $i \in \mathbb{N}$. Also note that, since $i_{k_i} \leq i$ (by definition of k_i), and $\{u_t\}_{t=1}^{\infty}$ is a nondecreasing sequence, $u_{i_{k_i}} \leq u_i$ for every $i \in \mathbb{N}$. Together with (20), these facts imply that, on the event K, $\forall i \geq \iota_0$,

$$\sup_{u_i \leq m < \infty} \frac{1}{m} \sum_{t=1}^{m} \ell(f_i(X_t), f(X_t)) \leq \sup_{u_{i_{k_i}} \leq m < \infty} \frac{1}{m} \sum_{t=1}^{m} \ell(f_i(X_t), f(X_t)) \leq \alpha_i.$$

■

With these results in hand, we are finally ready for the proof of sufficiency of Condition 1 for strong universal inductive learning.

Lemma 25 $\mathcal{C}_1 \subseteq \mathrm{SUIL}$.

Proof Suppose $\mathbb{X} \in \mathcal{C}_1$. Lemma 23 implies that there exists a sequence $\{\mathcal{G}_i\}_{i=1}^{\infty}$ of finite sets of measurable functions with $\mathcal{G}_1 \subseteq \mathcal{G}_2 \subseteq \cdots$ such that, for every measurable function $f^\star : \mathcal{X} \to \mathcal{Y}$, $\lim\limits_{i\to\infty} \min\limits_{g_i \in \mathcal{G}_i} \mathbb{E}[\hat{\mu}_{\mathbb{X}}(\ell(g_i(\cdot), f^\star(\cdot)))] = 0$. Furthermore, applying Lemma 21 to the sequence of sets $\{\ell(f(\cdot), g(\cdot)) : f, g \in \mathcal{G}_i\}$, with $\gamma_i = 4^{1-i}\bar{\ell}$, we find that there exist (nonrandom) nondecreasing sequences $\{m_i\}_{i=1}^{\infty}$ and $\{i_n\}_{n=1}^{\infty}$ in $\mathbb{N}$ with $m_i \to \infty$ and $i_n \to \infty$ such that $\forall n \in \mathbb{N}$,

$$\mathbb{E}\left[\max_{f,g\in\mathcal{G}_{i_n}} \left| \hat{\mu}_{\mathbb{X}}(\ell(f(\cdot), g(\cdot))) - \max_{m_{i_n} \le m \le n} \frac{1}{m}\sum_{t=1}^{m} \ell(f(X_t), g(X_t)) \right|\right] \le \gamma_{i_n}. \tag{21}$$

Let $I = \{i_n : n \in \mathbb{N}\}$, and for each $i \in I$, denote $n_i = \min\{n \in \mathbb{N} : i_n = i\}$. Markov's inequality and (21) imply

$$\begin{aligned}
&\sum_{i\in I} \mathbb{P}\left(\max_{f,g\in\mathcal{G}_i} \left| \hat{\mu}_{\mathbb{X}}(\ell(f(\cdot), g(\cdot))) - \max_{m_i \le m \le n_i} \frac{1}{m}\sum_{t=1}^{m} \ell(f(X_t), g(X_t)) \right| > \sqrt{\gamma_i}\right) \\
&\le \sum_{i\in I} \frac{1}{\sqrt{\gamma_i}} \mathbb{E}\left[\max_{f,g\in\mathcal{G}_i} \left| \hat{\mu}_{\mathbb{X}}(\ell(f(\cdot), g(\cdot))) - \max_{m_i \le m \le n_i} \frac{1}{m}\sum_{t=1}^{m} \ell(f(X_t), g(X_t)) \right|\right] \\
&\le \sum_{i\in I} \sqrt{\gamma_i} \le \sum_{i=1}^{\infty} 2^{1-i}\sqrt{\bar{\ell}} = 2\sqrt{\bar{\ell}} < \infty.
\end{aligned}$$

Therefore, the Borel-Cantelli Lemma implies that there exists an event K' of probability one, on which $\exists \iota_0 \in \mathbb{N}$ such that $\forall i \in I$ with $i \ge \iota_0$,

$$\max_{f,g\in\mathcal{G}_i} \left(\hat{\mu}_{\mathbb{X}}(\ell(f(\cdot), g(\cdot))) - \max_{m_i \le m \le n_i} \frac{1}{m}\sum_{t=1}^{m} \ell(f(X_t), g(X_t)) \right) \le \sqrt{\gamma_i}. \tag{22}$$

Additionally, note that $\forall n \in \mathbb{N}$, $n \ge n_{i_n}$, so that $\forall f, g \in \mathcal{G}_{i_n}$,

$$\max_{m_{i_n} \le m \le n} \frac{1}{m}\sum_{t=1}^{m} \ell(f(X_t), g(X_t)) \ge \max_{m_{i_n} \le m \le n_{i_n}} \frac{1}{m}\sum_{t=1}^{m} \ell(f(X_t), g(X_t)). \tag{23}$$

Furthermore, since $i_n \to \infty$, on the event K', $\exists \nu_1 \in \mathbb{N}$ such that $\forall n \ge \nu_1$, we have $i_n \ge \iota_0$, so that (22) and (23) imply

$$\max_{f,g\in\mathcal{G}_{i_n}} \left(\hat{\mu}_{\mathbb{X}}(\ell(f(\cdot), g(\cdot))) - \max_{m_{i_n} \le m \le n} \frac{1}{m}\sum_{t=1}^{m} \ell(f(X_t), g(X_t)) \right) \le \sqrt{\gamma_{i_n}}. \tag{24}$$

Now consider using the inductive learning rule $\hat{f}_n$ defined in (10), with $\mathcal{F}_n = \mathcal{G}_{i_n}$ and $\hat{m}_n = m_{i_n}$ for each $n \in \mathbb{N}$, and $\{\varepsilon_n\}_{n=1}^{\infty}$ an arbitrary sequence in $[0, \infty)$ with $\varepsilon_n \to 0$. Fix any measurable function $f^\star : \mathcal{X} \to \mathcal{Y}$. Note that, since $\lim\limits_{n\to\infty} i_n = \lim\limits_{i\to\infty} m_i = \infty$, we have $\lim\limits_{n\to\infty} \hat{m}_n = \infty$. Furthermore, since i_n is nondecreasing with $\lim\limits_{n\to\infty} i_n = \infty$, the above guarantees for the sequence $\{\mathcal{G}_i\}_{i=1}^{\infty}$ imply $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots$ and $\lim\limits_{n\to\infty} \min\limits_{g\in\mathcal{F}_n} \mathbb{E}[\hat{\mu}_{\mathbb{X}}(\ell(g(\cdot), f^\star(\cdot)))] = 0$. Therefore, Lemma 24 implies that there exists a (nonrandom) sequence $\{f_n^\star\}_{n=1}^{\infty}$ with

$f_n^\star \in \mathcal{F}_n$ for each $n \in \mathbb{N}$, a (nonrandom) sequence $\{\alpha_n\}_{n=1}^{\infty}$ in $(0,\infty)$ with $\alpha_n \to 0$, and an event K of probability one, on which $\exists \nu_0 \in \mathbb{N}$ such that $\forall n \geq \nu_0$,

$$\sup_{\hat{m}_n \leq m < \infty} \frac{1}{m} \sum_{t=1}^{m} \ell(f_n^\star(X_t), f^\star(X_t)) \leq \alpha_n. \tag{25}$$

For brevity, denote $\hat{g}_n(\cdot) = \hat{f}_n(X_{1:n}, f^\star(X_{1:n}), \cdot)$ for every $n \in \mathbb{N}$. Note that, by the definition of $\hat{f}_n$ from (10) and the fact that $f_n^\star \in \mathcal{F}_n$, $\forall n \in \mathbb{N}$,

$$\max_{\hat{m}_n \leq m \leq n} \frac{1}{m} \sum_{t=1}^{m} \ell(\hat{g}_n(X_t), f^\star(X_t)) \leq \varepsilon_n + \max_{\hat{m}_n \leq m \leq n} \frac{1}{m} \sum_{t=1}^{m} \ell(f_n^\star(X_t), f^\star(X_t)).$$

Thus, on the event K, $\forall n \in \mathbb{N}$ with $n \geq \nu_0$, (25) implies

$$\max_{\hat{m}_n \leq m \leq n} \frac{1}{m} \sum_{t=1}^{m} \ell(\hat{g}_n(X_t), f^\star(X_t)) \leq \varepsilon_n + \alpha_n. \tag{26}$$

On the other hand, suppose the event $K \cap K'$ occurs, fix any $n \in \mathbb{N}$ with $n \geq \max\{\nu_0, \nu_1\}$, and fix any $f \in \mathcal{F}_n$ satisfying

$$\hat{\mu}_{\mathbb{X}}(\ell(f(\cdot), f^\star(\cdot))) > 3\alpha_n + \sqrt{\gamma_{i_n}} + \varepsilon_n, \tag{27}$$

if such a function f exists in $\mathcal{F}_n$. The triangle inequality implies

$$\begin{aligned}
&\max_{\hat{m}_n \leq m \leq n} \frac{1}{m} \sum_{t=1}^{m} \ell(f(X_t), f^\star(X_t)) \\
&\geq \max_{\hat{m}_n \leq m \leq n} \frac{1}{m} \sum_{t=1}^{m} \left(\ell(f(X_t), f_n^\star(X_t)) - \ell(f_n^\star(X_t), f^\star(X_t))\right) \\
&\geq \max_{\hat{m}_n \leq m \leq n} \frac{1}{m} \sum_{t=1}^{m} \ell(f(X_t), f_n^\star(X_t)) \; - \sup_{\hat{m}_n \leq m < \infty} \frac{1}{m} \sum_{t=1}^{m} \ell(f_n^\star(X_t), f^\star(X_t)).
\end{aligned}$$

Since the event K holds and $n \geq \nu_0$, (25) implies this last expression is no smaller than

$$\max_{\hat{m}_n \leq m \leq n} \frac{1}{m} \sum_{t=1}^{m} \ell(f(X_t), f_n^\star(X_t)) - \alpha_n.$$

Furthermore, since both f and $f_n^\star$ are elements of $\mathcal{F}_n$, and since the event K' holds and $n \geq \nu_1$, the inequality (24) (together with the definitions of $\mathcal{F}_n = \mathcal{G}_{i_n}$ and $\hat{m}_n = m_{i_n}$) implies the above is at least as large as

$$\hat{\mu}_{\mathbb{X}}(\ell(f(\cdot), f_n^\star(\cdot))) - \sqrt{\gamma_{i_n}} - \alpha_n. \tag{28}$$

By the triangle inequality, $\ell(f(\cdot), f_n^\star(\cdot)) + \ell(f_n^\star(\cdot), f^\star(\cdot)) \geq \ell(f(\cdot), f^\star(\cdot))$. Combined with subadditivity and monotonicity of $\hat{\mu}_{\mathbb{X}}$ (Lemma 8), this implies

$$\hat{\mu}_{\mathbb{X}}(\ell(f(\cdot), f_n^\star(\cdot))) + \hat{\mu}_{\mathbb{X}}(\ell(f_n^\star(\cdot), f^\star(\cdot))) \geq \hat{\mu}_{\mathbb{X}}(\ell(f(\cdot), f_n^\star(\cdot)) + \ell(f_n^\star(\cdot), f^\star(\cdot))) \geq \hat{\mu}_{\mathbb{X}}(\ell(f(\cdot), f^\star(\cdot))).$$

Subtracting $\hat{\mu}_{\mathbb{X}}(\ell(f_n^\star(\cdot), f^\star(\cdot)))$ from these expressions implies that (28) is no smaller than

$$\begin{aligned}&\hat{\mu}_{\mathbb{X}}(\ell(f(\cdot), f^\star(\cdot))) - \hat{\mu}_{\mathbb{X}}(\ell(f_n^\star(\cdot), f^\star(\cdot))) - \sqrt{\gamma_{i_n}} - \alpha_n \\ &\geq \hat{\mu}_{\mathbb{X}}(\ell(f(\cdot), f^\star(\cdot))) \; - \sup_{\hat{m}_n \leq m < \infty} \frac{1}{m} \sum_{t=1}^{m} \ell(f_n^\star(X_t), f^\star(X_t)) - \sqrt{\gamma_{i_n}} - \alpha_n.\end{aligned}$$

Since the event K holds and $n \geq \nu_0$, (25) implies this is at least as large as

$$\hat{\mu}_{\mathbb{X}}(\ell(f(\cdot), f^\star(\cdot))) - \sqrt{\gamma_{i_n}} - 2\alpha_n.$$

Since f was chosen to satisfy (27), this is strictly greater than $3\alpha_n + \sqrt{\gamma_{i_n}} + \varepsilon_n - \sqrt{\gamma_{i_n}} - 2\alpha_n = \varepsilon_n + \alpha_n$. Altogether, we have that

$$\max_{\hat{m}_n \leq m \leq n} \frac{1}{m} \sum_{t=1}^{m} \ell(f(X_t), f^\star(X_t)) > \varepsilon_n + \alpha_n.$$

Together with (26), this implies that $\hat{g}_n \neq f$. Since this is true of any $f \in \mathcal{F}_n$ satisfying (27) (if any such f exists in $\mathcal{F}_n$), and since $\hat{g}_n \in \mathcal{F}_n$, it follows that $\hat{g}_n$ is a function $f \in \mathcal{F}_n$ *not* satisfying (27). Altogether, we have that on the event $K \cap K'$, $\forall n \in \mathbb{N}$ with $n \geq \max\{\nu_0, \nu_1\}$,

$$\hat{\mu}_{\mathbb{X}}(\ell(\hat{g}_n(\cdot), f^\star(\cdot))) \leq 3\alpha_n + \sqrt{\gamma_{i_n}} + \varepsilon_n.$$

In particular, recall that $i_n \to \infty$ and $\gamma_i \to 0$, so that $\lim\limits_{n\to\infty} \sqrt{\gamma_{i_n}} = 0$. Thus, since $\lim\limits_{n\to\infty} \alpha_n = \lim\limits_{n\to\infty} \varepsilon_n = 0$ (by their definitions), and since $\max\{\nu_0, \nu_1\} < \infty$, we have that on the event $K \cap K'$,

$$\lim_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(\hat{f}_n, f^\star; n) = \lim_{n\to\infty} \hat{\mu}_{\mathbb{X}}(\ell(\hat{g}_n(\cdot), f^\star(\cdot))) \leq \lim_{n\to\infty} 3\alpha_n + \sqrt{\gamma_{i_n}} + \varepsilon_n = 0.$$

Since the event $K \cap K'$ has probability one (by the union bound), and $\hat{\mathcal{L}}_{\mathbb{X}}$ is nonnegative, this establishes that $\hat{\mathcal{L}}_{\mathbb{X}}(\hat{f}_n, f^\star; n) \to 0$ (a.s.). Since this argument applies to *any* measurable $f^\star : \mathcal{X} \to \mathcal{Y}$, this establishes that $\hat{f}_n$ is strongly universally consistent under $\mathbb{X}$, so that $\mathbb{X} \in$ SUIL. Since this argument applies to *any* $\mathbb{X} \in \mathcal{C}_1$, this completes the proof that $\mathcal{C}_1 \subseteq$ SUIL. ■

Combining Lemmas 18, 19, and 25 completes the proof of Theorem 7.

Interestingly, we may note that the *only* reliance of the above proof of Lemma 25 on the assumption $\mathbb{X} \in \mathcal{C}_1$ is in the existence of the sequence $\{\mathcal{G}_i\}_{i=1}^{\infty}$ via Lemma 23: that is, we have in fact established that any $\mathbb{X}$ for which there exists a sequence $\{\mathcal{G}_i\}_{i=1}^{\infty}$ with these properties admits strong universal inductive learning, so that the existence of such a sequence implies $\mathbb{X} \in$ SUIL. Together with Theorem 7 (implying $\mathcal{C}_1 =$ SUIL) and Lemma 23 (implying $\mathbb{X} \in \mathcal{C}_1$ suffices for such a sequence $\{\mathcal{G}_i\}_{i=1}^{\infty}$ to exist), this establishes that $\mathcal{C}_1$ is in fact *equivalent* to the set of processes for which such a sequence exists (and hence so are SUIL and SUAL, via Theorem 7). Thus, we have yet another useful equivalent way of expressing Condition 1. This is stated formally in the following corollary.

Corollary 26 *A process* $\mathbb{X}$ *satisfies Condition 1 if and only if there exists a sequence* $\{\mathcal{F}_i\}_{i=1}^{\infty}$ *of nonempty finite sets of measurable functions* $\mathcal{X} \to \mathcal{Y}$ *such that, for every measurable* $f : \mathcal{X} \to \mathcal{Y}$, $\lim_{i\to\infty} \min_{f_i \in \mathcal{F}_i} \mathbb{E}[\hat{\mu}_{\mathbb{X}}(\ell(f_i(\cdot), f(\cdot)))] = 0$.

Indeed, we may further observe that, since Condition 1 does not involve $\mathcal{Y}$ or ℓ, applying the above equivalence to the special case of $\mathcal{Y} = \{0, 1\}$ and $\ell(y, y') = \mathbb{1}[y \neq y']$ admits another simple equivalent condition. Specifically, in that special case, one can easily verify that there exists a sequence $\{\mathcal{F}_i\}_{i=1}^{\infty}$ as described in Corollary 26 if and only if there exists a countable set $\mathcal{T}_2 \subseteq \mathcal{B}$ with $\forall A \in \mathcal{B}$, $\inf_{G\in\mathcal{T}_2} \mathbb{E}[\hat{\mu}_{\mathbb{X}}(G \triangle A)] = 0$, as guaranteed by the set $\mathcal{T}_1$ from Lemma 22. Thus, for any process $\mathbb{X}$, the existence of such a set $\mathcal{T}_2$ is also provably equivalent to Condition 1.

5. Optimistically Universal Learning

This section presents the proofs of two results on optimistically universal learning: Theorems 5 and 6 stated in Section 1.2. For the first of these, we propose a new general self-adaptive learning rule, and prove that it is optimistically universal: that is, it is strongly universally consistent under *every* process admitting strong universal self-adaptive learning. For the second of these theorems, we prove that there is no optimistically universal inductive learning rule. Together, these results imply that the additional capability of self-adaptive learning rules to adjust their predictor based on the unlabeled test data is crucial for optimistically universal learning.

5.1 Existence of Optimistically Universal Self-Adaptive Learning Rules

We now present the construction of an optimistically universal self-adaptive learning rule. Fix a sequence $\{\mathcal{F}_i\}_{i=1}^{\infty}$ of nonempty finite sets of measurable functions $\mathcal{X} \to \mathcal{Y}$ with $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots$ such that $\forall \mathbb{X} \in \mathcal{C}_1$, for every measurable $f : \mathcal{X} \to \mathcal{Y}$, $\lim_{i\to\infty} \min_{f_i\in\mathcal{F}_i} \mathbb{E}[\hat{\mu}_{\mathbb{X}}(\ell(f_i(\cdot), f(\cdot)))] = 0$. Recall that such a sequence $\{\mathcal{F}_i\}_{i=1}^{\infty}$ is guaranteed to exist by Lemma 23. Let $\{u_i\}_{i=1}^{\infty}$ be an arbitrary nondecreasing sequence in $\mathbb{N}$ with $u_i \to \infty$ and $u_1 = 1$, and let $\{\gamma_i\}_{i=1}^{\infty}$ be an arbitrary sequence in $(0, \infty)$ with $\gamma_1 \geq \bar{\ell}$ and $\gamma_i \to 0$. Let $\{x_i\}_{i=1}^{\infty}$ be any sequence in $\mathcal{X}$ and let $\{y_i\}_{i=1}^{\infty}$ be any sequence in $\mathcal{Y}$. For each $n, m \in \mathbb{N}$ with $m \geq n$, let

$$\hat{i}_{n,m}(x_{1:m}) = \max\Bigg\{ i \in \mathbb{N} : u_i \leq n \text{ and } \max_{f,g\in\mathcal{F}_i} \left(\max_{u_i \leq s \leq m} \frac{1}{s} \sum_{t=1}^{s} \ell(f(x_t), g(x_t)) - \max_{u_i \leq s \leq n} \frac{1}{s} \sum_{t=1}^{s} \ell(f(x_t), g(x_t)) \right) \leq \gamma_i \Bigg\}.$$

This is a well-defined positive integer, since our constraints on u_1 and γ_1 guarantee that the set of i values on the right hand side is nonempty, while the fact that $u_i \to \infty$ implies this set of i values is finite (and hence has a maximum element). Let $\{\varepsilon_n\}_{n=1}^{\infty}$ be an arbitrary sequence in $[0, \infty)$ such that $\varepsilon_n \to 0$. Finally, for every $n, m \in \mathbb{N}$ with $m \geq n$, define the function $\hat{f}_{n,m}(x_{1:m}, y_{1:n}, \cdot)$ as

$$\underset{f \in \mathcal{F}_{\hat{i}_{n,m}(x_{1:m})}}{\operatorname{argmin}^{\varepsilon_n}} \max_{u_{\hat{i}_{n,m}(x_{1:m})} \leq s \leq n} \frac{1}{s} \sum_{t=1}^{s} \ell(f(x_t), y_t). \tag{29}$$

Since the sets $\mathcal{F}_i$ are finite, one can easily verify that the selection in the $\operatorname{argmin}^{\varepsilon_n}$ above can be chosen in a way that makes $\hat{f}_{n,m}$ a measurable function. For completeness, for every $m \in \mathbb{N} \cup \{0\}$, also define $\hat{f}_{0,m}(x_{1:m}, \{\}, \cdot)$ as an arbitrary element of $\mathcal{F}_1$ (chosen identically for every m and $x_{1:m}$), which is then also a measurable function. Thus, the function $\hat{f}_{n,m}$ defines a self-adaptive learning rule. We have the following theorem for this $\hat{f}_{n,m}$.

Theorem 27 *The self-adaptive learning rule $\hat{f}_{n,m}$ is optimistically universal.*

Proof The proof proceeds along similar lines to that of Lemma 25, except using the data-dependent values $\hat{i}_{n,m}(X_{1:m})$ in place of the distribution-dependent sequence i_n from the proof of Lemma 25. Fix any $\mathbb{X} \in \mathcal{C}_1$ and any measurable $f^\star : \mathcal{X} \to \mathcal{Y}$.

Note that, for any given $i \in \mathbb{N}$ and $f, g \in \mathcal{F}_i$, $\max_{u_i \leq s \leq m} \frac{1}{s} \sum_{t=1}^{s} \ell(f(X_t), g(X_t))$ is nondecreasing in m, so that $\forall n \in \mathbb{N}$, $\hat{i}_{n,m}(X_{1:m})$ is nonincreasing in m. Since $\hat{i}_{n,m}(X_{1:m})$ is always positive, this implies $\hat{i}_{n,m}(X_{1:m})$ converges as $m \to \infty$; in particular, since $\hat{i}_{n,m}(X_{1:m}) \in \mathbb{N}$, this implies $\forall n \in \mathbb{N}$, $\exists m_n^* \in \mathbb{N}$ with $m_n^* \geq n$ such that $\forall m \geq m_n^*$, $\hat{i}_{n,m}(X_{1:m}) = \hat{i}_{n,m_n^*}(X_{1:m_n^*})$. For brevity, let us define $\hat{i}_n = \hat{i}_{n,m_n^*}(X_{1:m_n^*})$. By definition of $\hat{i}_{n,m}(X_{1:m})$, we have that every $m \geq m_n^*$ satisfies

$$\max_{f,g \in \mathcal{F}_{\hat{i}_n}} \left(\max_{u_{\hat{i}_n} \leq s \leq m} \frac{1}{s} \sum_{t=1}^{s} \ell(f(X_t), g(X_t)) - \max_{u_{\hat{i}_n} \leq s \leq n} \frac{1}{s} \sum_{t=1}^{s} \ell(f(X_t), g(X_t)) \right) \leq \gamma_{\hat{i}_n}.$$

Taking the limiting case as $m \to \infty$, together with monotonicity of the max function, this implies

$$\max_{f,g \in \mathcal{F}_{\hat{i}_n}} \left(\sup_{u_{\hat{i}_n} \leq s < \infty} \frac{1}{s} \sum_{t=1}^{s} \ell(f(X_t), g(X_t)) - \max_{u_{\hat{i}_n} \leq s \leq n} \frac{1}{s} \sum_{t=1}^{s} \ell(f(X_t), g(X_t)) \right) \leq \gamma_{\hat{i}_n}. \tag{30}$$

Furthermore, for each $i \in \mathbb{N}$, since $\mathcal{F}_i$ is finite, continuity of the max function implies

$$\begin{aligned}
&\limsup_{n \to \infty} \max_{f,g \in \mathcal{F}_i} \left(\max_{u_i \leq s \leq m_n^*} \frac{1}{s} \sum_{t=1}^{s} \ell(f(X_t), g(X_t)) - \max_{u_i \leq s \leq n} \frac{1}{s} \sum_{t=1}^{s} \ell(f(X_t), g(X_t)) \right) \\
&\leq \limsup_{n \to \infty} \max_{f,g \in \mathcal{F}_i} \left(\max_{u_i \leq s < \infty} \frac{1}{s} \sum_{t=1}^{s} \ell(f(X_t), g(X_t)) - \max_{u_i \leq s \leq n} \frac{1}{s} \sum_{t=1}^{s} \ell(f(X_t), g(X_t)) \right) \\
&= \max_{f,g \in \mathcal{F}_i} \left(\max_{u_i \leq s < \infty} \frac{1}{s} \sum_{t=1}^{s} \ell(f(X_t), g(X_t)) - \lim_{n \to \infty} \max_{u_i \leq s \leq n} \frac{1}{s} \sum_{t=1}^{s} \ell(f(X_t), g(X_t)) \right) = 0 < \gamma_i.
\end{aligned}$$

Together with finiteness of every u_i, this implies

$$\lim_{n \to \infty} \hat{i}_n = \infty. \tag{31}$$

Next note that, by our choices of the sequences $\{\mathcal{F}_i\}_{i=1}^{\infty}$ and $\{u_i\}_{i=1}^{\infty}$, Lemma 24 implies that there exists a (nonrandom) sequence $\{f_i^\star\}_{i=1}^{\infty}$, with $f_i^\star \in \mathcal{F}_i$ for each $i \in \mathbb{N}$, a (nonrandom) sequence $\{\alpha_i\}_{i=1}^{\infty}$ in $(0,\infty)$ with $\alpha_i \to 0$, and an event K of probability one, on which $\exists \iota_0 \in \mathbb{N}$ such that $\forall i \geq \iota_0$,

$$\sup_{u_i \leq s < \infty} \frac{1}{s} \sum_{t=1}^{s} \ell(f_i^\star(X_t), f^\star(X_t)) \leq \alpha_i.$$

In particular, since $\lim\limits_{n\to\infty} \hat{i}_n = \infty$ by (31), this implies that, on the event K, $\exists \nu_0 \in \mathbb{N}$ such that $\forall n \geq \nu_0$, we have $\hat{i}_n \geq \iota_0$, so that the above implies

$$\sup_{u_{\hat{i}_n} \leq s < \infty} \frac{1}{s} \sum_{t=1}^{s} \ell(f_{\hat{i}_n}^\star(X_t), f^\star(X_t)) \leq \alpha_{\hat{i}_n}. \tag{32}$$

For brevity, for every $n, m \in \mathbb{N}$ with $m \geq n$, define $\hat{g}_{n,m}(\cdot) = \hat{f}_{n,m}(X_{1:m}, f^\star(X_{1:n}), \cdot)$. Since every $m \geq m_n^*$ has $\hat{i}_{n,m}(X_{1:m}) = \hat{i}_{n,m_n^*}(X_{1:m_n^*})$, the definition of $\hat{f}_{n,m}$ implies that any $m \geq m_n^*$ also has $\hat{g}_{n,m} = \hat{g}_{n,m_n^*}$ (recalling the remark following the definition of $\mathrm{argmin}^{\varepsilon}$, regarding consistency among multiple evaluations). Denote $\hat{g}_n = \hat{g}_{n,m_n^*}$. Combining the definition of $\hat{f}_{n,m_n^*}$ with (32) we have that, on the event K, $\forall n \in \mathbb{N}$ with $n \geq \nu_0$,

$$\begin{aligned}\max_{u_{\hat{i}_n} \leq s \leq n} \frac{1}{s} \sum_{t=1}^{s} \ell(\hat{g}_n(X_t), f^\star(X_t)) &\leq \max_{u_{\hat{i}_n} \leq s \leq n} \frac{1}{s} \sum_{t=1}^{s} \ell(f_{\hat{i}_n}^\star(X_t), f^\star(X_t)) + \varepsilon_n \\ &\leq \sup_{u_{\hat{i}_n} \leq s < \infty} \frac{1}{s} \sum_{t=1}^{s} \ell(f_{\hat{i}_n}^\star(X_t), f^\star(X_t)) + \varepsilon_n \leq \alpha_{\hat{i}_n} + \varepsilon_n. \end{aligned} \tag{33}$$

Now suppose the event K occurs, fix any $n \in \mathbb{N}$ with $n \geq \nu_0$, and fix any $f \in \mathcal{F}_{\hat{i}_n}$ satisfying

$$\sup_{u_{\hat{i}_n} \leq s < \infty} \frac{1}{s} \sum_{t=1}^{s} \ell(f(X_t), f^\star(X_t)) > 3\alpha_{\hat{i}_n} + \gamma_{\hat{i}_n} + \varepsilon_n, \tag{34}$$

if such a function f exists in $\mathcal{F}_{\hat{i}_n}$. The triangle inequality implies

$$\begin{aligned}&\max_{u_{\hat{i}_n} \leq s \leq n} \frac{1}{s} \sum_{t=1}^{s} \ell(f(X_t), f^\star(X_t)) \\ &\geq \max_{u_{\hat{i}_n} \leq s \leq n} \frac{1}{s} \sum_{t=1}^{s} \Big(\ell(f(X_t), f_{\hat{i}_n}^\star(X_t)) - \ell(f_{\hat{i}_n}^\star(X_t), f^\star(X_t)) \Big) \\ &\geq \max_{u_{\hat{i}_n} \leq s \leq n} \frac{1}{s} \sum_{t=1}^{s} \ell(f(X_t), f_{\hat{i}_n}^\star(X_t)) \; - \sup_{u_{\hat{i}_n} \leq s < \infty} \frac{1}{s} \sum_{t=1}^{s} \ell(f_{\hat{i}_n}^\star(X_t), f^\star(X_t)).\end{aligned}$$

Since the event K holds and $n \geq \nu_0$, (32) implies the expression on this last line is at least as large as

$$\max_{u_{\hat{i}_n} \leq s \leq n} \frac{1}{s} \sum_{t=1}^{s} \ell(f(X_t), f_{\hat{i}_n}^\star(X_t)) - \alpha_{\hat{i}_n}.$$

Since both f and $f^\star_{\hat{i}_n}$ are elements of $\mathcal{F}_{\hat{i}_n}$, (30) implies that the above expression is at least as large as

$$\sup_{u_{\hat{i}_n} \le s < \infty} \frac{1}{s} \sum_{t=1}^{s} \ell(f(X_t), f^\star_{\hat{i}_n}(X_t)) - \gamma_{\hat{i}_n} - \alpha_{\hat{i}_n}. \tag{35}$$

By the triangle inequality, $\ell(f(X_t), f^\star_{\hat{i}_n}(X_t)) \ge \ell(f(X_t), f^\star(X_t)) - \ell(f^\star_{\hat{i}_n}(X_t), f^\star(X_t))$ for every t. Together with monotonicity of the supremum function, this implies

$$\begin{aligned}
&\sup_{u_{\hat{i}_n} \le s < \infty} \frac{1}{s} \sum_{t=1}^{s} \ell(f(X_t), f^\star_{\hat{i}_n}(X_t)) \\
&\ge \sup_{u_{\hat{i}_n} \le s < \infty} \frac{1}{s} \sum_{t=1}^{s} \Big(\ell(f(X_t), f^\star(X_t)) - \ell(f^\star_{\hat{i}_n}(X_t), f^\star(X_t)) \Big) \\
&\ge \sup_{u_{\hat{i}_n} \le s < \infty} \frac{1}{s} \sum_{t=1}^{s} \ell(f(X_t), f^\star(X_t)) \; - \sup_{u_{\hat{i}_n} \le s < \infty} \frac{1}{s} \sum_{t=1}^{s} \ell(f^\star_{\hat{i}_n}(X_t), f^\star(X_t)).
\end{aligned} \tag{36}$$

Since the event K holds and $n \ge \nu_0$, (32) implies the second term in (36) is at most $\alpha_{\hat{i}_n}$, while (34) implies the first term is strictly greater than $3\alpha_{\hat{i}_n} + \gamma_{\hat{i}_n} + \varepsilon_n$. Together, we have that (36) is strictly greater than $2\alpha_{\hat{i}_n} + \gamma_{\hat{i}_n} + \varepsilon_n$. Thus, (35) is strictly greater than $\alpha_{\hat{i}_n} + \varepsilon_n$.

Altogether, on the event K, for any $n \in \mathbb{N}$ with $n \ge \nu_0$, any $f \in \mathcal{F}_{\hat{i}_n}$ satisfying (34) has

$$\max_{u_{\hat{i}_n} \le s \le n} \frac{1}{s} \sum_{t=1}^{s} \ell(f(X_t), f^\star(X_t)) > \alpha_{\hat{i}_n} + \varepsilon_n,$$

which together with (33) implies $\hat{g}_n \ne f$. Since this is true of any $f \in \mathcal{F}_{\hat{i}_n}$ satisfying (34) (if any such f exists in $\mathcal{F}_{\hat{i}_n}$), and since $\hat{g}_n \in \mathcal{F}_{\hat{i}_n}$, it must be that $\hat{g}_n$ is a function in $\mathcal{F}_{\hat{i}_n}$ *not* satisfying (34). In summary, we have established that, on the event K, every $n \in \mathbb{N}$ with $n \ge \nu_0$ has

$$\sup_{u_{\hat{i}_n} \le s < \infty} \frac{1}{s} \sum_{t=1}^{s} \ell(\hat{g}_n(X_t), f^\star(X_t)) \le 3\alpha_{\hat{i}_n} + \gamma_{\hat{i}_n} + \varepsilon_n. \tag{37}$$

Now note that, for every $n \in \mathbb{N}$, since $\hat{g}_{n,m} = \hat{g}_n$ for every $m \ge m_n^*$, we have

$$\begin{aligned}
\hat{\mathcal{L}}_{\mathbb{X}}\Big(\hat{f}_{n,\cdot}, f^\star; n\Big) &= \limsup_{s\to\infty} \frac{1}{s+1} \sum_{m=n}^{n+s} \ell(\hat{g}_{n,m}(X_{m+1}), f^\star(X_{m+1})) \\
&\le \limsup_{s\to\infty} \frac{1}{s+1}(m_n^* - 1)\bar{\ell} + \frac{1}{s+1} \sum_{m=m_n^*}^{n+s} \ell(\hat{g}_n(X_{m+1}), f^\star(X_{m+1})) \\
&\le \limsup_{s\to\infty} \frac{n+s+1}{s+1} \frac{1}{n+s+1} \sum_{t=1}^{n+s+1} \ell(\hat{g}_n(X_t), f^\star(X_t)) \\
&= \limsup_{s\to\infty} \frac{1}{s} \sum_{t=1}^{s} \ell(\hat{g}_n(X_t), f^\star(X_t)) \le \sup_{u_{\hat{i}_n} \le s < \infty} \frac{1}{s} \sum_{t=1}^{s} \ell(\hat{g}_n(X_t), f^\star(X_t)).
\end{aligned}$$

Combined with (37), this implies that, on the event K, every $n \in \mathbb{N}$ with $n \ge \nu_0$ satisfies

$$\hat{\mathcal{L}}_{\mathbb{X}}\Big(\hat{f}_{n,\cdot}, f^\star; n\Big) \le 3\alpha_{\hat{i}_n} + \gamma_{\hat{i}_n} + \varepsilon_n.$$

Recalling that $\lim_{i\to\infty} \alpha_i = \lim_{i\to\infty} \gamma_i = \lim_{n\to\infty} \varepsilon_n = 0$ by their definitions, and that $\lim_{n\to\infty} \hat{i}_n = \infty$ by (31), we have that on the event K,

$$\lim_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}\left(\hat{f}_{n,\cdot}, f^\star; n\right) \leq \lim_{n\to\infty} 3\alpha_{\hat{i}_n} + \gamma_{\hat{i}_n} + \varepsilon_n = 0.$$

Recalling that the event K has probability one, and $\hat{\mathcal{L}}_{\mathbb{X}}$ is nonnegative, this establishes that $\hat{\mathcal{L}}_{\mathbb{X}}\left(\hat{f}_{n,\cdot}, f^\star; n\right) \to 0$ (a.s.). Since this argument applies to *any* measurable $f^\star : \mathcal{X} \to \mathcal{Y}$, this establishes that $\hat{f}_{n,m}$ is strongly universally consistent under $\mathbb{X}$. Furthermore, since this argument applies to *any* $\mathbb{X} \in \mathcal{C}_1$, and Theorem 7 implies SUAL $= \mathcal{C}_1$, this completes the proof that $\hat{f}_{n,m}$ is strongly universally consistent under every $\mathbb{X} \in$ SUAL: i.e., that $\hat{f}_{n,m}$ is optimistically universal. ■

An immediate consequence of Theorem 27 is that there *exist* optimistically universal self-adaptive learning rules, so that this also completes the proof of Theorem 5 stated in Section 1.2.

5.2 Nonexistence of Optimistically Universal Inductive Learning Rules

Given the positive result above on optimistically universal self-adaptive learning, it is natural to wonder whether the same is true of *inductive* learning. However, it turns out this is *not* the case. In fact, we find below that there do not even exist inductive learning rules that are strongly universally consistent under every $\mathbb{X}$ with *convergent relative frequencies*, which form a proper subset of SUIL (recall the discussion in Section 3). We begin with the following result (restated from Section 1.2). For technical reasons, throughout Section 5.2 we assume that $(\mathcal{X}, \mathcal{T})$ is a Polish space; for instance, $\mathbb{R}^p$ satisfies this for any $p \in \mathbb{N}$, under the usual Euclidean topology.

Theorem 6 (restated) *There* does not exist *an optimistically universal inductive learning rule, if $\mathcal{X}$ is uncountable.*

Before presenting the proof, we first have a technical lemma regarding a basic fact about nonatomic probability measures.

Lemma 28 *For any nonatomic probability measure π_0 on $\mathcal{X}$, there exists a sequence $\{R_k\}_{k=1}^{\infty}$ in $\mathcal{B}$ such that, $\forall k \in \mathbb{N}$, $\pi_0(R_k) = 1/2$, and $\forall A \in \mathcal{B}$, $\lim_{k\to\infty} \pi_0(A \cap R_k) = (1/2)\pi_0(A)$.*

Proof Denote by λ the Lebesgue measure on $\mathbb{R}$. First, note that since $(\mathcal{X}, \mathcal{T})$ is a Polish space, $(\mathcal{X}, \mathcal{B})$ is a *standard Borel space* (in the sense of Srivastava, 1998). In particular, since π_0 is nonatomic, this implies that there exists a Borel isomorphism $\psi : \mathcal{X} \to [0,1]$ such that, for every Borel subset B of $[0,1]$, $\pi_0(\psi^{-1}(B)) = \lambda(B)$ (see e.g., Srivastava, 1998, Theorem 3.4.23).

For each $k \in \mathbb{N}$ and each $i \in \mathbb{Z}$, define $C_{k,i} = \left[(i-1)2^{-k}, i2^{-k}\right)$, let $B_k = \bigcup_{i\in\mathbb{Z}} C_{k,2i}$, and define $R_k = \psi^{-1}(B_k \cap [0,1])$. Note that each $B_k \cap [0,1]$ is a Borel subset of $[0,1]$, so that measurability of ψ implies $R_k \in \mathcal{B}$; furthermore, $\pi_0(R_k) = \pi_0(\psi^{-1}(B_k \cap [0,1])) = \lambda(B_k \cap [0,1]) = 1/2$, as required.

Now fix any set $A \in \mathcal{B}$, and let $B \subseteq [0,1]$ be the Borel subset of $[0,1]$ with $A = \psi^{-1}(B)$ (which exists by the bimeasurability property of ψ). Since λ is a *regular* measure (e.g., Cohn, 1980, Proposition 1.4.1), for any $\varepsilon > 0$, there exists an *open* set U_ε with $B \subseteq U_\varepsilon \subseteq \mathbb{R}$ such that $\lambda(U_\varepsilon \setminus B) < \varepsilon$. As any open subset of $\mathbb{R}$ is a union of countably many pairwise-disjoint open intervals (e.g., Kolmogorov and Fomin, 1975, Section 6, Theorem 6), we let $(a_1, b_1), (a_2, b_2), \ldots$ be a sequence of disjoint open intervals ($a_i \in [-\infty, \infty)$, $b_i \in (-\infty, \infty]$) with $U_\varepsilon = \bigcup_{i=1}^{\infty} (a_i, b_i)$; for notational simplicity, we suppose this sequence is infinite, which can always be achieved by adding an infinite number of empty intervals (a_i, b_i) with $a_i = b_i \in \mathbb{R}$. Since $U_\varepsilon \setminus \bigcup_{i=1}^{j} (a_i, b_i) \downarrow \emptyset$ as $j \to \infty$, and since $\lambda(U_\varepsilon) = \lambda(U_\varepsilon \setminus B) + \lambda(B) < \varepsilon + 1 < \infty$, continuity of finite measures implies $\lim_{j\to\infty} \lambda\left(U_\varepsilon \setminus \bigcup_{i=1}^{j} (a_i, b_i)\right) = 0$ (e.g., Schervish, 1995, Theorem A.19). In particular, for any $\delta > 0$, $\exists j_\delta \in \mathbb{N}$ such that $\lambda\left(U_\varepsilon \setminus \bigcup_{i=1}^{j_\delta} (a_i, b_i)\right) < \delta/2$. Now let $k_\delta = \left\lceil \log_2\left(\frac{4j_\delta}{\delta}\right)\right\rceil$. Since $\lambda(U_\varepsilon) < \infty$, it must be that every $a_i > -\infty$ and every $b_i < \infty$. Furthermore, letting $\bar{a}_i = \min\{t2^{-k_\delta} : a_i < t2^{-k_\delta}, t \in \mathbb{Z}\}$ and $\bar{b}_i = \max\{t2^{-k_\delta} : b_i > t2^{-k_\delta}, t \in \mathbb{Z}\}$, we have that

$$\lambda\left((a_i, b_i) \setminus \bigcup \{C_{k_\delta,t} : C_{k_\delta,t} \subseteq (a_i, b_i), t \in \mathbb{Z}\}\right) \le |\bar{a}_i - a_i| + |b_i - \bar{b}_i| \le 2 \cdot 2^{-k_\delta} \le \frac{\delta}{2j_\delta}.$$

Thus,

$$\begin{aligned}
&\lambda\Big(U_\varepsilon \setminus \bigcup \{C_{k_\delta,t} : C_{k_\delta,t} \subseteq U_\varepsilon, t \in \mathbb{Z}\}\Big)\\
&\le \lambda\left(U_\varepsilon \setminus \bigcup_{i=1}^{j_\delta}(a_i, b_i)\right) + \lambda\left(\bigcup_{i=1}^{j_\delta}(a_i, b_i) \setminus \bigcup \{C_{k_\delta,t} : C_{k_\delta,t} \subseteq U_\varepsilon, t \in \mathbb{Z}\}\right)\\
&< \delta/2 + \sum_{i=1}^{j_\delta} \lambda\Big((a_i, b_i) \setminus \bigcup \{C_{k_\delta,t} : C_{k_\delta,t} \subseteq U_\varepsilon, t \in \mathbb{Z}\}\Big)\\
&\le \delta/2 + \sum_{i=1}^{j_\delta} \lambda\Big((a_i, b_i) \setminus \bigcup \{C_{k_\delta,t} : C_{k_\delta,t} \subseteq (a_i, b_i), t \in \mathbb{Z}\}\Big) \le \delta/2 + \sum_{i=1}^{j_\delta} \frac{\delta}{2j_\delta} = \delta. \qquad (38)
\end{aligned}$$

Now note that, for every $k > k_\delta$ and $i \in \mathbb{Z}$, each $j \in \mathbb{Z}$ has either $C_{k,j} \subseteq C_{k_\delta,i}$ or $C_{k,j} \cap C_{k_\delta,i} = \emptyset$, and moreover each j has $C_{k,2j} \subseteq C_{k_\delta,i}$ if and only if $C_{k,2j-1} \subseteq C_{k_\delta,i}$ (the smallest j with $C_{k,j} \subseteq C_{k_\delta,i}$ has $(j-1)2^{-k} = (i-1)2^{-k_\varepsilon}$, which implies j is an *odd* number because $k > k_\varepsilon$; similarly, the largest j with $C_{k,j} \subseteq C_{k_\delta,i}$ has $j2^{-k} = i2^{-k_\varepsilon}$ and is therefore *even*), so that

$$\lambda(B_k \cap C_{k_\delta,i}) = \lambda\Big(\bigcup\{C_{k,2j} : C_{k,2j} \subseteq C_{k_\delta,i}, j \in \mathbb{Z}\}\Big) = (1/2)\lambda(C_{k_\delta,i}),$$

and hence (by disjointness of the $C_{k_\delta,i}$ sets)

$$\lambda\Big(B_k \cap \bigcup\{C_{k_\delta,i} : C_{k_\delta,i} \subseteq U_\varepsilon, i \in \mathbb{Z}\}\Big) = \sum_{i\in\mathbb{Z}:C_{k_\delta,i}\subseteq U_\varepsilon} \lambda(B_k \cap C_{k_\delta,i})$$
$$= \sum_{i\in\mathbb{Z}:C_{k_\delta,i}\subseteq U_\varepsilon} (1/2)\lambda(C_{k_\delta,i}) = (1/2)\lambda\Big(\bigcup\{C_{k_\delta,i} : C_{k_\delta,i} \subseteq U_\varepsilon, i \in \mathbb{Z}\}\Big).$$

Therefore $\forall k > k_\delta$,

$$\begin{aligned}&\lambda(U_\varepsilon \cap B_k)\\ &= \lambda\Big(B_k \cap \bigcup\{C_{k_\delta,i} : C_{k_\delta,i} \subseteq U_\varepsilon, i \in \mathbb{Z}\}\Big) + \lambda\Big(B_k \cap U_\varepsilon \setminus \bigcup\{C_{k_\delta,i} : C_{k_\delta,i} \subseteq U_\varepsilon, i \in \mathbb{Z}\}\Big)\\ &= (1/2)\lambda\Big(\bigcup\{C_{k_\delta,i} : C_{k_\delta,i} \subseteq U_\varepsilon, i \in \mathbb{Z}\}\Big) + \lambda\Big(B_k \cap U_\varepsilon \setminus \bigcup\{C_{k_\delta,i} : C_{k_\delta,i} \subseteq U_\varepsilon, i \in \mathbb{Z}\}\Big).\end{aligned} \tag{39}$$

The first term in (39) equals $(1/2)\left(\lambda(U_\varepsilon) - \lambda(U_\varepsilon \setminus \bigcup\{C_{k_\delta,i} : C_{k_\delta,i} \subseteq U_\varepsilon, i \in \mathbb{Z}\})\right)$, which by (38) is greater than $(1/2)\lambda(U_\varepsilon) - \delta/2$. Furthermore, the second term in (39) is no smaller than 0, and no greater than $\lambda(U_\varepsilon \setminus \bigcup\{C_{k_\delta,i} : C_{k_\delta,i} \subseteq U_\varepsilon, i \in \mathbb{Z}\})$. Thus,

$$\begin{aligned}&(1/2)\lambda(U_\varepsilon) - \delta/2 < \lambda(U_\varepsilon \cap B_k)\\ &\le (1/2)\Big(\lambda(U_\varepsilon) - \lambda\Big(U_\varepsilon \setminus \bigcup\{C_{k_\delta,i} : C_{k_\delta,i} \subseteq U_\varepsilon, i \in \mathbb{Z}\}\Big)\Big) + \lambda\Big(U_\varepsilon \setminus \bigcup\{C_{k_\delta,i} : C_{k_\delta,i} \subseteq U_\varepsilon, i \in \mathbb{Z}\}\Big)\\ &= (1/2)\Big(\lambda(U_\varepsilon) + \lambda\Big(U_\varepsilon \setminus \bigcup\{C_{k_\delta,i} : C_{k_\delta,i} \subseteq U_\varepsilon, i \in \mathbb{Z}\}\Big)\Big) < (1/2)\lambda(U_\varepsilon) + \delta/2,\end{aligned}$$

where this last inequality is by (38).

Since this holds for every $k > k_\delta$, and k_δ is finite for every $\delta \in (0, 1)$, we have $\forall \delta \in (0, 1)$,

$$(1/2)\lambda(U_\varepsilon) - \delta/2 \le \liminf_{k\to\infty} \lambda(U_\varepsilon \cap B_k) \le \limsup_{k\to\infty} \lambda(U_\varepsilon \cap B_k) \le (1/2)\lambda(U_\varepsilon) + \delta/2,$$

and taking the limit as $\delta \to 0$ implies

$$\lim_{k\to\infty} \lambda(U_\varepsilon \cap B_k) = (1/2)\lambda(U_\varepsilon).$$

This further implies that

$$\limsup_{k\to\infty} \lambda(B \cap B_k) \le \lim_{k\to\infty} \lambda(U_\varepsilon \cap B_k) = (1/2)\lambda(U_\varepsilon) < (1/2)\lambda(B) + \varepsilon/2,$$

and

$$\begin{aligned}\liminf_{k\to\infty} \lambda(B \cap B_k) &\ge \lim_{k\to\infty} \lambda(U_\varepsilon \cap B_k) - \lambda(U_\varepsilon \setminus B) = (1/2)\lambda(U_\varepsilon) - \lambda(U_\varepsilon \setminus B)\\ &= (1/2)\lambda(B) - (1/2)\lambda(U_\varepsilon \setminus B) > (1/2)\lambda(B) - \varepsilon/2.\end{aligned}$$

Since these inequalities hold for every $\varepsilon > 0$, taking the limit as $\varepsilon \to 0$ reveals that

$$\lim_{k\to\infty} \lambda(B \cap B_k) = (1/2)\lambda(B).$$

Furthermore, since $\psi^{-1}(B) \cap \psi^{-1}(B_k \cap [0,1]) = \psi^{-1}(B \cap B_k \cap [0,1]) = \psi^{-1}(B \cap B_k)$ for every $k \in \mathbb{N}$, this implies that

$$\begin{aligned}\lim_{k\to\infty} \pi_0(A \cap R_k) &= \lim_{k\to\infty} \pi_0(\psi^{-1}(B) \cap \psi^{-1}(B_k \cap [0,1])) = \lim_{k\to\infty} \pi_0(\psi^{-1}(B \cap B_k)) \\ &= \lim_{k\to\infty} \lambda(B \cap B_k) = (1/2)\lambda(B) = (1/2)\pi_0(\psi^{-1}(B)) = (1/2)\pi_0(A).\end{aligned}$$

Since this argument holds $\forall A \in \mathcal{B}$, this completes the proof. ■

We are now ready for the proof of Theorem 6. The proof is partly inspired by that of a related (but somewhat different) result of Nobel (1999), based on a technique of Adams and Nobel (1998). Specifically, Nobel (1999) proves that there is no universally consistent learning rule for all *joint* processes $(\mathbb{X}, \mathbb{Y})$ that are stationary and ergodic. In contrast, we are interested in learning under a fixed target function $f^\star$, and as such the construction of Nobel (1999) needs to be modified for our purposes. However, the proof below does preserve the essential elements of the cutting and stacking argument of Adams and Nobel (1998), though generalized to suit our abstract setting. While the processes $\mathbb{X}$ we construct do not have the property of stationarity from the original proof of Nobel (1999), they *do* satisfy ergodicity (indeed, they are *product* processes) and are contained in CRF.

Proof of Theorem 6 Fix any inductive learning rule f_n. We begin by constructing the process $\mathbb{X}$. Since $\mathcal{X}$ is uncountable, and $(\mathcal{X}, \mathcal{T})$ is a Polish space, there exists a nonatomic probability measure π_0 on $\mathcal{X}$ (with respect to $\mathcal{B}$) (see Parthasarathy, 1967, Chapter 2, Theorem 8.1). Furthermore, fixing any such nonatomic π_0, Lemma 28 implies there exists a sequence $\{R_k\}_{k=1}^{\infty}$ in $\mathcal{B}$ such that, $\forall k \in \mathbb{N}$, $\pi_0(R_k) = 1/2$, and $\forall A \in \mathcal{B}$, $\lim_{k\to\infty} \pi_0(A \cap R_k) = (1/2)\pi_0(A)$. Also define $R_0 = \emptyset$. Define random variables $U_{k,j}$ (for all $k, j \in \mathbb{N}$), $V_{k,j}$ (for all $k, j \in \mathbb{N}$), and W_j (for all $j \in \mathbb{N}$), all mutually independent (and independent from $\{f_n\}_{n\in\mathbb{N}}$), with distributions specified as follows. For each $k, j \in \mathbb{N}$, $U_{k,j}$ has distribution $\pi_0(\cdot|\mathcal{X} \setminus R_k)$, while $V_{k,j}$ has distribution $\pi_0(\cdot|R_k)$. For each $j \in \mathbb{N}$, W_j has distribution π_0. Denote $\mathbf{U} = \{U_{k,j}\}_{k,j\in\mathbb{N}}$, $\mathbf{V} = \{V_{k,j}\}_{k,j\in\mathbb{N}}$, $\mathbf{W} = \{W_j\}_{j\in\mathbb{N}}$.

Fix any $y_0, y_1 \in \mathcal{Y}$ with $\ell(y_0, y_1) > 0$. For any array $\mathbf{v} = \{v_{k,j}\}_{k,j\in\mathbb{N}}$, and any $K \in \mathbb{N}$, denote $\mathbf{v}_{<K} = \{v_{k,j}\}_{k,j\in\mathbb{N},k<K}$, and denote $\mathbf{v}_K = \{v_{K,j}\}_{j\in\mathbb{N}}$. Then, for any arrays $\mathbf{u} = \{u_{k,j}\}_{k,j\in\mathbb{N}}$ and $\mathbf{v} = \{v_{k,j}\}_{k,j\in\mathbb{N}}$ in $\mathcal{X}$, any sequence $\mathbf{w} = \{w_j\}_{j\in\mathbb{N}}$ in $\mathcal{X}$, and any $K \in \mathbb{N}$, define

$$f_K^\star(x; \mathbf{u}_{<K}, \mathbf{v}_{<K}, \mathbf{w}) = \begin{cases} y_0, & \text{if } x \in (\mathbf{v}_{<K} \cup R_K) \setminus (\mathbf{w} \cup \mathbf{u}_{<K}) \\ y_1, & \text{otherwise} \end{cases}$$

and

$$f_0^\star(x; \mathbf{v}) = \begin{cases} y_0, & \text{if } x \in \mathbf{v} \\ y_1, & \text{otherwise} \end{cases},$$

where, for notational simplicity, in these definitions we treat $\mathbf{v}_{<K}$, $\mathbf{w}$, $\mathbf{u}_{<K}$, $\mathbf{v}$ as the *sets* of the distinct values in the respective arrays.

Now, for any $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$ as above, inductively define values $X_i^{(k)}(\mathbf{u}_{<k},\mathbf{u}_k,\mathbf{v}_{<k},\mathbf{v}_k,\mathbf{w})$ as follows. Let $n_0 = 0$. For this inductive definition, suppose that for some $k \in \mathbb{N}$ the value $n_{k-1} \in \mathbb{N}$ and the values $\{X_i^{(k-1)}(\mathbf{u}_{<k-1}, \mathbf{u}_{k-1}, \mathbf{v}_{<k-1}, \mathbf{v}_{k-1}, \mathbf{w}) : i \in \mathbb{N}, i \leq n_{k-1}\}$

are already defined (taking this to be trivially satisfied in the case $k = 1$, wherein this is an empty sequence). For each $i \in \mathbb{N}$ with $i \leq n_{k-1}$, define $\tilde{X}_i^{(k)}(\mathbf{u}_{<k}, \mathbf{u}_k, \mathbf{v}_{<k}, \mathbf{v}_k, \mathbf{w})$ and $X_i^{(k)}(\mathbf{u}_{<k}, \mathbf{u}_k, \mathbf{v}_{<k}, \mathbf{v}_k, \mathbf{w})$ both equal to $X_i^{(k-1)}(\mathbf{u}_{<k-1}, \mathbf{u}_k, \mathbf{v}_{<k-1}, \mathbf{v}_{k-1}, \mathbf{w})$. Then, for each $i \in \mathbb{N}$, define $\tilde{X}^{(k)}_{n_{k-1}+k(i-1)+1}(\mathbf{u}_{<k}, \mathbf{u}_k, \mathbf{v}_{<k}, \mathbf{v}_k, \mathbf{w}) = v_{k,n_{k-1}+k(i-1)+1}$, and for each $j \in \mathbb{N}$ with $2 \leq j \leq k$, define $\tilde{X}^{(k)}_{n_{k-1}+k(i-1)+j}(\mathbf{u}_{<k}, \mathbf{u}_k, \mathbf{v}_{<k}, \mathbf{v}_k, \mathbf{w}) = u_{k,n_{k-1}+k(i-1)+j}$. To simplify notation, for each $i \in \mathbb{N}$, abbreviate $\hat{X}_i^{(k)} = \tilde{X}_i^{(k)}(\mathbf{U}_{<k}, \mathbf{U}_k, \mathbf{V}_{<k}, \mathbf{V}_k, \mathbf{W})$. If $\exists n \in \mathbb{N}$ with $n > n_{k-1}$ such that

$$\mathbb{P}\left(\pi_0\left(\left\{x : \ell\left(f_n\left(\hat{X}^{(k)}_{1:n}, f^\star_k\left(\hat{X}^{(k)}_{1:n}; \mathbf{U}_{<k}, \mathbf{V}_{<k}, \mathbf{W}\right), x\right), y_0\right) \geq \ell(y_0, y_1)/2\right\}\right) \geq 3/4\right) < 2^{-k}, \tag{40}$$

then fix some such value of n, and $\forall i \in \{n_{k-1}+1, \ldots, n\}$ define $X_i^{(k)}(\mathbf{u}_{<k}, \mathbf{u}_k, \mathbf{v}_{<k}, \mathbf{v}_k, \mathbf{w}) = \tilde{X}_i^{(k)}(\mathbf{u}_{<k}, \mathbf{u}_k, \mathbf{v}_{<k}, \mathbf{v}_k, \mathbf{w})$. Furthermore, for each $i \in \mathbb{N}$ with $n + 1 \leq i \leq n^2$, define $X_i^{(k)}(\mathbf{u}_{<k}, \mathbf{u}_k, \mathbf{v}_{<k}, \mathbf{v}_k, \mathbf{w}) = w_i$. Finally, define $n_k = n^2$. Otherwise, if no such n satisfies (40), then $\forall i \in \mathbb{N}$ with $i > n_{k-1}$, define $X_i^{(k)}(\mathbf{u}_{<k}, \mathbf{u}_k, \mathbf{v}_{<k}, \mathbf{v}_k, \mathbf{w}) = \tilde{X}_i^{(k)}(\mathbf{u}_{<k}, \mathbf{u}_k, \mathbf{v}_{<k}, \mathbf{v}_k, \mathbf{w})$, in which case the inductive definition is complete (upon reaching the smallest value of k for which no such n exists). Note that, since we do not condition on any variables in (40), the values n_k are *not* random.

Now we consider two cases. First, suppose there is a maximum value k^* of $k \in \mathbb{N}$ for which n_{k-1} is defined. In this case, $\nexists n \in \mathbb{N}$ with $n > n_{k^*-1}$ satisfying (40) with $k = k^*$, and furthermore $X_i^{(k^*)}(\mathbf{u}_{<k^*}, \mathbf{u}_{k^*}, \mathbf{v}_{<k^*}, \mathbf{v}_{k^*}, \mathbf{w}) = \tilde{X}_i^{(k^*)}(\mathbf{u}_{<k^*}, \mathbf{u}_{k^*}, \mathbf{v}_{<k^*}, \mathbf{v}_{k^*}, \mathbf{w})$ for every $i \in \mathbb{N}$, and every $\mathbf{u}$, $\mathbf{v}$, and $\mathbf{w}$. Next note that, by the law of total probability and basic limit theorems for probabilities (see e.g., Ash and Doléans-Dade, 2000), denoting $\mathbf{Q}_{k^*} = (\mathbf{U}_{<k^*}, \mathbf{V}_{<k^*}, \mathbf{W})$ for brevity,

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{P}\left(\limsup_{n\to\infty}\left\{\pi_0\left(\left\{x : \ell\left(f_n\left(\hat{X}^{(k^*)}_{1:n}, f^\star_{k^*}\left(\hat{X}^{(k^*)}_{1:n}; \mathbf{Q}_{k^*}\right), x\right), y_0\right) \geq \ell(y_0, y_1)/2\right\}\right) \geq 3/4\right\} \middle| \mathbf{Q}_{k^*}\right)\right] \\
&= \mathbb{P}\left(\limsup_{n\to\infty}\left\{\pi_0\left(\left\{x : \ell\left(f_n\left(\hat{X}^{(k^*)}_{1:n}, f^\star_{k^*}\left(\hat{X}^{(k^*)}_{1:n}; \mathbf{Q}_{k^*}\right), x\right), y_0\right) \geq \ell(y_0, y_1)/2\right\}\right) \geq 3/4\right\}\right) \\
&\geq \limsup_{n\to\infty}\mathbb{P}\left(\pi_0\left(\left\{x : \ell\left(f_n\left(\hat{X}^{(k^*)}_{1:n}, f^\star_{k^*}\left(\hat{X}^{(k^*)}_{1:n}; \mathbf{Q}_{k^*}\right), x\right), y_0\right) \geq \ell(y_0, y_1)/2\right\}\right) \geq 3/4\right).
\end{aligned}$$

The negation of (40) implies this last expression is at least 2^{-k^*} (noting that the negation of (40) holds for *every* $n > n_{k^*-1}$ in the present case). In particular, since the $U_{k,j}$,$V_{k',j'}$, and $W_{j''}$ variables are all independent, this implies $\exists \mathbf{u}, \mathbf{v}, \mathbf{w}$ such that, taking $X_i = X_i^{(k^*)}(\mathbf{u}_{<k^*}, \mathbf{U}_{k^*}, \mathbf{v}_{<k^*}, \mathbf{V}_{k^*}, \mathbf{w})$ for every $i \in \mathbb{N}$, and $f^\star(\cdot) = f^\star_{k^*}(\cdot; \mathbf{u}_{<k^*}, \mathbf{v}_{<k^*}, \mathbf{w})$, we have

$$\mathbb{P}\left(\limsup_{n\to\infty}\left\{\pi_0(\{x : \ell(f_n(X_{1:n}, f^\star(X_{1:n}), x), y_0) \geq \ell(y_0, y_1)/2\}) \geq 3/4\right\}\right) \geq 2^{-k^*}.$$

Define the event

$$E' = \left\{\limsup_{n\to\infty}\pi_0(\{x \in R_{k^*} : \ell(f_n(X_{1:n}, f^\star(X_{1:n}), x), y_0) \geq \ell(y_0, y_1)/2\}) \geq 1/4\right\}.$$

Since $\pi_0(R_{k^*}) = 1/2$, we have that

$$\begin{aligned}&\limsup_{n\to\infty} \{\pi_0(\{x : \ell(f_n(X_{1:n}, f^\star(X_{1:n}), x), y_0) \geq \ell(y_0, y_1)/2\}) \geq 3/4\} \\ &\subseteq \limsup_{n\to\infty} \{\pi_0(\{x \in R_{k^*} : \ell(f_n(X_{1:n}, f^\star(X_{1:n}), x), y_0) \geq \ell(y_0, y_1)/2\}) \geq 1/4\} \subseteq E',\end{aligned}$$

so that E' has probability at least 2^{-k^*}. Also let E denote the event that $\forall k, j \in \mathbb{N}$, $V_{k,j} \notin \{w_{j'} : j' \in \mathbb{N}\} \cup \{u_{k',j'} : k', j' \in \mathbb{N}\}$; note that, since π_0 is nonatomic, and hence so is each $\pi_0(\cdot|R_k)$ (since $\pi_0(R_k) > 0$), E has probability one.

Denote $t_i = n_{k^*-1} + k^*(i-1) + 1$ for each $i \in \mathbb{N}$, and let $I_{k^*} = \{t_i : i \in \mathbb{N}\}$. Note that, since every $V_{k^*,j} \in R_{k^*}$ and every $t \in I_{k^*}$ has $X_t = V_{k^*,t}$ (by definition), on the event E, every $t \in I_{k^*}$ has $f^\star(X_t) = y_0$ (by definition of $f^\star$). Therefore, on the event E, every $n \in \mathbb{N}$ with $n > n_{k^*-1}$ has

$$\begin{aligned}\hat{\mathcal{L}}_{\mathbb{X}}(f_n, f^\star; n) &\geq \limsup_{m\to\infty} \frac{1}{m} \sum_{t=n+1}^{n+m} \mathbb{1}_{I_{k^*}}(t) \ell(f_n(X_{1:n}, f^\star(X_{1:n}), X_t), f^\star(X_t)) \\ &= \limsup_{m\to\infty} \frac{1}{m} \sum_{t=n+1}^{n+m} \mathbb{1}_{I_{k^*}}(t) \ell(f_n(X_{1:n}, f^\star(X_{1:n}), X_t), y_0).\end{aligned}$$

Since $k^* \sum_{t=n+1}^{n+m} \mathbb{1}_{I_{k^*}}(t) > m - k^*$, letting $i_n = \max\{i \in \mathbb{N} : t_i \leq n\}$, this last line is at least as large as

$$\begin{aligned}&\limsup_{q\to\infty} \frac{1}{k^* q + k^*} \sum_{s=1}^{q} \ell\big(f_n\big(X_{1:n}, f^\star(X_{1:n}), X_{t_{i_n+s}}\big), y_0\big) \\ &= \limsup_{q\to\infty} \frac{1}{k^* q} \sum_{s=1}^{q} \ell\big(f_n\big(X_{1:n}, f^\star(X_{1:n}), X_{t_{i_n+s}}\big), y_0\big) \\ &\geq \limsup_{q\to\infty} \frac{1}{k^* q} \sum_{s=1}^{q} \mathbb{1}_{[\ell(y_0,y_1)/2,\infty)}\big(\ell\big(f_n\big(X_{1:n}, f^\star(X_{1:n}), X_{t_{i_n+s}}\big), y_0\big)\big) \frac{\ell(y_0, y_1)}{2}.\end{aligned}$$

Furthermore, the subsequence $\{X_{t_{i_n+s}}\}_{s=1}^{\infty}$ is a sequence of independent random variables with distribution $\pi_0(\cdot|R_{k^*})$ (namely, a subsequence of $\mathbf{V}_{k^*}$), also independent from the rest of the sequence $\{X_t : t \notin \{t_{i_n+s} : s \in \mathbb{N}\}\}$ and f_n. This implies that

$$\big\{\mathbb{1}_{[\ell(y_0,y_1)/2,\infty)}\big(\ell\big(f_n\big(X_{1:n}, f^\star(X_{1:n}), X_{t_{i_n+s}}\big), y_0\big)\big)\big\}_{s=1}^{\infty}$$

is a sequence of conditionally i.i.d. Bernoulli random variables (given $X_{1:n}$ and f_n). Thus, $\forall n \in \mathbb{N}$ with $n > n_{k^*-1}$, by the strong law of large numbers (applied under the conditional distribution given $X_{1:n}$ and f_n) and the law of total probability, there is an event E''_n of probability one such that, on $E \cap E''_n$,

$$\begin{aligned}
&\limsup_{q \to \infty} \frac{1}{k^* q} \sum_{s=1}^{q} \mathbb{1}_{[\ell(y_0,y_1)/2,\infty)}\left(\ell\left(f_n\left(X_{1:n}, f^\star(X_{1:n}), X_{t_{i_n+s}}\right), y_0\right)\right) \frac{\ell(y_0,y_1)}{2} \\
&= \frac{\ell(y_0,y_1)}{2k^*} \pi_0\left(\left\{x : \ell(f_n(X_{1:n}, f^\star(X_{1:n}), x), y_0) \geq \frac{\ell(y_0,y_1)}{2}\right\} \middle| R_{k^*}\right) \\
&= \frac{\ell(y_0,y_1)}{k^*} \pi_0\left(\left\{x \in R_{k^*} : \ell(f_n(X_{1:n}, f^\star(X_{1:n}), x), y_0) \geq \frac{\ell(y_0,y_1)}{2}\right\}\right).
\end{aligned}$$

Combining this with the above, we have that on the event $E \cap E' \cap \bigcap_{n > n_{k^*-1}} E''_n$,

$$\begin{aligned}
&\limsup_{n \to \infty} \hat{\mathcal{L}}_{\mathbb{X}}(f_n, f^\star; n) \\
&\geq \frac{\ell(y_0,y_1)}{k^*} \limsup_{n \to \infty} \pi_0\left(\left\{x \in R_{k^*} : \ell(f_n(X_{1:n}, f^\star(X_{1:n}), x), y_0) \geq \frac{\ell(y_0,y_1)}{2}\right\}\right) \geq \frac{\ell(y_0,y_1)}{4k^*}.
\end{aligned}$$

Since $\frac{\ell(y_0,y_1)}{4k^*} > 0$, and since $E \cap E' \cap \bigcap_{n > n_{k^*-1}} E''_n$ has probability at least $2^{-k^*} > 0$ (by the union bound), and since $f^\star$ is clearly a measurable function, this implies that f_n is not strongly universally consistent under the process $\mathbb{X}$ defined here.

To complete this first case, we argue that $\mathbb{X} \in$ SUIL; in fact, we will show the stronger result that $\mathbb{X} \in$ CRF. Note that for every $t > n_{k^*-1}$, either $t \in I_{k^*}$, in which case $X_t = V_{k^*,t}$, or else $X_t = U_{k^*,t}$. Thus, for any $n > n_{k^*-1}$ and $A \in \mathcal{B}$,

$$\frac{1}{n} \sum_{t=n_{k^*-1}+1}^{n} \mathbb{1}_A(X_t) = \frac{1}{n} \sum_{t=n_{k^*-1}+1}^{n} \left(\mathbb{1}_{I_{k^*}}(t)\mathbb{1}_A(V_{k^*,t}) + \mathbb{1}_{\mathbb{N} \setminus I_{k^*}}(t)\mathbb{1}_A(U_{k^*,t})\right).$$

Thus, since $0 \leq \frac{1}{n} \sum_{t=1}^{n_{k^*-1}} \mathbb{1}_A(X_t) \leq \frac{n_{k^*-1}}{n}$ implies $\lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n_{k^*-1}} \mathbb{1}_A(X_t) = 0$, we have

$$\lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{1}_A(X_t) = \lim_{n \to \infty} \frac{1}{n} \sum_{t=n_{k^*-1}+1}^{n} \mathbb{1}_{I_{k^*}}(t)\mathbb{1}_A(V_{k^*,t}) + \lim_{n \to \infty} \frac{1}{n} \sum_{t=n_{k^*-1}+1}^{n} \mathbb{1}_{\mathbb{N} \setminus I_{k^*}}(t)\mathbb{1}_A(U_{k^*,t}).$$

For any $n \in \mathbb{N}$ with $n > n_{k^*-1}$, denoting by $q_n = \sum_{t=n_{k^*-1}+1}^{n} \mathbb{1}_{I_{k^*}}(t)$, we have $n - n_{k^*-1} \leq k^* q_n < n - n_{k^*-1} + k^*$. Therefore

$$\lim_{n \to \infty} \frac{k^* q_n}{n} = 1, \tag{41}$$

so that

$$\begin{aligned}
\lim_{n \to \infty} \frac{1}{n} \sum_{t=n_{k^*-1}+1}^{n} \mathbb{1}_{I_{k^*}}(t)\mathbb{1}_A(V_{k^*,t}) &= \lim_{n \to \infty} \frac{k^* q_n}{n} \frac{1}{k^* q_n} \sum_{t=n_{k^*-1}+1}^{n} \mathbb{1}_{I_{k^*}}(t)\mathbb{1}_A(V_{k^*,t}) \\
&= \frac{1}{k^*} \lim_{n \to \infty} \frac{1}{q_n} \sum_{i=1}^{q_n} \mathbb{1}_A(V_{k^*,t_i}) = \frac{1}{k^*} \lim_{s \to \infty} \frac{1}{s} \sum_{i=1}^{s} \mathbb{1}_A(V_{k^*,t_i}).
\end{aligned}$$

Since the variables $\{V_{k^*,t_i}\}_{i=1}^{\infty}$ are independent $\pi_0(\cdot|R_{k^*})$-distributed random variables, the strong law of large numbers implies that with probability one, the rightmost expression above equals $(1/k^*)\pi_0(A|R_{k^*})$. Likewise, for any $n \in \mathbb{N}$ with $n > n_{k^*-1}$, denoting by $q'_n = \sum_{t=n_{k^*-1}+1}^{n} \mathbb{1}_{\mathbb{N}\setminus I_{k^*}}(t)$ and noting that $q'_n = n - n_{k^*-1} - q_n$, (41) implies

$$\lim_{n\to\infty} \frac{q'_n}{n} = \lim_{n\to\infty} \frac{k^*(n - n_{k^*-1} - q_n)}{k^* n} = \frac{k^*-1}{k^*}.$$

Therefore, if $k^* > 1$, denoting by $t'_1, t'_2, \ldots$ the enumeration of the elements of $\{t \in \mathbb{N}\setminus I_{k^*} : t > n_{k^*-1}\}$ in increasing order, we have that

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n} \sum_{t=n_{k^*-1}+1}^{n} \mathbb{1}_{\mathbb{N}\setminus I_{k^*}}(t)\mathbb{1}_A(U_{k^*,t}) &= \lim_{n\to\infty} \frac{q'_n}{n}\frac{1}{q'_n} \sum_{t=n_{k^*-1}+1}^{n} \mathbb{1}_{\mathbb{N}\setminus I_{k^*}}(t)\mathbb{1}_A(U_{k^*,t}) \\ &= \frac{k^*-1}{k^*} \lim_{n\to\infty} \frac{1}{q'_n} \sum_{i=1}^{q'_n} \mathbb{1}_A(U_{k^*,t'_i}) = \frac{k^*-1}{k^*} \lim_{s\to\infty} \frac{1}{s} \sum_{i=1}^{s} \mathbb{1}_A(U_{k^*,t'_i}).\end{aligned}$$

Since the variables $\{U_{k^*,t'_i}\}_{i=1}^{\infty}$ are independent $\pi_0(\cdot|\mathcal{X}\setminus R_{k^*})$-distributed random variables, the strong law of large numbers implies that on an event of probability one, the rightmost expression above equals $\frac{k^*-1}{k^*}\pi_0(A|\mathcal{X}\setminus R_{k^*})$, so that on this event,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=n_{k^*-1}+1}^{n} \mathbb{1}_{\mathbb{N}\setminus I_{k^*}}(t)\mathbb{1}_A(U_{k^*,t}) = \frac{k^*-1}{k^*}\pi_0(A|\mathcal{X}\setminus R_{k^*}).$$

Although this argument was specific to $k^* > 1$, the above equality is trivially also satisfied if $k^* = 1$, so that the conclusion holds for any $k^* \in \mathbb{N}$.

By the union bound, both of the above events occur simultaneously with probability one, so that altogether we have that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{1}_A(X_t) = \frac{1}{k^*}\pi_0(A|R_{k^*}) + \frac{k^*-1}{k^*}\pi_0(A|\mathcal{X}\setminus R_{k^*}) \text{ (a.s.).} \tag{42}$$

In particular, this also establishes that the limit of the expression on the left hand side *exists* almost surely. Since this conclusion holds for any choice of $A \in \mathcal{B}$, we have that $\mathbb{X} \in \text{CRF}$. Thus, since Theorem 17 of Section 3 establishes that $\text{CRF} \subseteq \mathcal{C}_1$, we have that $\mathbb{X} \in \mathcal{C}_1$, and since Theorem 7 establishes that $\text{SUIL} = \mathcal{C}_1$, this implies $\mathbb{X} \in \text{SUIL}$. Therefore, in this first case, we conclude that the inductive learning rule f_n is not optimistically universal.

Next, let us examine the second case, wherein n_k is defined for every $k \in \mathbb{N}\cup\{0\}$, so that $\{n_k\}_{k=0}^{\infty}$ is an infinite increasing sequence of nonnegative integers. In this case, for every $k \in \mathbb{N}$, (40) and the definition of n_k imply that, denoting by $\mathbf{Q}_k = (\mathbf{U}_{<k}, \mathbf{V}_{<k}, \mathbf{W})$,

$$\mathbb{P}\Big(\pi_0\Big(\Big\{x : \ell\Big(f_{\sqrt{n_k}}\Big(\hat{X}^{(k)}_{1:\sqrt{n_k}}, f^{\star}_k\Big(\hat{X}^{(k)}_{1:\sqrt{n_k}}; \mathbf{Q}_k\Big), x\Big), y_0\Big) \geq \ell(y_0,y_1)/2\Big\}\Big) \geq 3/4\Big) < 2^{-k}.$$

By the monotone convergence theorem and linearity of expectations, combined with the law of total probability, this implies

$$\begin{aligned}&\mathbb{E}\left[\sum_{k=1}^{\infty}\mathbb{P}\left(\pi_0\left(\left\{x:\ell\left(f_{\sqrt{n_k}}\left(\hat{X}^{(k)}_{1:\sqrt{n_k}},f^\star_k\left(\hat{X}^{(k)}_{1:\sqrt{n_k}};\mathbf{Q}_k\right),x\right),y_0\right)\geq\ell(y_0,y_1)/2\right\}\right)\geq 3/4\middle|\mathbf{V}\right)\right]\\&=\sum_{k=1}^{\infty}\mathbb{P}\left(\pi_0\left(\left\{x:\ell\left(f_{\sqrt{n_k}}\left(\hat{X}^{(k)}_{1:\sqrt{n_k}},f^\star_k\left(\hat{X}^{(k)}_{1:\sqrt{n_k}};\mathbf{Q}_k\right),x\right),y_0\right)\geq\ell(y_0,y_1)/2\right\}\right)\geq 3/4\right)<1.\end{aligned}$$

In particular, this implies that with probability one,

$$\sum_{k=1}^{\infty}\mathbb{P}\left(\pi_0\left(\left\{x:\ell\left(f_{\sqrt{n_k}}\left(\hat{X}^{(k)}_{1:\sqrt{n_k}},f^\star_k\left(\hat{X}^{(k)}_{1:\sqrt{n_k}};\mathbf{Q}_k\right),x\right),y_0\right)\geq\ell(y_0,y_1)/2\right\}\right)\geq 3/4\middle|\mathbf{V}\right)<\infty.$$

Since $\mathbf{U}$, $\mathbf{W}$, and $\{f_n\}_{n\in\mathbb{N}}$ are independent from $\mathbf{V}$, and since every $k,j\in\mathbb{N}$ has $V_{k,j}$ with distribution $\pi_0(\cdot|R_k)$ and hence $V_{k,j}\in R_k$, this implies $\exists\mathbf{v}$ with $v_{k,j}\in R_k$ for every $k,j\in\mathbb{N}$, such that, defining $X_i=X^{(k)}_i(\mathbf{U}_{<k},\mathbf{U}_k,\mathbf{v}_{<k},\mathbf{v}_k,\mathbf{W})$ for every $k\in\mathbb{N}$ and $i\in\{n_{k-1}+1,\ldots,n_k\}$,

$$\sum_{k=1}^{\infty}\mathbb{P}\left(\pi_0\left(\left\{x:\ell\left(f_{\sqrt{n_k}}\left(X_{1:\sqrt{n_k}},f^\star_k\left(X_{1:\sqrt{n_k}};\mathbf{U}_{<k},\mathbf{v}_{<k},\mathbf{W}\right),x\right),y_0\right)\geq\ell(y_0,y_1)/2\right\}\right)\geq 3/4\right)<\infty.$$

The Borel-Cantelli Lemma then implies that there exists an event H' of probability one, on which $\exists k_0\in\mathbb{N}$ such that, $\forall k\in\mathbb{N}$ with $k>k_0$,

$$\pi_0\left(\left\{x:\ell\left(f_{\sqrt{n_k}}\left(X_{1:\sqrt{n_k}},f^\star_k\left(X_{1:\sqrt{n_k}};\mathbf{U}_{<k},\mathbf{v}_{<k},\mathbf{W}\right),x\right),y_0\right)\geq\ell(y_0,y_1)/2\right\}\right)<3/4.$$

Next, let H denote the event that $\{W_j:j\in\mathbb{N}\}\cap\{v_{k,j}:k,j\in\mathbb{N}\}=\emptyset$ and $\{U_{k,j}:k,j\in\mathbb{N}\}\cap\{v_{k,j}:k,j\in\mathbb{N}\}=\emptyset$. Note that, since π_0 is nonatomic, and so is $\pi_0(\cdot|\mathcal{X}\setminus R_k)$ for every $k\in\mathbb{N}$, H has probability one. Furthermore, for every $k\in\mathbb{N}$, by definition of $f^\star_k$, $\forall j\in\mathbb{N}$, $f^\star_k(W_j;\mathbf{U}_{<k},\mathbf{v}_{<k},\mathbf{W})=y_1$, and $\forall k',j\in\mathbb{N}$ with $k'<k$, $f^\star_k(U_{k',j};\mathbf{U}_{<k},\mathbf{v}_{<k},\mathbf{W})=y_1$. Also, for every $j\in\mathbb{N}$, the distribution of $U_{k,j}$ is $\pi_0(\cdot|\mathcal{X}\setminus R_k)$, and therefore we have $U_{k,j}\notin R_k$; together with the definition of $f^\star_k$, this implies $f^\star_k(U_{k,j};\mathbf{U}_{<k},\mathbf{v}_{<k},\mathbf{W})=y_1$ on the event H. The definition of $f^\star_k$ further implies that, on H, for every $k',k,j\in\mathbb{N}$ with $k'<k$, $f^\star_k(v_{k',j};\mathbf{U}_{<k},\mathbf{v}_{<k},\mathbf{W})=y_0$. Also, since $v_{k,j}\in R_k$ for every $k,j\in\mathbb{N}$, on the event H, every $k,j\in\mathbb{N}$ has $f^\star_k(v_{k,j};\mathbf{U}_{<k},\mathbf{v}_{<k},\mathbf{W})=y_0$. Furthermore, by definition of $f^\star_0$, every $k,j\in\mathbb{N}$ has $f^\star_0(v_{k,j};\mathbf{v})=y_0$, and on the event H, every $j\in\mathbb{N}$ has $f^\star_0(W_j;\mathbf{v})=y_1$, and $\forall k,j\in\mathbb{N}$, $f^\star_0(U_{k,j};\mathbf{v})=y_1$. Altogether we have that, on the event H, every $k',k,j\in\mathbb{N}$ with $k'\leq k$ has $f^\star_k(v_{k',j};\mathbf{U}_{<k},\mathbf{v}_{<k},\mathbf{W})=f^\star_0(v_{k',j};\mathbf{v})$, $f^\star_k(U_{k',j};\mathbf{U}_{<k},\mathbf{v}_{<k},\mathbf{W})=f^\star_0(U_{k',j};\mathbf{v})$, and $f^\star_k(W_j;\mathbf{U}_{<k},\mathbf{v}_{<k},\mathbf{W})=f^\star_0(W_j;\mathbf{v})$. In particular, note that for any $k\in\mathbb{N}$, every $i\in\left\{1,\ldots,\sqrt{n_k}\right\}$ has $X_i\in\{v_{k',i}:k'\leq k\}\cup\{U_{k',i}:k'\leq k\}\cup\{W_i\}$, so that, on the event H, $f^\star_k(X_i;\mathbf{U}_{<k},\mathbf{v}_{<k},\mathbf{W})=f^\star_0(X_i;\mathbf{v})$. Thus, taking $f^\star(\cdot)=f^\star_0(\cdot;\mathbf{v})$, on the event $H\cap H'$, $\forall k\in\mathbb{N}$ with $k>k_0$,

$$\pi_0\left(\left\{x:\ell\left(f_{\sqrt{n_k}}\left(X_{1:\sqrt{n_k}},f^\star(X_{1:\sqrt{n_k}}),x\right),y_0\right)\geq\ell(y_0,y_1)/2\right\}\right)<3/4.$$

Any $y \in \mathcal{Y}$ with $\ell(y, y_1) < \ell(y_0, y_1)/2$ necessarily has $\ell(y, y_0) > \ell(y, y_0) + \ell(y, y_1) - \ell(y_0, y_1)/2 \geq \ell(y_0, y_1) - \ell(y_0, y_1)/2 = \ell(y_0, y_1)/2$, where the second inequality is due to the triangle inequality for ℓ. Therefore, on $H \cap H'$, $\forall k \in \mathbb{N}$ with $k > k_0$,

$$\begin{aligned}
&\pi_0\left(\left\{x : \ell\left(f_{\sqrt{n_k}}\left(X_{1:\sqrt{n_k}}, f^\star(X_{1:\sqrt{n_k}}), x\right), y_1\right) \geq \ell(y_0, y_1)/2\right\}\right) \\
&= 1 - \pi_0\left(\left\{x : \ell\left(f_{\sqrt{n_k}}\left(X_{1:\sqrt{n_k}}, f^\star(X_{1:\sqrt{n_k}}), x\right), y_1\right) < \ell(y_0, y_1)/2\right\}\right) \\
&\geq 1 - \pi_0\left(\left\{x : \ell\left(f_{\sqrt{n_k}}\left(X_{1:\sqrt{n_k}}, f^\star(X_{1:\sqrt{n_k}}), x\right), y_0\right) > \ell(y_0, y_1)/2\right\}\right) > 1/4. \qquad (43)
\end{aligned}$$

Now fix any $k, k' \in \mathbb{N}$ with $k' \geq k$ and $k' > 1$ (which implies $n_{k'} > \sqrt{n_{k'}}$), and note that every $t \in \{\sqrt{n_{k'}} + 1, \ldots, n_{k'}\}$ has $X_t = W_t$; on H, this implies $f^\star(X_t) = y_1$. Thus, on the event H,

$$\begin{aligned}
&\frac{1}{n_{k'} - \sqrt{n_{k'}}} \sum_{t=\sqrt{n_{k'}}+1}^{n_{k'}} \ell\left(f_{\sqrt{n_k}}\left(X_{1:\sqrt{n_k}}, f^\star\left(X_{1:\sqrt{n_k}}\right), X_t\right), f^\star(X_t)\right) \\
&= \frac{1}{n_{k'} - \sqrt{n_{k'}}} \sum_{t=\sqrt{n_{k'}}+1}^{n_{k'}} \ell\left(f_{\sqrt{n_k}}\left(X_{1:\sqrt{n_k}}, f^\star\left(X_{1:\sqrt{n_k}}\right), X_t\right), y_1\right) \\
&\geq \frac{1}{n_{k'} - \sqrt{n_{k'}}} \sum_{t=\sqrt{n_{k'}}+1}^{n_{k'}} \mathbb{1}_{[\ell(y_0,y_1)/2,\infty)}\left(\ell\left(f_{\sqrt{n_k}}\left(X_{1:\sqrt{n_k}}, f^\star\left(X_{1:\sqrt{n_k}}\right), X_t\right), y_1\right)\right) \frac{\ell(y_0, y_1)}{2}.
\end{aligned}$$

Furthermore, the fact that $\{X_t\}_{t=\sqrt{n_{k'}}+1}^{n_{k'}} = \{W_t\}_{t=\sqrt{n_{k'}}+1}^{n_{k'}}$ also implies that $\{X_t\}_{t=\sqrt{n_{k'}}+1}^{n_{k'}}$ are independent π_0-distributed random variables, also independent from $X_{1:\sqrt{n_k}}$ (since $k \leq k'$) and $f_{\sqrt{n_k}}$. Therefore, Hoeffding's inequality (applied under the conditional distribution given $X_{1:\sqrt{n_k}}$ and $f_{\sqrt{n_k}}$) and the law of total probability imply that, on an event $H''_{k,k'}$ of probability at least $1 - \frac{1}{(k')^3}$,

$$\begin{aligned}
&\frac{1}{n_{k'} - \sqrt{n_{k'}}} \sum_{t=\sqrt{n_{k'}}+1}^{n_{k'}} \mathbb{1}_{[\ell(y_0,y_1)/2,\infty)}\left(\ell\left(f_{\sqrt{n_k}}\left(X_{1:\sqrt{n_k}}, f^\star\left(X_{1:\sqrt{n_k}}\right), X_t\right), y_1\right)\right) \\
&\geq \pi_0\left(\left\{x : \ell\left(f_{\sqrt{n_k}}\left(X_{1:\sqrt{n_k}}, f^\star\left(X_{1:\sqrt{n_k}}\right), x\right), y_1\right) \geq \ell(y_0, y_1)/2\right\}\right) - \sqrt{\frac{(3/2)\ln(k')}{n_{k'} - \sqrt{n_{k'}}}}.
\end{aligned}$$

Combining with (43) we have that, on the event $H \cap H' \cap \bigcap_{k' \in \mathbb{N}\setminus\{1\}} \bigcap_{k \leq k'} H''_{k,k'}$, every $k, k' \in \mathbb{N}$ with $k' \geq k > k_0$ satisfy

$$\begin{aligned}
&\frac{1}{n_{k'} - \sqrt{n_{k'}}} \sum_{t=\sqrt{n_{k'}}+1}^{n_{k'}} \ell\left(f_{\sqrt{n_k}}\left(X_{1:\sqrt{n_k}}, f^\star\left(X_{1:\sqrt{n_k}}\right), X_t\right), f^\star(X_t)\right) \\
&\qquad > \frac{\ell(y_0, y_1)}{2}\left(\frac{1}{4} - \sqrt{\frac{(3/2)\ln(k')}{n_{k'} - \sqrt{n_{k'}}}}\right).
\end{aligned}$$

Since n_k is strictly increasing in k, we have that on $H \cap H' \cap \bigcap_{k' \in \mathbb{N} \setminus \{1\}} \bigcap_{k \le k'} H''_{k,k'}$,

$$\begin{aligned}
\limsup_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(f_n, f^\star; n) &\ge \limsup_{k\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}\left(f_{\sqrt{n_k}}, f^\star; \sqrt{n_k}\right) \\
&= \limsup_{k\to\infty} \limsup_{m\to\infty} \frac{1}{m} \sum_{t=1}^{m} \ell\left(f_{\sqrt{n_k}}\left(X_{1:\sqrt{n_k}}, f^\star\left(X_{1:\sqrt{n_k}}\right), X_t\right), f^\star(X_t)\right) \\
&\ge \limsup_{k\to\infty} \limsup_{k'\to\infty} \frac{1}{n_{k'}} \sum_{t=\sqrt{n_{k'}}+1}^{n_{k'}} \ell\left(f_{\sqrt{n_k}}\left(X_{1:\sqrt{n_k}}, f^\star\left(X_{1:\sqrt{n_k}}\right), X_t\right), f^\star(X_t)\right) \\
&\ge \limsup_{k\to\infty} \limsup_{k'\to\infty} \frac{n_{k'} - \sqrt{n_{k'}}}{n_{k'}} \frac{\ell(y_0, y_1)}{2} \left(\frac{1}{4} - \sqrt{\frac{(3/2)\ln(k')}{n_{k'} - \sqrt{n_{k'}}}}\right). \qquad (44)
\end{aligned}$$

Since $n_{k'}$ is strictly increasing in k', we have that for any $k' \ge 4$, $0 \le \frac{(3/2)\ln(k')}{n_{k'} - \sqrt{n_{k'}}} \le \frac{3\ln(n_{k'})}{n_{k'}}$, which converges to 0 as $k' \to \infty$. Furthermore, $\frac{n_{k'} - \sqrt{n_{k'}}}{n_{k'}} = 1 - \frac{1}{\sqrt{n_{k'}}}$, which converges to 1 as $k' \to \infty$. Therefore, the expression in (44) equals $\ell(y_0, y_1)/8$. By the union bound, the event $H \cap H' \cap \bigcap_{k' \in \mathbb{N} \setminus \{1\}} \bigcap_{k \le k'} H''_{k,k'}$ has probability at least

$$1 - \sum_{k' \in \mathbb{N} \setminus \{1\}} \sum_{k \le k'} \frac{1}{(k')^3} = 1 - \sum_{k' \in \mathbb{N} \setminus \{1\}} \frac{1}{(k')^2} = 2 - \frac{\pi^2}{6} > 0,$$

so that there is a nonzero probability that $\limsup_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(f_n, f^\star; n) > 0$. In particular, since $f^\star$ is clearly a measurable function, this implies that the inductive learning rule f_n is not strongly universally consistent under $\mathbb{X}$.

It remains to show that the process $\mathbb{X}$ defined above for this second case is an element of SUIL; again, we will in fact establish the stronger fact that $\mathbb{X} \in$ CRF. For this, for each $k \in \mathbb{N}$, let $J_k = \{n_{k-1} + (i-1)k + 1 : i \in \mathbb{N}, n_{k-1} + (i-1)k + 1 \le \sqrt{n_k}\}$. For any $n \in \mathbb{N}$, denote $k_n = \max\{k \in \mathbb{N} : n_{k-1} < n\}$; this is well-defined, since $n_0 = 0$ (so that this set of k values is nonempty), and n_k is strictly increasing (so that this set of k values is finite, and hence has a maximum value). Note that, since n_k is finite for every k, it follows that $k_n \to \infty$. Fix any $A \in \mathcal{B}$. By the construction of the process above, we have that, $\forall n \in \mathbb{N}$,

$$\frac{1}{n} \sum_{t=1}^{n} \mathbb{1}_A(X_t) = \frac{1}{n} \sum_{k=1}^{k_n} \left(\left(\sum_{t=n_{k-1}+1}^{\min\{\sqrt{n_k}, n\}} \left(\mathbb{1}_{J_k}(t)\mathbb{1}_A(v_{k,t}) + \mathbb{1}_{\mathbb{N} \setminus J_k}(t)\mathbb{1}_A(U_{k,t})\right)\right) + \sum_{t=\sqrt{n_k}+1}^{\min\{n_k, n\}} \mathbb{1}_A(W_t)\right). \qquad (45)$$

By Kolmogorov's strong law of large numbers, with probability one we have

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{k_n} \left(\left(\sum_{t=n_{k-1}+1}^{\min\{\sqrt{n_k}, n\}} \left(\mathbb{1}_{J_k}(t)\left(\mathbb{1}_A(v_{k,t}) - \mathbb{1}_A(v_{k,t})\right) + \mathbb{1}_{\mathbb{N} \setminus J_k}(t)\left(\mathbb{1}_A(U_{k,t}) - \pi_0(A | \mathcal{X} \setminus R_k)\right)\right)\right) + \sum_{t=\sqrt{n_k}+1}^{\min\{n_k, n\}} \left(\mathbb{1}_A(W_t) - \pi_0(A)\right)\right) = 0. \qquad (46)$$

We therefore focus on establishing convergence of

$$\frac{1}{n}\sum_{k=1}^{k_n}\left(\left(\sum_{t=n_{k-1}+1}^{\min\{\sqrt{n_k},n\}}\left(\mathbb{1}_{J_k}(t)\mathbb{1}_A(v_{k,t})+\mathbb{1}_{\mathbb{N}\setminus J_k}(t)\pi_0(A|\mathcal{X}\setminus R_k)\right)\right)+\sum_{t=\sqrt{n_k}+1}^{\min\{n_k,n\}}\pi_0(A)\right). \quad (47)$$

Note that, for any $k,n\in\mathbb{N}$ with $n>n_{k-1}$,

$$|J_k\cap\{n_{k-1}+1,\ldots,\min\{\sqrt{n_k},n\}\}|=\left\lceil\frac{\min\{\sqrt{n_k},n\}-n_{k-1}}{k}\right\rceil\leq\frac{n}{k}+1,$$

and that $\max(J_{k-1})\leq\sqrt{n_{k-1}}$ for any $k>1$. Thus,

$$\begin{aligned}0\leq\frac{1}{n}\sum_{k=1}^{k_n}\sum_{t=n_{k-1}+1}^{\min\{\sqrt{n_k},n\}}\mathbb{1}_{J_k}(t)\mathbb{1}_A(v_{k,t})&\leq\frac{1}{n}\sum_{k=1}^{k_n}\sum_{t=n_{k-1}+1}^{\min\{\sqrt{n_k},n\}}\mathbb{1}_{J_k}(t)\\&\leq\frac{\sqrt{n_{k_n-1}}}{n}+\frac{1}{n}\left(\frac{n}{k_n}+1\right)=\frac{\sqrt{n_{k_n-1}}}{n}+\frac{1}{k_n}+\frac{1}{n}.\end{aligned} \quad (48)$$

By definition of k_n, this rightmost expression at most $\frac{1}{\sqrt{n}}+\frac{1}{k_n}+\frac{1}{n}$, which has limit 0 as $n\to\infty$ since $k_n\to\infty$. Thus,

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{k_n}\sum_{t=n_{k-1}+1}^{\min\{\sqrt{n_k},n\}}\mathbb{1}_{J_k}(t)\mathbb{1}_A(v_{k,t})=0. \quad (49)$$

By the definition of the R_k sequence, for any $\varepsilon\in(0,1)$, $\exists k_\varepsilon\in\mathbb{N}$ such that, $\forall k\geq k_\varepsilon$, $|\pi_0(A\cap R_k)-(1/2)\pi_0(A)|<\varepsilon/2$. For any $k\geq k_\varepsilon$, we have $\pi_0(A|\mathcal{X}\setminus R_k)=2\pi_0(A\cap(\mathcal{X}\setminus R_k))=2(\pi_0(A)-\pi_0(A\cap R_k))\in(2(\pi_0(A)-(1/2)\pi_0(A)-\varepsilon/2),2(\pi_0(A)-(1/2)\pi_0(A)+\varepsilon/2))=(\pi_0(A)-\varepsilon,\pi_0(A)+\varepsilon)$. Thus, for any $n\in\mathbb{N}$ with $k_n\geq k_\varepsilon$, we have that

$$\begin{aligned}&\frac{1}{n}\sum_{k=1}^{k_n}\left(\sum_{t=n_{k-1}+1}^{\min\{\sqrt{n_k},n\}}\mathbb{1}_{\mathbb{N}\setminus J_k}(t)\pi_0(A|\mathcal{X}\setminus R_k)+\sum_{t=\sqrt{n_k}+1}^{\min\{n_k,n\}}\pi_0(A)\right)\\&\geq-\varepsilon+\frac{1}{n}\sum_{k=k_\varepsilon}^{k_n}\sum_{t=n_{k-1}+1}^{\min\{n_k,n\}}\mathbb{1}_{\mathbb{N}\setminus J_k}(t)\pi_0(A)\geq-\varepsilon+\frac{1}{n}\sum_{k=k_\varepsilon}^{k_n}\sum_{t=n_{k-1}+1}^{\min\{n_k,n\}}(\pi_0(A)-\mathbb{1}_{J_k}(t))\\&\geq-\varepsilon-\left(\frac{1}{n}\sum_{k=1}^{k_n}\sum_{t=n_{k-1}+1}^{\min\{n_k,n\}}\mathbb{1}_{J_k}(t)\right)+\left(\frac{1}{n}\sum_{k=k_\varepsilon}^{k_n}\sum_{t=n_{k-1}+1}^{\min\{n_k,n\}}\pi_0(A)\right)\\&=-\varepsilon-\left(\frac{1}{n}\sum_{k=1}^{k_n}\sum_{t=n_{k-1}+1}^{\min\{\sqrt{n_k},n\}}\mathbb{1}_{J_k}(t)\right)+\left(1-\frac{n_{k_\varepsilon-1}}{n}\right)\pi_0(A)\end{aligned}$$

and

$$\begin{aligned}
&\frac{1}{n}\sum_{k=1}^{k_n}\left(\sum_{t=n_{k-1}+1}^{\min\{\sqrt{n_k},n\}} \mathbb{1}_{\mathbb{N}\setminus J_k}(t)\pi_0(A|\mathcal{X}\setminus R_k) + \sum_{t=\sqrt{n_k}+1}^{\min\{n_k,n\}} \pi_0(A)\right)\\
&\leq \frac{n_{k_\varepsilon-1}}{n} + \frac{1}{n}\sum_{k=k_\varepsilon}^{k_n}\left(\sum_{t=n_{k-1}+1}^{\min\{\sqrt{n_k},n\}} \pi_0(A|\mathcal{X}\setminus R_k) + \sum_{t=\sqrt{n_k}+1}^{\min\{n_k,n\}} \pi_0(A)\right)\\
&\leq \frac{n_{k_\varepsilon-1}}{n} + \frac{1}{n}\sum_{k=k_\varepsilon}^{k_n}\sum_{t=n_{k-1}+1}^{\min\{n_k,n\}} (\pi_0(A)+\varepsilon) \leq \frac{n_{k_\varepsilon-1}}{n} + \pi_0(A) + \varepsilon.
\end{aligned}$$

As mentioned above, the rightmost expression in (48) has limit 0. Therefore, the inequalities in (48) also imply that $\lim\limits_{n\to\infty} \frac{1}{n}\sum\limits_{k=1}^{k_n}\sum\limits_{t=n_{k-1}+1}^{\min\{\sqrt{n_k},n\}} \mathbb{1}_{J_k}(t) = 0$. Furthermore, for any fixed $\varepsilon \in (0,1)$, $\lim\limits_{n\to\infty} \frac{n_{k_\varepsilon-1}}{n} = 0$. Thus, we have that

$$\begin{aligned}
\pi_0(A) - \varepsilon &\leq \liminf_{n\to\infty} \frac{1}{n}\sum_{k=1}^{k_n}\left(\sum_{t=n_{k-1}+1}^{\min\{\sqrt{n_k},n\}} \mathbb{1}_{\mathbb{N}\setminus J_k}(t)\pi_0(A|\mathcal{X}\setminus R_k) + \sum_{t=\sqrt{n_k}+1}^{\min\{n_k,n\}} \pi_0(A)\right)\\
&\leq \limsup_{n\to\infty} \frac{1}{n}\sum_{k=1}^{k_n}\left(\sum_{t=n_{k-1}+1}^{\min\{\sqrt{n_k},n\}} \mathbb{1}_{\mathbb{N}\setminus J_k}(t)\pi_0(A|\mathcal{X}\setminus R_k) + \sum_{t=\sqrt{n_k}+1}^{\min\{n_k,n\}} \pi_0(A)\right) \leq \pi_0(A) + \varepsilon.
\end{aligned}$$

Taking the limit as $\varepsilon \to 0$ reveals that

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{k_n}\left(\sum_{t=n_{k-1}+1}^{\min\{\sqrt{n_k},n\}} \mathbb{1}_{\mathbb{N}\setminus J_k}(t)\pi_0(A|\mathcal{X}\setminus R_k) + \sum_{t=\sqrt{n_k}+1}^{\min\{n_k,n\}} \pi_0(A)\right) = \pi_0(A),$$

which also establishes that the limit exists. Combined with (49), (46), and (45), we have

$$\frac{1}{n}\sum_{t=1}^{n} \mathbb{1}_A(X_t) \to \pi_0(A) \text{ (a.s.).} \tag{50}$$

In particular, this implies that the limit of the left hand side *exists* almost surely. Since this holds for any choice of $A \in \mathcal{B}$, we have that $\mathbb{X} \in \mathrm{CRF}$. Since Theorem 17 of Section 3 establishes that $\mathrm{CRF} \subseteq \mathcal{C}_1$, this also implies $\mathbb{X} \in \mathcal{C}_1$, and since Theorem 7 establishes that $\mathrm{SUIL} = \mathcal{C}_1$, this further implies $\mathbb{X} \in \mathrm{SUIL}$. Thus, in this second case as well, we conclude that the inductive learning rule f_n is not optimistically universal. Since any inductive learning rule f_n satisfies one of these two cases, this completes the proof that no inductive learning rule is optimistically universal. ■

Combining this result with a simple technique for learning in countable spaces, we immediately have the following corollary.

Corollary 29 *There exists an optimistically universal inductive learning rule if and only if $\mathcal{X}$ is countable.*

Proof The "only if" part of the claim follows immediately from Theorem 6. For the "if" part, consider a simple inductive learning rule $\hat{f}_n$, defined as follows. For any $n \in \mathbb{N}$, $x_{1:n} \in \mathcal{X}^n$, $y_{1:n} \in \mathcal{Y}^n$, and $x \in \mathcal{X}$, if $x \in \{x_1, \ldots, x_n\}$, then letting $i(x; x_{1:n}) = \min\{i \in \{1, \ldots, n\} : x_i = x\}$, we define $\hat{f}_n(x_{1:n}, y_{1:n}, x) = y_{i(x;x_{1:n})}$. The value $\hat{f}_n(x_{1:n}, y_{1:n}, x)$ can be defined arbitrarily when $x \notin \{x_1, \ldots, x_n\}$. In other words, this method simply *memorizes* the observed data points (x_i, y_i), $i \in \{1, \ldots, n\}$, and if the test point x is among the observed x_i points, it simply reports the corresponding memorized y_i value.

Suppose $\mathcal{X}$ is countable, and enumerate its elements $\mathcal{X} = \{z_1, z_2, \ldots\}$ (or in the case of finite $|\mathcal{X}|$, $\mathcal{X} = \{z_1, z_2, \ldots, z_{|\mathcal{X}|}\}$). For each $k \in \mathbb{N}$ with $k \le |\mathcal{X}|$, let $A_k = \{z_k\}$; if $|\mathcal{X}| < \infty$, let $A_k = \emptyset$ for all $k \in \mathbb{N}$ with $k > |\mathcal{X}|$. Fix any $\mathbb{X} \in \mathcal{C}_1$. By Lemma 13,

$$\lim_{i\to\infty} \hat{\mu}_{\mathbb{X}}\left(\bigcup_{k\ge i} A_k \right) = 0 \text{ (a.s.).}$$

Combined with Lemma 14, this implies

$$\lim_{n\to\infty} \hat{\mu}_{\mathbb{X}}\Big(\bigcup \{A_i : X_{1:n} \cap A_i = \emptyset\} \Big) = 0 \text{ (a.s.).}$$

From the definition of $\hat{f}_n$, for each $n \in \mathbb{N}$, any $f^\star : \mathcal{X} \to \mathcal{Y}$, and each $z_i \in \mathcal{X}$, if $\hat{f}_n(X_{1:n}, f^\star(X_{1:n}), z_i) \neq f^\star(z_i)$, then necessarily $z_i \notin \{X_1, \ldots, X_n\}$. Therefore,

$$\bigcup \{A_i : X_{1:n} \cap A_i = \emptyset\} = \mathcal{X} \setminus \{X_1, \ldots, X_n\} \supseteq \{z_i : \hat{f}_n(X_{1:n}, f^\star(X_{1:n}), z_i) \neq f^\star(z_i)\}.$$

Combining this with Lemma 8 (for homogeneity and monotonicity of $\hat{\mu}_{\mathbb{X}}$), we have that for any measurable $f^\star : \mathcal{X} \to \mathcal{Y}$,

$$\begin{aligned} \lim_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(\hat{f}_n, f^\star; n) &\le \lim_{n\to\infty} \hat{\mu}_{\mathbb{X}}\Big(\mathbb{1}_{\{x : \hat{f}_n(X_{1:n}, f^\star(X_{1:n}), x) \neq f^\star(x)\}}(\cdot)\bar{\ell} \Big) \\ &= \bar{\ell} \lim_{n\to\infty} \hat{\mu}_{\mathbb{X}}\Big(\{x : \hat{f}_n(X_{1:n}, f^\star(X_{1:n}), x) \neq f^\star(x)\} \Big) \\ &\le \bar{\ell} \lim_{n\to\infty} \hat{\mu}_{\mathbb{X}}\Big(\bigcup \{A_i : X_{1:n} \cap A_i = \emptyset\} \Big) = 0 \text{ (a.s.).} \end{aligned}$$

Thus, since $\hat{\mathcal{L}}_{\mathbb{X}}$ is nonnegative, $\hat{f}_n$ is strongly universally consistent under every $\mathbb{X} \in \mathcal{C}_1$. Recalling that (by Theorem 7) SUIL $= \mathcal{C}_1$, this completes the proof. ∎

It is worth noting here that the proof of Theorem 6 can be made somewhat simpler if we only wish to directly establish the theorem statement. Specifically, the variables $V_{k,j}$ there can be replaced by i.i.d. π_0 samples, while the $U_{k,j}$ variables can all be set equal to some fixed point $x_0 \in \mathcal{X}$; in this case, the sets R_k are not needed (replaced by $\mathcal{X} \setminus \{x_0\}$), and several of the definitions can be simplified (e.g., the $f_k^\star$ functions can all be replaced by a fixed function $f_1^\star$, which simply outputs y_0 except on w_j and x_0 points, where it outputs y_1). The general approach to the proof of inconsistency remains essentially unchanged. One can easily verify that the resulting process does satisfy Condition 1; however, it does not necessarily have convergent relative frequencies (specifically, in the second case discussed in the proof). The details of this simpler proof are left as an exercise for the interested reader. We have chosen the more-involved proof presented above so that the inductive learning rule is shown to not be universally consistent even under processes with convergent relative frequencies. Indeed, the constructed processes are in fact *product* processes, and the limit in (8) is a (non-random) probability measure, described by either (42) or (50). Formally, we have established the following corollary.

Corollary 30 *If $\mathcal{X}$ is uncountable, then there does not exist an inductive learning rule that is strongly universally consistent under every $\mathbb{X} \in$ CRF.*

6. Online Learning

In this section, we discuss the *online learning* setting, establishing a number of results related to the following question (restated from Section 1.2) on the existence of optimistically universal learning rules.

Open Problem 1 (restated) *Does there exist an optimistically universal online learning rule?*

We approach this question and related issues in an analogous fashion to the above discussion of self-adaptive and inductive learning. However, unlike the results on self-adaptive and inductive learning, the results presented here are only partial, and leave open a number of interesting core questions, including the above open problem.

After introducing some useful lemmas on online aggregation techniques in Section 6.1, we begin the discussion of universally consistent online learning in Section 6.2 with the subject of concisely characterizing the family of processes SUOL. We propose a concise condition (Condition 2) for a process $\mathbb{X}$, and prove that it is generally a *necessary* condition: i.e., it is satisfied by any $\mathbb{X}$ that admits strong universal online learning. We also argue that it is a *sufficient* condition in the case that $\mathcal{X}$ is countable or that $\mathbb{X}$ is deterministic, but we leave open the question of whether this condition is a sufficient condition for $\mathbb{X}$ to admit strong universal online learning in the *general* case (Open Problem 2). Following this, in Section 6.3, we address the relation between admission of strong universal online learning and admission of strong universal self-adaptive learning. We specifically establish that the latter implies the former, but not vice versa (when $\mathcal{X}$ is infinite): that is, SUAL $\subset$ SUOL with *strict* inclusion, which establishes a separation of SUOL from SUAL and SUIL. Although lacking a general concise (provable) characterization of SUOL, we are at least able to show, in Section 6.4, that the family SUOL is invariant to the choice of loss function ℓ (as was true of SUIL and SUAL above, from their equivalence to $\mathcal{C}_1$ in Theorem 7), under the additional restriction that ℓ is totally bounded. We also argue that SUOL is invariant to the choice of ℓ among losses that are *not* totally bounded, but we leave open the question of whether these two SUOL families are equal (Open Problem 3).

6.1 Online Aggregation

Before getting into the new results of the present work on online learning, we first introduce some supporting lemmas based on a well-known aggregation technique from the literature on online learning with arbitrary sequences. The first lemma is a regret guarantee for a weighted averaging prediction algorithm. The technique and analysis are taken from classic works in the theory of online learning (Vovk, 1990, 1992; Littlestone and Warmuth, 1994; Cesa-Bianchi, Freund, Haussler, Helmbold, Schapire, and Warmuth, 1997; Kivinen and Warmuth, 1999; Singer and Feder, 1999; Györfi and Lugosi, 2002). For completeness, we include a brief proof: a version of this classic argument.

Lemma 31 *For each $n \in \mathbb{N}$, let $\{z_{n,i}\}_{i=1}^{\infty}$ be a sequence of values in $[0,1]$, and let $\{p_i\}_{i=1}^{\infty}$ be a sequence in $(0,1)$ with $\sum_{i=1}^{\infty} p_i = 1$. Fix a finite constant $b \in (0,1)$. For each $n, i \in \mathbb{N}$, define $L_{n,i} = \frac{1}{n}\sum_{t=1}^{n} z_{t,i}$. Then for each $i \in \mathbb{N}$, define $w_{1,i} = v_{1,i} = p_i$, and for each $n \in \mathbb{N} \setminus \{1\}$, define $w_{n,i} = p_i b^{(n-1)L_{(n-1),i}}$, and $v_{n,i} = w_{n,i} / \sum_{i=1}^{\infty} w_{n,i}$. Finally, for each $n \in \mathbb{N}$, define $\bar{z}_n = \sum_{i=1}^{\infty} v_{n,i} z_{n,i}$. Then for every $n \in \mathbb{N}$,*

$$\frac{1}{n}\sum_{t=1}^{n} \bar{z}_t \leq \inf_{i \in \mathbb{N}} \left(\frac{\ln(1/b)}{1-b} L_{n,i} + \frac{1}{(1-b)n} \ln\left(\frac{1}{p_i}\right) \right).$$

Proof Denote $W_n = \sum_{i=1}^{\infty} w_{n,i}$ for each $n \in \mathbb{N}$. Then note that $\forall n \in \mathbb{N}$, $W_{n+1} = \sum_{i=1}^{\infty} w_{n,i} b^{z_{n,i}} = W_n \sum_{i=1}^{\infty} v_{n,i} b^{z_{n,i}}$. Noting that $b^{z_{n,i}} \leq 1 - (1-b)z_{n,i}$, we find that

$$\frac{W_{n+1}}{W_n} \leq \sum_{i=1}^{\infty} v_{n,i}(1 - (1-b)z_{n,i}) = 1 - (1-b)\bar{z}_n.$$

Since $W_1 = 1$, by induction we have $W_{n+1} \leq \prod_{t=1}^{n}(1-(1-b)\bar{z}_t)$. Noting that $\ln(1-(1-b)\bar{z}_t) \leq -(1-b)\bar{z}_t$, we have that $\ln(W_{n+1}) \leq \sum_{t=1}^{n} \ln(1-(1-b)\bar{z}_t) \leq -(1-b)\sum_{t=1}^{n} \bar{z}_t$. Therefore, for any $n \in \mathbb{N}$,

$$\begin{aligned} \sum_{t=1}^{n} \bar{z}_t &\leq \frac{1}{1-b} \ln\left(\frac{1}{W_{n+1}}\right) = \frac{1}{1-b} \ln\left(\frac{1}{\sum_{i=1}^{\infty} p_i b^{nL_{n,i}}}\right) \\ &\leq \frac{1}{1-b} \ln\left(\frac{1}{\sup_{i \in \mathbb{N}} p_i b^{nL_{n,i}}}\right) = \inf_{i \in \mathbb{N}} \left(\frac{\ln(1/b)}{1-b} nL_{n,i} + \frac{1}{1-b} \ln\left(\frac{1}{p_i}\right) \right). \end{aligned}$$

Dividing the leftmost and rightmost expressions by n completes the proof. ■

For our purposes, we will need the following implication of this lemma.

Lemma 32 *For any sequence* $\left\{\hat{h}_n^{(i)}\right\}_{i=1}^{\infty}$ *of online learning rules, there exists an online learning rule* $\hat{f}_n$ *such that, for any process* $\mathbb{X}$ *and any function* $f^\star : \mathcal{X} \to \mathcal{Y}$*, if, with probability one, there exists a sequence* $\{i_n\}_{n=1}^{\infty}$ *in* $\mathbb{N}$ *with* $\ln(i_n) = o(n)$ *s.t.* $\lim\limits_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(\hat{h}_{\cdot}^{(i_n)}, f^\star; n) = 0$*, then* $\lim\limits_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(\hat{f}_{\cdot}, f^\star; n) = 0$ *(a.s.).*

Proof Fix any sequences $\mathbf{x} = \{x_n\}_{n=1}^{\infty}$ in $\mathcal{X}$ and $\mathbf{y} = \{y_n\}_{n=1}^{\infty}$ in $\mathcal{Y}$. For each $n, i \in \mathbb{N}$, define $\hat{z}_{n,i}(x_{1:n}, y_{1:n}) = \ell(\hat{h}_{n-1}^{(i)}(x_{1:(n-1)}, y_{1:(n-1)}, x_n), y_n)/\bar{\ell}$ (which may be random, if $\hat{h}_{n-1}^{(i)}$ is a randomized learning rule). For each $i \in \mathbb{N}$, let $p_i = \frac{6}{\pi^2 i^2}$, and note that $\sum\limits_{i=1}^{\infty} p_i = 1$. Fix any $b \in (0,1)$, and for $n, i \in \mathbb{N}$ define $v_{n,i}$ as in Lemma 31, for these p_i values, and for $z_{n,i} = \hat{z}_{n,i}(x_{1:n}, y_{1:n}) \in [0,1]$ for each $n, i \in \mathbb{N}$. Finally, define $\bar{z}_n(x_{1:n}, y_{1:n}) = \sum\limits_{i=1}^{\infty} v_{n,i} \hat{z}_{n,i}(x_{1:n}, y_{1:n})$. From this point, there are two possible routes toward defining the online learning rule $\hat{f}_n$, depending on whether we involve randomization. In the simplest definition, when predicting for x_{n+1}, we could simply sample an index i (independently) according to the distribution specified by $\{v_{(n+1),i}\}_{i=1}^{\infty}$, and take the $\hat{h}_n^{(i)}$ learning rules's prediction. It is fairly straightforward to relate the expected performance of this method to the quantities $\bar{z}_t(x_{1:t}, y_{1:t})$ and then apply Lemma 31 (see e.g., Littlestone and Warmuth, 1994), together with concentration inequalities to argue that the bound from Lemma 31 almost surely becomes valid in the limit of $n \to \infty$. However, instead of this approach, we will analyze a method that avoids randomization. Specifically, let $\{\varepsilon_n\}_{n=0}^{\infty}$ be any sequence in $(0, \infty)$ with $\varepsilon_n \to 0$, and for each $n \in \mathbb{N} \cup \{0\}$, define[4]

$$\hat{f}_n(x_{1:n}, y_{1:n}, x_{n+1}) = \operatorname*{argmin}_{y \in \mathcal{Y}}{}^{\varepsilon_n} \sum_{i=1}^{\infty} v_{(n+1),i} \ell\Big(y, \hat{h}_n^{(i)}(x_{1:n}, y_{1:n}, x_{n+1})\Big).$$

We use this definition for any n and any such sequences $\mathbf{x}$ and $\mathbf{y}$, so that this completes the definition of $\hat{f}_n$. With this definition, for any $t \in \mathbb{N} \cup \{0\}$ and sequences $\mathbf{x}$ and $\mathbf{y}$, by the triangle inequality and the fact that $\sum\limits_{i=1}^{\infty} v_{(t+1),i} = 1$, we have that

$$\begin{aligned} &\ell\Big(\hat{f}_t(x_{1:t}, y_{1:t}, x_{t+1}), y_{t+1}\Big) = \ell\Big(\hat{f}_t(x_{1:t}, y_{1:t}, x_{t+1}), y_{t+1}\Big) \sum_{i=1}^{\infty} v_{(t+1),i} \\ &\leq \sum_{i=1}^{\infty} v_{(t+1),i} \ell\Big(\hat{f}_t(x_{1:t}, y_{1:t}, x_{t+1}), \hat{h}_t^{(i)}(x_{1:t}, y_{1:t}, x_{t+1})\Big) + \sum_{i=1}^{\infty} v_{(t+1),i} \ell\Big(\hat{h}_t^{(i)}(x_{1:t}, y_{1:t}, x_{t+1}), y_{t+1}\Big). \end{aligned}$$

4. Here we suppose the $\operatorname*{argmin}_{y \in \mathcal{Y}}^{\varepsilon_n}$ selection is implemented in a way that renders this function measurable. This is clearly always possible in our context. For instance, it would suffice to consider an enumeration of a countable dense subset of $\mathcal{Y}$ (which exists by the separability assumption) and then choose the first y in this enumeration satisfying the ε_n excess criterion in the definition of $\operatorname*{argmin}_{y \in \mathcal{Y}}^{\varepsilon_n}$.

Then the definition of $\hat{f}_t$ guarantees this is at most

$$\begin{aligned}&\varepsilon_t + \inf_{y \in \mathcal{Y}} \sum_{i=1}^{\infty} v_{(t+1),i} \ell\Big(y, \hat{h}_t^{(i)}(x_{1:t}, y_{1:t}, x_{t+1})\Big) + \sum_{i=1}^{\infty} v_{(t+1),i} \ell\Big(\hat{h}_t^{(i)}(x_{1:t}, y_{1:t}, x_{t+1}), y_{t+1}\Big)\\ &\leq \varepsilon_t + 2\sum_{i=1}^{\infty} v_{(t+1),i} \ell\Big(\hat{h}_t^{(i)}(x_{1:t}, y_{1:t}, x_{t+1}), y_{t+1}\Big) = \varepsilon_t + 2\bar{\ell}\bar{z}_{t+1}(x_{1:(t+1)}, y_{1:(t+1)}),\end{aligned}$$

so that

$$\frac{1}{n}\sum_{t=0}^{n-1} \ell\Big(\hat{f}_t(x_{1:t}, y_{1:t}, x_{t+1}), y_{t+1}\Big) \leq \frac{1}{n}\sum_{t=0}^{n-1} \left(\varepsilon_t + 2\bar{\ell}\bar{z}_{t+1}(x_{1:(t+1)}, y_{1:(t+1)})\right).$$

Together with Lemma 31, we have that

$$\begin{aligned}&\frac{1}{n}\sum_{t=0}^{n-1} \ell\Big(\hat{f}_t(x_{1:t}, y_{1:t}, x_{t+1}), y_{t+1}\Big) \qquad (51)\\ &\leq \left(\frac{1}{n}\sum_{t=0}^{n-1}\varepsilon_t\right) + 2\inf_{i \in \mathbb{N}} \left(\frac{\ln(1/b)}{1-b}\left(\frac{1}{n}\sum_{t=0}^{n-1} \ell\Big(\hat{h}_t^{(i)}(x_{1:t}, y_{1:t}, x_{t+1}), y_{t+1}\Big)\right) + \frac{\bar{\ell}}{(1-b)n}\ln\!\left(\frac{1}{p_i}\right)\right).\end{aligned}$$

Now fix $\mathbb{X}$ and $f^\star$ such that, with probability one, there exists a sequence $\{i_n\}_{n=1}^{\infty}$ in $\mathbb{N}$ with $\ln(i_n) = o(n)$ such that $\lim_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(\hat{h}_\cdot^{(i_n)}, f^\star; n) = 0$. Then, on the event that this occurs, the inequality in (51) implies

$$\begin{aligned}\hat{\mathcal{L}}_{\mathbb{X}}(\hat{f}_\cdot, f^\star; n) &\leq \left(\frac{1}{n}\sum_{t=0}^{n-1}\varepsilon_t\right) + 2\inf_{i\in\mathbb{N}}\left(\frac{\ln(1/b)}{1-b}\hat{\mathcal{L}}_{\mathbb{X}}(\hat{h}_\cdot^{(i)}, f^\star; n) + \frac{\bar{\ell}}{(1-b)n}\ln\!\left(\frac{1}{p_i}\right)\right)\\ &\leq \left(\frac{1}{n}\sum_{t=0}^{n-1}\varepsilon_t\right) + 2\left(\frac{\ln(1/b)}{1-b}\hat{\mathcal{L}}_{\mathbb{X}}(\hat{h}_\cdot^{(i_n)}, f^\star; n) + \frac{2\bar{\ell}}{(1-b)n}\ln(i_n) + \frac{\bar{\ell}}{(1-b)n}\ln\!\left(\frac{\pi^2}{6}\right)\right).\end{aligned}$$

Since $\varepsilon_t \to 0$ implies $\lim_{n\to\infty} \frac{1}{n}\sum_{t=0}^{n-1}\varepsilon_t = 0$, and since $\lim_{n\to\infty}\hat{\mathcal{L}}_{\mathbb{X}}(\hat{h}_\cdot^{(i_n)}, f^\star; n) = 0$ and $\lim_{n\to\infty}\frac{1}{n}\ln(i_n) = 0$ in this context, and $\hat{\mathcal{L}}_{\mathbb{X}}$ is nonnegative, it follows that $\lim_{n\to\infty}\hat{\mathcal{L}}_{\mathbb{X}}(\hat{f}_\cdot, f^\star; n) = 0$ on this event. ∎

The next lemma provides a technical fact useful in the proofs of the theorems below.

Lemma 33 *Suppose $\{\beta_{i,n}\}_{i,n\in\mathbb{N}}$ is an array of values in $[0,\infty)$ such that $\lim_{i\to\infty}\limsup_{n\to\infty}\beta_{i,n} = 0$, and that $\{j_n\}_{n=1}^{\infty}$ is a sequence in $\mathbb{N}$ with $j_n \to \infty$. Then there exists a sequence $\{i_n\}_{n=1}^{\infty}$ in $\mathbb{N}$ such that $i_n \leq j_n$ for every $n \in \mathbb{N}$, and $\lim_{n\to\infty}\beta_{i_n,n} = 0$.*

Proof For each $i \in \mathbb{N}$, let $n_i \in \mathbb{N}$ be such that $\sup_{n \geq n_i}\beta_{i,n} \leq \frac{1}{i} + \limsup_{n\to\infty}\beta_{i,n}$; such an n_i is guaranteed to exist by the definition of the lim sup. For each $n \in \mathbb{N}$ with $n < n_1$, define

$i_n = 1$, and for each $n \in \mathbb{N}$ with $n \geq n_1$, define $i_n = \max\{i \in \{1, \ldots, j_n\} : n \geq n_i\}$. By definition, we clearly have $i_n \leq j_n$ for every $n \in \mathbb{N}$. Furthermore, by definition, we have $n \geq n_{i_n}$ for every $n \geq n_1$, so that $\beta_{i_n,n} \leq \frac{1}{i_n} + \limsup_{n' \to \infty} \beta_{i_n,n'}$. Finally, since n_i is finite for each $i \in \mathbb{N}$, and $j_n \to \infty$, we have $i_n \to \infty$. Altogether, we have

$$\limsup_{n \to \infty} \beta_{i_n,n} \leq \limsup_{n \to \infty} \left(\frac{1}{i_n} + \limsup_{n' \to \infty} \beta_{i_n,n'} \right) \leq \limsup_{i \to \infty} \left(\frac{1}{i} + \limsup_{n \to \infty} \beta_{i,n} \right) = 0.$$

Since $\liminf_{n \to \infty} \beta_{i_n,n} \geq 0$ by nonnegativity of the $\beta_{i,n}$ values, the result follows. ∎

6.2 Toward Concisely Characterizing SUOL

We begin the discussion of universally consistent online learning with the subject of concisely characterizing the family of processes SUOL. Specifically, we consider the following candidate for such a characterization. Though we succeed in establishing its necessity, determining whether it is also sufficient will be left as an open problem.

Condition 2 *For every sequence $\{A_k\}_{k=1}^{\infty}$ of disjoint elements of $\mathcal{B}$,*

$$|\{k \in \mathbb{N} : X_{1:T} \cap A_k \neq \emptyset\}| = o(T) \text{ (a.s.)}.$$

Denote by $\mathcal{C}_2$ the set of all processes $\mathbb{X}$ satisfying Condition 2. With the aim of concisely characterizing the family of processes SUOL, we consider now the specific question of whether $\text{SUOL} = \mathcal{C}_2$. Formally, we make partial progress toward resolving the following question, which remains open at this writing.

Open Problem 2 *Is* $\text{SUOL} = \mathcal{C}_2$*?*

In this subsection, we show that in general, $\text{SUOL} \subseteq \mathcal{C}_2$, and that equality holds when $\mathcal{X}$ is countable. Equality also holds for the intersections of these sets with the family of deterministic processes.

We begin with the first of these claims. First, as was true of $\mathcal{C}_1$, we can also state Condition 2 in an alternative equivalent form, which makes the necessity of Condition 2 for learning more immediately clear.

Lemma 34 *A process $\mathbb{X}$ satisfies Condition 2 if and only if, for every sequence $\{A_i\}_{i=1}^{\infty}$ of disjoint elements of $\mathcal{B}$ with $\bigcup_{i=1}^{\infty} A_i = \mathcal{X}$, denoting by $i(x)$ the unique index $i \in \mathbb{N}$ with $x \in A_i$ (for each $x \in \mathcal{X}$),*

$$\limsup_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\left[X_{1:(t-1)} \cap A_{i(X_t)} = \emptyset\right] = 0 \text{ (a.s.)}.$$

Proof First note that, for any sequence $\{A_k\}_{k=1}^{\infty}$ of disjoint sets in $\mathcal{B}$, defining $B_1 = \mathcal{X} \setminus \bigcup_{k=1}^{\infty} A_k$ and $B_k = A_{k-1}$ for integers $k \geq 2$, we have that $\{B_k\}_{k=1}^{\infty}$ is a sequence of disjoint sets in $\mathcal{B}$ with $\bigcup_{k=1}^{\infty} B_k = \mathcal{X}$, and $|\{k : X_{1:T} \cap A_k \neq \emptyset\}| \leq |\{k : X_{1:T} \cap B_k \neq \emptyset\}|$, so that if $|\{k : X_{1:T} \cap B_k \neq \emptyset\}| = o(T)$ (a.s.), then $|\{k : X_{1:T} \cap A_k \neq \emptyset\}| = o(T)$ (a.s.) as well. Thus, the set of processes $\mathbb{X}$ satisfying Condition 2 remains unchanged if we restrict the disjoint sequences $\{A_k\}_{k=1}^{\infty}$ to those satisfying $\bigcup_{k=1}^{\infty} A_k = \mathcal{X}$.

Now fix any process $\mathbb{X}$ and any sequence $\{A_i\}_{i=1}^{\infty}$ of disjoint elements of $\mathcal{B}$ with $\bigcup_{i=1}^{\infty} A_i = \mathcal{X}$, and let $i(x)$ be as in the lemma statement. Then note that, for any $T \in \mathbb{N}$,

$$|\{i \in \mathbb{N} : X_{1:T} \cap A_i \neq \emptyset\}| = \mathbb{1}\left[X_{1:(T-1)} \cap A_{i(X_T)} = \emptyset\right] + \left|\{i \in \mathbb{N} : X_{1:(T-1)} \cap A_i \neq \emptyset\}\right|.$$

By induction (taking $T = 1$ in the above equality for the base case), this implies that $\forall T \in \mathbb{N}$,

$$|\{i \in \mathbb{N} : X_{1:T} \cap A_i \neq \emptyset\}| = \sum_{t=1}^{T} \mathbb{1}\left[X_{1:(t-1)} \cap A_{i(X_t)} = \emptyset\right].$$

In particular, this implies that $\sum_{t=1}^{T} \mathbb{1}\left[X_{1:(t-1)} \cap A_{i(X_t)} = \emptyset\right] = o(T)$ (a.s.) if and only if $|\{i \in \mathbb{N} : X_{1:T} \cap A_i \neq \emptyset\}| = o(T)$ (a.s.). Since this equivalence holds for any choice of disjoint sequence $\{A_i\}_{i=1}^{\infty}$ in $\mathcal{B}$ with $\bigcup_{i=1}^{\infty} A_i = \mathcal{X}$, the lemma follows. ■

With this lemma in hand, we can now prove the following theorem, which establishes that Condition 2 is *necessary* for a process to admit strong universal online learning.

Theorem 35 SUOL $\subseteq \mathcal{C}_2$.

Proof This proof follows essentially the same outline as that of Lemma 19. We prove the result in the contrapositive. Suppose $\mathbb{X} \notin \mathcal{C}_2$. By Lemma 34, there exists a disjoint sequence $\{A_i\}_{i=1}^{\infty}$ in $\mathcal{B}$ with $\bigcup_{i=1}^{\infty} A_i = \mathcal{X}$ such that, letting $i(x)$ denote the unique $i \in \mathbb{N}$ with $x \in A_i$ (for each $x \in \mathcal{X}$), we have, with probability strictly greater than 0,

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\left[X_{1:(t-1)} \cap A_{i(X_t)} = \emptyset\right] > 0.$$

Furthermore, since the left hand side is always nonnegative, this also implies (see e.g., Ash and Doléans-Dade, 2000, Theorem 1.6.6)

$$\mathbb{E}\left[\limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\left[X_{1:(t-1)} \cap A_{i(X_t)} = \emptyset\right]\right] > 0. \tag{52}$$

Now take any two distinct values $y_0, y_1 \in \mathcal{Y}$, and (as we did in the proof of Lemma 19) for each $\kappa \in [0,1)$, $i \in \mathbb{N}$, and $x \in A_i$, denoting $\kappa_i = \lfloor 2^i \kappa \rfloor - 2\lfloor 2^{i-1}\kappa \rfloor \in \{0,1\}$, define $f_{\kappa}^{\star}(x) = y_{\kappa_i}$. Also denote $i_t = i(X_t)$ for every $t \in \mathbb{N}$, and for any $n \in \mathbb{N} \cup \{0\}$, let $\bar{\mathcal{A}}(X_{1:n}) = \bigcup\{A_i : X_{1:n} \cap A_i = \emptyset\}$.

Now fix any online learning rule g_n, and for brevity define $f_n^{\kappa}(\cdot) = g_n(X_{1:n}, f_\kappa^\star(X_{1:n}), \cdot)$ for each $n \in \mathbb{N}$. Then

$$\sup_{\kappa\in[0,1)} \mathbb{E}\left[\limsup_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(g_\cdot, f_\kappa^\star; n)\right] \geq \int_0^1 \mathbb{E}\left[\limsup_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(g_\cdot, f_\kappa^\star; n)\right] \mathrm{d}\kappa$$
$$\geq \int_0^1 \mathbb{E}\left[\limsup_{n\to\infty} \frac{1}{n}\sum_{t=0}^{n-1} \ell\left(f_t^\kappa(X_{t+1}), f_\kappa^\star(X_{t+1})\right) \mathbb{1}_{\bar{\mathcal{A}}(X_{1:t})}(X_{t+1})\right] \mathrm{d}\kappa.$$

By Fubini's theorem, this is equal

$$\mathbb{E}\left[\int_0^1 \limsup_{n\to\infty} \frac{1}{n}\sum_{t=0}^{n-1} \ell(f_t^\kappa(X_{t+1}), f_\kappa^\star(X_{t+1})) \mathbb{1}_{\bar{\mathcal{A}}(X_{1:t})}(X_{t+1})\mathrm{d}\kappa\right].$$

Since ℓ is bounded, Fatou's lemma implies this is at least as large as

$$\mathbb{E}\left[\limsup_{n\to\infty} \int_0^1 \frac{1}{n}\sum_{t=0}^{n-1} \ell(f_t^\kappa(X_{t+1}), f_\kappa^\star(X_{t+1})) \mathbb{1}_{\bar{\mathcal{A}}(X_{1:t})}(X_{t+1})\mathrm{d}\kappa\right],$$

and linearity of integration implies this equals

$$\mathbb{E}\left[\limsup_{n\to\infty} \frac{1}{n}\sum_{t=0}^{n-1} \mathbb{1}_{\bar{\mathcal{A}}(X_{1:t})}(X_{t+1}) \int_0^1 \ell(f_t^\kappa(X_{t+1}), f_\kappa^\star(X_{t+1}))\,\mathrm{d}\kappa\right]. \tag{53}$$

For any $t \in \mathbb{N}\cup\{0\}$, the value of $f_t^\kappa(X_{t+1})$ is a function of $\mathbb{X}$ and $\kappa_{i_1}, \ldots, \kappa_{i_t}$. Therefore, for any $t \in \mathbb{N}\cup\{0\}$ with $X_{t+1} \in \bar{\mathcal{A}}(X_{1:t})$, the value of $f_t^\kappa(X_{t+1})$ is functionally independent of $\kappa_{i_{t+1}}$. Thus, for any $t \in \mathbb{N}\cup\{0\}$, letting $K \sim \mathrm{Uniform}([0,1))$ be independent of $\mathbb{X}$ and g_t, if $X_{t+1} \in \bar{\mathcal{A}}(X_{1:t})$, we have

$$\int_0^1 \ell(f_t^\kappa(X_{t+1}), f_\kappa^\star(X_{t+1}))\,\mathrm{d}\kappa = \mathbb{E}\left[\ell\left(f_t^K(X_{t+1}), f_K^\star(X_{t+1})\right)\middle|\mathbb{X}, g_t\right]$$
$$= \mathbb{E}\left[\mathbb{E}\left[\ell\left(g_t(X_{1:t}, \{y_{K_{i_j}}\}_{j=1}^t, X_{t+1}), y_{K_{t+1}}\right)\middle|\mathbb{X}, g_t, K_{i_1}, \ldots, K_{i_t}\right]\middle|\mathbb{X}, g_t\right]$$
$$= \mathbb{E}\left[\sum_{b\in\{0,1\}} \frac{1}{2}\ell\left(g_t(X_{1:t}, \{y_{K_{i_j}}\}_{j=1}^t, X_{t+1}), y_b\right)\middle|\mathbb{X}, g_t\right].$$

By the triangle inequality, this is no smaller than $\mathbb{E}\left[\frac{1}{2}\ell(y_0, y_1)\middle|\mathbb{X}, g_t\right] = \frac{1}{2}\ell(y_0, y_1)$, so that (53) is at least as large as

$$\mathbb{E}\left[\limsup_{n\to\infty} \frac{1}{n}\sum_{t=0}^{n-1} \mathbb{1}_{\bar{\mathcal{A}}(X_{1:t})}(X_{t+1})\frac{1}{2}\ell(y_0, y_1)\right]$$
$$= \frac{1}{2}\ell(y_0, y_1)\mathbb{E}\left[\limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \mathbb{1}\left[X_{1:(t-1)} \cap A_{i(X_t)} = \emptyset\right]\right] > 0,$$

where this last inequality is immediate from (52) and the fact that (since ℓ is a metric) $\ell(y_0, y_1) > 0$. Altogether, we have that

$$\sup_{\kappa \in [0,1)} \mathbb{E}\left[\limsup_{n \to \infty} \hat{\mathcal{L}}_{\mathbb{X}}(g_\cdot, f_\kappa^\star; n)\right] > 0.$$

In particular, this implies $\exists \kappa \in [0, 1)$ such that $\mathbb{E}\left[\limsup_{n \to \infty} \hat{\mathcal{L}}_{\mathbb{X}}(g_\cdot, f_\kappa^\star; n)\right] > 0$. Since any random variable equal 0 (a.s.) necessarily has expected value 0, this further implies that with probability strictly greater than 0, $\limsup_{n \to \infty} \hat{\mathcal{L}}_{\mathbb{X}}(g_\cdot, f_\kappa^\star; n) > 0$. Thus, g_n is not strongly universally consistent. Since g_n was an arbitrary online learning rule, we conclude that there does not exist an online learning rule that is strongly universally consistent under $\mathbb{X}$: that is, $\mathbb{X} \notin$ SUOL. Since this argument holds for any $\mathbb{X} \notin \mathcal{C}_2$, the theorem follows. ∎

Although this work falls short of establishing equivalence between SUOL and $\mathcal{C}_2$ in the general case (i.e., positively resolving Open Problem 2 in general), we do show this equivalence in the special case of *countable* $\mathcal{X}$, and indeed also positively resolve Open Problem 1 for countable $\mathcal{X}$ in the process. Note that, in this special case, Condition 2 simplifies to the condition that the number of distinct points $x \in \mathcal{X}$ occurring in the sequence $X_{1:T}$ is $o(T)$ almost surely. Specifically, we have the following result.

Theorem 36 *If $\mathcal{X}$ is countable, then Condition 2 is necessary and sufficient for a process $\mathbb{X}$ to admit strong universal online learning: that is,* SUOL $= \mathcal{C}_2$*. Moreover, if $\mathcal{X}$ is countable, then there exists an optimistically universal online learning rule.*

Proof Suppose $\mathcal{X}$ is countable. For the first claim, since we already know SUOL $\subseteq \mathcal{C}_2$ from Theorem 35, it suffices to show $\mathcal{C}_2 \subseteq$ SUOL, for this special case. We will establish this fact, while simultaneously establishing the second claim, by showing that there is an online learning rule that is strongly universally consistent under every $\mathbb{X} \in \mathcal{C}_2$ (which thereby also establishes that every such process is in SUOL). Toward this end, fix any $y_0 \in \mathcal{Y}$, and define an online learning rule f_n such that, for each $n \in \mathbb{N} \cup \{0\}$, $\forall x_{1:(n+1)} \in \mathcal{X}^{n+1}$, $\forall y_{1:n} \in \mathcal{Y}^n$, if $x_{n+1} = x_i$ for some $i \in \{1, \ldots, n\}$, then $f_n(x_{1:n}, y_{1:n}, x_{n+1}) = y_i$ for the smallest $i \in \{1, \ldots, n\}$ with $x_{n+1} = x_i$, and otherwise $f_n(x_{1:n}, y_{1:n}, x_{n+1}) = y_0$. The key property of f_n here is that it is *memorization-based*, in that any previously-observed point's response y will be faithfully reproduced if that point is encountered again later in the sequence. The specific fact that it evaluates to y_0 in the case of a previously-unseen point is unimportant in this context, and this case can in fact be defined arbitrarily (subject to the function f_n being measurable) without affecting the result (and similarly for the choice to break ties to favor smaller indices).

Now fix any $\mathbb{X} \in \mathcal{C}_2$ and any measurable function $f^\star : \mathcal{X} \to \mathcal{Y}$. Note that any $i, t \in \mathbb{N}$ with $i \le t$ and $X_{t+1} = X_i$ has $f^\star(X_{t+1}) = f^\star(X_i)$, so that $\ell(f_t(X_{1:t}, f^\star(X_{1:t}), X_{t+1}), f^\star(X_{t+1})) =$

$\ell(f^\star(X_i), f^\star(X_{t+1})) = 0$. Therefore, we have

$$\limsup_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(f_\cdot, f^\star; n) = \limsup_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \ell(f_t(X_{1:t}, f^\star(X_{1:t}), X_{t+1}), f^\star(X_{t+1}))$$
$$\leq \limsup_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \bar{\ell} \mathbb{1}[\nexists i \in \{1, \ldots, t\} : X_{t+1} = X_i] = \limsup_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \bar{\ell} \mathbb{1}[X_{1:(t-1)} \cap \{X_t\} = \emptyset]. \tag{54}$$

Since $\mathcal{X}$ is countable, we can enumerate its elements as $z_1, z_2, \ldots$ (or $z_1, \ldots, z_{|\mathcal{X}|}$, in the case of finite $|\mathcal{X}|$). Then let $A_i = \{z_i\}$ for each $i \in \mathbb{N}$ with $i \leq |\mathcal{X}|$, and if $|\mathcal{X}| < \infty$ then let $A_i = \emptyset$ for every $i > |\mathcal{X}|$. Note that $\{A_i\}_{i=1}^{\infty}$ is a sequence of disjoint elements of $\mathcal{B}$ with $\bigcup_{i=1}^{\infty} A_i = \mathcal{X}$. Furthermore, letting $i(x)$ be defined as in Lemma 34, we have $A_{i(x)} = \{x\}$ for each $x \in \mathcal{X}$. Therefore, since $\mathbb{X} \in \mathcal{C}_2$, Lemma 34 implies that the rightmost expression in (54) equals 0 almost surely. Since this argument holds for any choice of $f^\star$, we conclude that f_n is strongly universally consistent under $\mathbb{X}$. Furthermore, since this holds for any choice of $\mathbb{X} \in \mathcal{C}_2$, the theorem follows. ■

The following corollary on deterministic processes is also implied, via a reduction to the case of countable $\mathcal{X}$.

Corollary 37 *For any deterministic process* $\mathbb{X}$, *Condition 2 is necessary and sufficient for* $\mathbb{X}$ *to admit strong universal online learning: that is,* $\mathbb{X} \in \mathrm{SUOL}$ *if and only if* $\mathbb{X} \in \mathcal{C}_2$.

Proof Sketch This result follows from essentially the same proof used for Theorem 36, except using the distinct entries of the sequence $\mathbb{X}$ as the z_i sequence there (and taking the rest of the space, not occurring in the sequence, as a separate irrelevant A_i set in the application of Lemma 34 there). Alternatively, it can also be established via a reduction to the case of countable $\mathcal{X}$. Specifically, fix any deterministic process $\mathbb{X}$, and let $\mathcal{X}_{\mathbb{X}}$ denote the set of *distinct* points $x \in \mathcal{X}$ appearing in the sequence $\mathbb{X}$. Note that $\mathcal{X}_{\mathbb{X}}$ is countable, and that (with a slight abuse of notation) $\mathbb{X}$ may be thought of as a sequence of $\mathcal{X}_{\mathbb{X}}$-valued random variables. Furthermore, it is straightforward to show that $\mathbb{X}$ satisfies Condition 2 for the space $\mathcal{X}_{\mathbb{X}}$ if and only if it satisfies Condition 2 for the original space $\mathcal{X}$ (since only the intersections of the sets A_i with $\mathcal{X}_{\mathbb{X}}$ are relevant for checking this condition). Thus, since Theorem 36 holds for *any* countable space $\mathcal{X}$, applying it to the space $\mathcal{X}_{\mathbb{X}}$, we have that $\mathbb{X}$ admits strong universal online learning if and only if $\mathbb{X}$ satisfies Condition 2. ■

6.3 Relation of Online Learning to Inductive and Self-Adaptive Learning

Next, we turn to addressing the relation between admission of strong universal online learning and admission of strong universal inductive or self-adaptive learning. Specifically, we find that the latter implies the former, but *not* vice versa (if $\mathcal{X}$ is infinite), so that admission of strong universal online learning is a strictly more general condition. To show this, since we have established in Theorem 7 that SUIL = SUAL, it suffices to argue that

SUAL $\subseteq$ SUOL, with *strict* inclusion if $|\mathcal{X}| = \infty$: that is, SUOL $\setminus$ SUAL $\neq \emptyset$. For this we have the following theorem.

Theorem 38 SUAL $\subseteq$ SUOL*, and the inclusion is* strict *iff* $|\mathcal{X}| = \infty$*.*

Proof We begin by showing SUAL $\subseteq$ SUOL. In fact, we will establish a stronger claim: that there exists a *single* online learning rule $\hat{f}_n$ that is strongly universally consistent for *every* $\mathbb{X} \in$ SUAL. Specifically, let $\hat{g}_{n,m}$ be an optimistically universal self-adaptive learning rule. The existence of such a rule was established in Theorem 5, and an explicit construction is given in (29), as established by Theorem 27. Now fix any $y_0 \in \mathcal{Y}$, and for each $i \in \mathbb{N}$ define an online learning rule $\hat{h}_n^{(i)}$ as follows. For each $n \in \mathbb{N} \cup \{0\}$, for any sequences $x_{1:(n+1)} \in \mathcal{X}^{n+1}$ and $y_{1:n} \in \mathcal{Y}^n$, if $n < i$, then define $\hat{h}_n^{(i)}(x_{1:n}, y_{1:n}, x_{n+1}) = y_0$, and if $n \geq i$, then define $\hat{h}_n^{(i)}(x_{1:n}, y_{1:n}, x_{n+1}) = \hat{g}_{i,n}(x_{1:n}, y_{1:i}, x_{n+1})$. It is easy to verify that measurability $\hat{h}_n^{(i)}$ follows from measurability of $\hat{g}_{i,n}$, so that this is a valid definition of an online learning rule.

Given this definition of the sequence $\left\{\hat{h}_n^{(i)}\right\}_{i=1}^{\infty}$, denote by $\hat{f}_n$ the online learning rule guaranteed to exist by Lemma 32 (defined explicitly in the proof above), satisfying the property described there relative to this sequence $\left\{\hat{h}_n^{(i)}\right\}_{i=1}^{\infty}$. Now fix any $\mathbb{X} \in$ SUAL and any measurable $f^\star : \mathcal{X} \to \mathcal{Y}$, and for each $i, n \in \mathbb{N}$, denote $\hat{\beta}_{i,n} = \hat{\mathcal{L}}_{\mathbb{X}}(\hat{h}_\cdot^{(i)}, f^\star; n)$. In particular, note that since ℓ is always finite, it holds that $\forall i \in \mathbb{N}$,

$$\begin{aligned}
\limsup_{n\to\infty} \hat{\beta}_{i,n} &= \limsup_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \ell\Big(\hat{h}_t^{(i)}(X_{1:t}, f^\star(X_{1:t}), X_{t+1}), f^\star(X_{t+1})\Big) \\
&= \limsup_{n\to\infty} \frac{1}{n} \sum_{t=i}^{n-1} \ell(\hat{g}_{i,t}(X_{1:t}, f^\star(X_{1:i}), X_{t+1}), f^\star(X_{t+1})) \\
&= \limsup_{n\to\infty} \frac{1}{n+1} \sum_{t=i}^{i+n} \ell(\hat{g}_{i,t}(X_{1:t}, f^\star(X_{1:i}), X_{t+1}), f^\star(X_{t+1})) = \hat{\mathcal{L}}_{\mathbb{X}}(\hat{g}_{i,\cdot}, f^\star; i).
\end{aligned}$$

Since $\hat{g}_{n,m}$ is strongly universally consistent under $\mathbb{X}$, it follows that $\lim_{i\to\infty} \limsup_{n\to\infty} \hat{\beta}_{i,n} = 0$ on some event E of probability one. In particular, on E, Lemma 33 implies that there exists a sequence $\{i_n\}_{i=1}^{\infty}$ in $\mathbb{N}$ with $i_n \leq n$ for every n, such that $\lim_{n\to\infty} \hat{\beta}_{i_n,n} = 0$. Therefore, since $\ln(i_n) \leq \ln(n) = o(n)$, the property of $\hat{f}_n$ guaranteed by Lemma 32 implies that $\lim_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(\hat{f}_\cdot, f^\star; n) = 0$ almost surely. Since this argument holds for any choice of $f^\star$, we conclude that $\hat{f}_n$ is strongly universally consistent under $\mathbb{X}$, and since this holds for any choice of $\mathbb{X} \in$ SUAL, it follows that SUAL $\subseteq$ SUOL.

SUOL and SUAL are trivially equal if $|\mathcal{X}| < \infty$, since then *every* process $\mathbb{X}$ is contained in $\mathcal{C}_1$, and Theorem 7 implies SUAL $= \mathcal{C}_1$, while we have just established that SUOL $\supseteq$ SUAL, so every process is contained in both SUAL and SUOL. Now consider the case $|\mathcal{X}| = \infty$. To see that SUOL $\setminus$ SUAL $\neq \emptyset$ in this case, in light of Corollary 37, together with Theorem 7, it suffices to construct a *deterministic* process in $\mathcal{C}_2 \setminus \mathcal{C}_1$. Toward this end, we let $\{z_i\}_{i=1}^{\infty}$ be an arbitrary sequence of distinct elements of $\mathcal{X}$, and define a deterministic process $\mathbb{X}$ as follows. For each $t \in \mathbb{N}$, define $i_t = \lfloor \log_2(2t) \rfloor$, and let $X_t = z_{i_t}$. For any sequence $\{A_k\}_{k=1}^{\infty}$ of disjoint elements of $\mathcal{B}$, and any $T \in \mathbb{N}$,

$$|\{k \in \mathbb{N} : X_{1:T} \cap A_k \neq \emptyset\}| \leq |\{i \in \mathbb{N} : X_{1:T} \cap \{z_i\} \neq \emptyset\}| = \lfloor \log_2(2T) \rfloor = o(T).$$

Therefore, $\mathbb{X} \in \mathcal{C}_2$. However, let $\{A_k\}_{k=1}^{\infty}$ denote any sequence of disjoint subsets of $\{z_i : i \in \mathbb{N}\}$, with $|A_k| = \infty$ for all $k \in \mathbb{N}$ (e.g., $A_k = \{z_{p_k^m} : m \in \mathbb{N}\}$, where p_k is the k^{th} prime number). Then note that every $i \in \mathbb{N}$ has $\frac{1}{2^i - 1} \sum_{t=1}^{2^i - 1} \mathbb{1}_{\{z_i\}}(X_t) = \frac{2^{i-1}}{2^i - 1} > \frac{1}{2}$. Thus, since each $|A_k| = \infty$, we have $\hat{\mu}_{\mathbb{X}}(A_k) \geq \frac{1}{2}$ for every $k \in \mathbb{N}$. But this implies $\forall i \in \mathbb{N}$, $\hat{\mu}_{\mathbb{X}}\left(\bigcup_{k \geq i} A_k\right) \geq \hat{\mu}_{\mathbb{X}}(A_i) \geq \frac{1}{2}$, so that $\lim_{i \to \infty} \hat{\mu}_{\mathbb{X}}\left(\bigcup_{k \geq i} A_k\right) \geq \frac{1}{2} > 0$. Together with Lemma 13, this implies $\mathbb{X} \notin \mathcal{C}_1$. ■

The proof of Theorem 38 actually establishes two additional results. First, since the online learning rule $\hat{f}_n$ constructed in the proof has no dependence on the distribution of the process $\mathbb{X}$ from SUAL, this proof also establishes the following corollary.

Corollary 39 *There exists an online learning rule that is strongly universally consistent under every* $\mathbb{X} \in$ SUAL*.*

Note that this is a weaker claim than would be required for positive resolution of Open Problem 1, since (as established by Theorem 38) the set of processes admitting strong universal online learning is a *strict* superset of the set of processes admitting strong universal self-adaptive learning (if $\mathcal{X}$ is infinite).

Second, since Theorem 35 establishes that SUOL $\subseteq \mathcal{C}_2$, and Theorem 7 establishes that SUAL $= \mathcal{C}_1$, Theorem 38 clearly also establishes that $\mathcal{C}_1 \subseteq \mathcal{C}_2$ (a fact that one can easily verify from their definitions as well). Furthermore, the above proof that the inclusion SUAL $\subseteq$ SUOL is strict if $|\mathcal{X}| = \infty$ establishes this fact by constructing a deterministic process $\mathbb{X} \in \mathcal{C}_2 \setminus \mathcal{C}_1$ (which thereby verifies the claim due to Corollary 37 and Theorem 7). Thus, it also establishes that the inclusion $\mathcal{C}_1 \subseteq \mathcal{C}_2$ is strict in the case $|\mathcal{X}| = \infty$. Also, as noted in the above proof, if $|\mathcal{X}| < \infty$, then $\mathcal{C}_1$ contains *every* process. Since $\mathcal{C}_1 \subseteq \mathcal{C}_2$, this clearly implies that if $|\mathcal{X}| < \infty$, then $\mathcal{C}_1 = \mathcal{C}_2$. Thus, the above proof also establishes the following result.

Corollary 40 $\mathcal{C}_1 \subseteq \mathcal{C}_2$*, and the inclusion is* strict *iff* $|\mathcal{X}| = \infty$*.*

6.4 Invariance of SUOL to the Choice of Loss Function

In this subsection, we are interested in the question of whether the family SUOL is invariant to the choice of loss function (subject to the basic constraints from Section 1.1). Recall that we established above that this property holds for the families SUIL and SUAL (as implied by their equivalence to $\mathcal{C}_1$ from Theorem 7, regardless of the choice of $(\mathcal{Y}, \ell)$). Furthermore, a positive resolution of Open Problem 2 would immediately imply this property for SUOL, since Condition 2 has no dependence on $(\mathcal{Y}, \ell)$. However, since Open Problem 2 remains open at this time, it is interesting to directly explore the question of invariance of SUOL to the choice of $(\mathcal{Y}, \ell)$. Specifically, we prove two relevant results. First, we show that SUOL is invariant to the choice of $(\mathcal{Y}, \ell)$, under the additional constraint that $(\mathcal{Y}, \ell)$ is *totally bounded*: that is, $\forall \varepsilon > 0$, $\exists \mathcal{Y}_\varepsilon \subseteq \mathcal{Y}$ s.t. $|\mathcal{Y}_\varepsilon| < \infty$ and $\sup_{y \in \mathcal{Y}} \inf_{y_\varepsilon \in \mathcal{Y}_\varepsilon} \ell(y_\varepsilon, y) \leq \varepsilon$. For instance, ℓ as the absolute loss with $\mathcal{Y}$ any bounded subset of $\mathbb{R}$ would satisfy this. In particular, this means that, in characterizing the family of processes SUOL for totally bounded losses, it suffices to characterize this set for the simplest case of *binary classification*: $(\mathcal{Y}, \ell) = (\{0,1\}, \ell_{01})$, where for any $\mathcal{Y}$ we generally denote by $\ell_{01} : \mathcal{Y}^2 \to [0, \infty)$ the 0-1 loss on $\mathcal{Y}$, defined by $\ell_{01}(y, y') = \mathbb{1}[y \neq y']$ for all $y, y' \in \mathcal{Y}$. Second, we also find that the set SUOL is invariant among (bounded, separable) losses that are *not* totally bounded (e.g., the 0-1 loss with $\mathcal{Y} = \mathbb{N}$). We leave open the question of whether or not these two SUOL sets are equal (Open Problem 3 below). We begin with the totally bounded case.

Theorem 41 *The set* SUOL *is invariant to the specification of* $(\mathcal{Y}, \ell)$*, subject to being totally bounded with* $\bar{\ell} > 0$*.*

Proof To disambiguate notation in this proof, for any metric space $(\mathcal{Y}', \ell')$, we denote by $\mathrm{SUOL}_{(\mathcal{Y}', \ell')}$ the family SUOL as it would be defined if $(\mathcal{Y}, \ell)$ were specified as $(\mathcal{Y}', \ell')$. As above, we define the measurable subsets of $\mathcal{Y}'$ as the elements of the Borel σ-algebra generated by the topology induced by ℓ'. Let ℓ_{01} be the 0-1 loss on $\{0,1\}$, as defined above. To establish the theorem, it suffices to verify the claim that $\mathrm{SUOL}_{(\mathcal{Y}', \ell')} = \mathrm{SUOL}_{(\{0,1\}, \ell_{01})}$ for all totally bounded metric spaces $(\mathcal{Y}', \ell')$ with $\sup_{y, y' \in \mathcal{Y}'} \ell'(y, y') > 0$. Fix any such $(\mathcal{Y}', \ell')$.

The inclusion $\mathrm{SUOL}_{(\mathcal{Y}', \ell')} \subseteq \mathrm{SUOL}_{(\{0,1\}, \ell_{01})}$ is quite straightforward, as follows. For any $\mathbb{X} \in \mathrm{SUOL}_{(\mathcal{Y}', \ell')}$, letting $\hat{f}_n$ be an online learning rule that is strongly universally consistent under $\mathbb{X}$ (for the specification $(\mathcal{Y}, \ell) = (\mathcal{Y}', \ell')$), we can define an online learning rule $\hat{f}_n^{01}$ for the specification $(\mathcal{Y}, \ell) = (\{0,1\}, \ell_{01})$ as follows. Let $z_0, z_1 \in \mathcal{Y}'$ be such that $\ell'(z_0, z_1) > 0$. For any $n \in \mathbb{N} \cup \{0\}$, and any sequences $x_{1:(n+1)}$ in $\mathcal{X}$ and $y_{1:n}$ in $\{0,1\}$, define a sequence $y'_{1:n}$ with $y'_i = z_{y_i}$ for each $i \in \{1, \ldots, n\}$, and then define $\hat{f}_n^{01}(x_{1:n}, y_{1:n}, x_{n+1}) = \operatorname*{argmin}_{y \in \{0,1\}} \ell'\Big(\hat{f}_n(x_{1:n}, y'_{1:n}, x_{n+1}), z_y\Big)$ (breaking ties in favor of $y = 0$). In particular, that $\hat{f}_n^{01}$ is a measurable function $\mathcal{X}^n \times \{0,1\}^n \times \mathcal{X} \to \{0,1\}$ follows immediately from measurability of $\hat{f}_n$. Then note that, for any measurable function $f : \mathcal{X} \to \{0,1\}$, defining $f' : \mathcal{X} \to \mathcal{Y}'$ as $f'(x) = z_{f(x)}$ (which is clearly also measurable), we have $\forall t \in \mathbb{N} \cup \{0\}$,

$$
\begin{aligned}
&\mathbb{1}\Big[\hat{f}_t^{01}(X_{1:t}, f(X_{1:t}), X_{t+1}) \neq f(X_{t+1})\Big] \\
&\leq \mathbb{1}\left[\ell'\Big(\hat{f}_t(X_{1:t}, f'(X_{1:t}), X_{t+1}), f'(X_{t+1})\Big) = \max_{y \in \{0,1\}} \ell'\Big(\hat{f}_t(X_{1:t}, f'(X_{1:t}), X_{t+1}), z_y\Big)\right] \\
&\leq \mathbb{1}\left[\ell'\Big(\hat{f}_t(X_{1:t}, f'(X_{1:t}), X_{t+1}), f'(X_{t+1})\Big) \geq \sum_{y \in \{0,1\}} \frac{1}{2} \ell'\Big(\hat{f}_t(X_{1:t}, f'(X_{1:t}), X_{t+1}), z_y\Big)\right] \\
&\leq \mathbb{1}\left[\ell'\Big(\hat{f}_t(X_{1:t}, f'(X_{1:t}), X_{t+1}), f'(X_{t+1})\Big) \geq \frac{1}{2} \ell'(z_0, z_1)\right] \\
&\leq \frac{2}{\ell'(z_0, z_1)} \ell'\Big(\hat{f}_t(X_{1:t}, f'(X_{1:t}), X_{t+1}), f'(X_{t+1})\Big),
\end{aligned}
$$

where the second-to-last inequality is due to the triangle inequality. Therefore, under the specification $(\mathcal{Y},\ell) = (\{0,1\},\ell_{01})$, we have

$$\limsup_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}\Big(\hat{f}^{01}_{\cdot}, f; n\Big) = \limsup_{n\to\infty} \frac{1}{n}\sum_{t=0}^{n-1} \mathbb{1}\Big[\hat{f}^{01}_t(X_{1:t}, f(X_{1:t}), X_{t+1}) \neq f(X_{t+1})\Big]$$
$$\leq \frac{2}{\ell'(z_0,z_1)} \limsup_{n\to\infty} \frac{1}{n}\sum_{t=0}^{n-1} \ell'\Big(\hat{f}_t(X_{1:t}, f'(X_{1:t}), X_{t+1}), f'(X_{t+1})\Big) = 0 \text{ (a.s.)},$$

where the last equality (to which the "almost surely" qualifier applies) is due to strong universal consistency of $\hat{f}_n$ (and the fact that z_0, z_1 were chosen to satisfy $\ell'(z_0,z_1) > 0$). Since this argument holds for any choice of measurable $f : \mathcal{X} \to \{0,1\}$, we conclude that $\hat{f}^{01}_n$ is strongly universally consistent under $\mathbb{X}$ (for the specification $(\mathcal{Y},\ell) = (\{0,1\},\ell_{01})$), so that $\mathbb{X} \in \mathrm{SUOL}_{(\{0,1\},\ell_{01})}$. Since this argument holds for any $\mathbb{X} \in \mathrm{SUOL}_{(\mathcal{Y}',\ell')}$, we conclude that $\mathrm{SUOL}_{(\mathcal{Y}',\ell')} \subseteq \mathrm{SUOL}_{(\{0,1\},\ell_{01})}$.

The proof of the converse inclusion is somewhat more involved. Specifically, fix any $\mathbb{X} \in \mathrm{SUOL}_{(\{0,1\},\ell_{01})}$, and let $\hat{f}^{01}_n$ be an online learning rule that is strongly universally consistent under $\mathbb{X}$ (for the specification $(\mathcal{Y},\ell) = (\{0,1\},\ell_{01})$). We then define an online learning rule $\hat{f}'_n$ for the specification $(\mathcal{Y},\ell) = (\mathcal{Y}',\ell')$ as follows. For each $\varepsilon > 0$, let $\mathcal{Y}'_\varepsilon \subseteq \mathcal{Y}'$ be such that $|\mathcal{Y}'_\varepsilon| < \infty$ and $\sup\limits_{y\in\mathcal{Y}'} \inf\limits_{y_\varepsilon\in\mathcal{Y}'_\varepsilon} \ell'(y_\varepsilon, y) \leq \varepsilon$, as guaranteed to exist by total boundedness. For each $y \in \mathcal{Y}'$, let $g_\varepsilon(y) = \operatorname*{argmin}\limits_{y_\varepsilon\in\mathcal{Y}'_\varepsilon} \ell'(y_\varepsilon, y)$, breaking ties to favor smaller indices in some fixed enumeration of $\mathcal{Y}'_\varepsilon$. Then, for each $y \in \mathcal{Y}'$ and each $y_\varepsilon \in \mathcal{Y}'_\varepsilon$, define $h^{(y_\varepsilon)}_\varepsilon(y) = \mathbb{1}[g_\varepsilon(y) = y_\varepsilon]$. One can easily verify that g_ε and $h^{(y_\varepsilon)}_\varepsilon$ are measurable functions, and furthermore that for every $y \in \mathcal{Y}'$, exactly one $y_\varepsilon \in \mathcal{Y}'_\varepsilon$ has $h^{(y_\varepsilon)}_\varepsilon(y) = 1$ while every $y'_\varepsilon \in \mathcal{Y}'_\varepsilon \setminus \{y_\varepsilon\}$ has $h^{(y'_\varepsilon)}_\varepsilon(y) = 0$.

For any $n \in \mathbb{N}\cup\{0\}$, and any sequences $x_{1:(n+1)}$ in $\mathcal{X}$ and $y_{1:n}$ in $\mathcal{Y}'$, define

$$\hat{f}^{(\varepsilon)}_n(x_{1:n}, y_{1:n}, x_{n+1}) = \operatorname*{argmax}_{y_\varepsilon\in\mathcal{Y}'_\varepsilon} \hat{f}^{01}_n(x_{1:n}, h^{(y_\varepsilon)}_\varepsilon(y_{1:n}), x_{n+1}),$$

breaking ties to favor y_ε with a smaller index in a fixed enumeration of $\mathcal{Y}'_\varepsilon$. Again, one can easily verify that $\hat{f}_n$ is a measurable function $\mathcal{X}^n \times (\mathcal{Y}')^n \times \mathcal{X} \to \mathcal{Y}'$, which follows immediately from measurability of $\hat{f}^{01}_n$, the $h^{(y_\varepsilon)}_\varepsilon$ functions, and the argmax. Thus, $\hat{f}^{(\varepsilon)}_n$ defines an online learning rule.

Now note that, for any measurable function $f : \mathcal{X} \to \mathcal{Y}'$, and each $y_\varepsilon \in \mathcal{Y}'_\varepsilon$, the composed function $x \mapsto h^{(y_\varepsilon)}_\varepsilon(f(x))$ is a measurable function $\mathcal{X} \to \{0,1\}$, and therefore (by strong universal consistency of $\hat{f}^{01}_n$) with probability one,

$$\limsup_{n\to\infty} \frac{1}{n}\sum_{t=0}^{n-1} \ell_{01}\Big(\hat{f}^{01}_t(X_{1:t}, h^{(y_\varepsilon)}_\varepsilon(f(X_{1:t})), X_{t+1}), h^{(y_\varepsilon)}_\varepsilon(f(X_{t+1}))\Big) = 0.$$

By the union bound, this holds simultaneously for all $y_\varepsilon \in \mathcal{Y}'_\varepsilon$ with probability one. We therefore have that, under the specification $(\mathcal{Y}, \ell) = (\mathcal{Y}', \ell')$,

$$\limsup_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}\Big(\hat{f}^{(\varepsilon)}_{\cdot}, f; n\Big) = \limsup_{n\to\infty} \frac{1}{n}\sum_{t=0}^{n-1} \ell\Big(\hat{f}^{(\varepsilon)}_t(X_{1:t}, f(X_{1:t}), X_{t+1}), f(X_{t+1})\Big)$$
$$\leq \limsup_{n\to\infty} \frac{1}{n}\sum_{t=0}^{n-1} \Big(\ell'(g_\varepsilon(f(X_{t+1})), f(X_{t+1})) + \bar{\ell}\mathbb{1}\Big[\hat{f}^{(\varepsilon)}_t(X_{1:t}, f(X_{1:t}), X_{t+1}) \neq g_\varepsilon(f(X_{t+1}))\Big]\Big)$$
$$\leq \limsup_{n\to\infty} \frac{1}{n}\sum_{t=0}^{n-1} \left(\varepsilon + \bar{\ell}\sum_{y_\varepsilon \in \mathcal{Y}'_\varepsilon} \ell_{01}\Big(\hat{f}^{01}_t(X_{1:t}, h^{(y_\varepsilon)}_\varepsilon(f(X_{1:t})), X_{t+1}), h^{(y_\varepsilon)}_\varepsilon(f(X_{t+1}))\Big)\right)$$
$$\leq \varepsilon + \bar{\ell}\sum_{y_\varepsilon \in \mathcal{Y}'_\varepsilon} \limsup_{n\to\infty} \frac{1}{n}\sum_{t=0}^{n-1} \ell_{01}\Big(\hat{f}^{01}_t(X_{1:t}, h^{(y_\varepsilon)}_\varepsilon(f(X_{1:t})), X_{t+1}), h^{(y_\varepsilon)}_\varepsilon(f(X_{t+1}))\Big) = \varepsilon \text{ (a.s.)},$$

where the inequality on this last line is due to finiteness of $|\mathcal{Y}'_\varepsilon|$.

We now apply this argument to values $\varepsilon \in \{1/i : i \in \mathbb{N}\}$. For any measurable $f^\star : \mathcal{X} \to \mathcal{Y}'$, for each $i, n \in \mathbb{N}$, denote $\beta^{f^\star}_{i,n} = \hat{\mathcal{L}}_{\mathbb{X}}\Big(\hat{f}^{(1/i)}_{\cdot}, f^\star; n\Big)$ (under the specification $(\mathcal{Y}, \ell) = (\mathcal{Y}', \ell')$). By the above argument, together with a union bound, on an event of probability one, we have

$$\lim_{i\to\infty} \limsup_{n\to\infty} \beta^{f^\star}_{i,n} \leq \lim_{i\to\infty} 1/i = 0.$$

Thus, since these $\beta^{f^\star}_{i,n}$ are also nonnegative, Lemma 33 implies that, on this event, there exists a sequence $\{i_n\}_{n=1}^\infty$ in $\mathbb{N}$, with $i_n \leq n$ for every $n \in \mathbb{N}$, such that $\lim_{n\to\infty} \beta^{f^\star}_{i_n,n} = 0$. Therefore, applying Lemma 32 to the sequence $\Big\{\hat{f}^{(1/i)}_n\Big\}_{i=1}^\infty$ of online learning rules, we conclude that there exists an online learning rule $\hat{f}_n$ such that, for this process $\mathbb{X}$, for any measurable $f^\star : \mathcal{X} \to \mathcal{Y}'$, under the specification $(\mathcal{Y}, \ell) = (\mathcal{Y}', \ell')$, $\lim_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}\Big(\hat{f}_{\cdot}, f^\star; n\Big) = 0$ almost surely: that is, $\hat{f}_n$ is strongly universally consistent under $\mathbb{X}$. In particular, this implies $\mathbb{X} \in \mathrm{SUOL}_{(\mathcal{Y}', \ell')}$. Since this argument holds for any $\mathbb{X} \in \mathrm{SUOL}_{(\{0,1\}, \ell_{01})}$, we conclude that $\mathrm{SUOL}_{(\{0,1\}, \ell_{01})} \subseteq \mathrm{SUOL}_{(\mathcal{Y}', \ell')}$. Combining this with the first part, we have that $\mathrm{SUOL}_{(\mathcal{Y}', \ell')} = \mathrm{SUOL}_{(\{0,1\}, \ell_{01})}$, and since these arguments apply to any totally bounded $(\mathcal{Y}', \ell')$ with $\sup_{y,y' \in \mathcal{Y}'} \ell'(y, y') > 0$, this completes the proof. ■

Next, we have the analogous result for losses that are *not* totally bounded.

Theorem 42 *The set* SUOL *is invariant to the specification of* $(\mathcal{Y}, \ell)$*, subject to being separable with* $\bar{\ell} < \infty$ *but* not *totally bounded.*

Proof This proof follows the same line as that of Theorem 41, but with a few important differences. We continue the notational conventions introduced there, but in this context we let ℓ_{01} denote the 0-1 loss on $\mathbb{N}$: that is, $\forall y, y' \in \mathbb{N}$, $\ell_{01}(y, y') = \mathbb{1}[y \neq y']$. To establish the theorem, it suffices to verify the claim that $\mathrm{SUOL}_{(\mathcal{Y}', \ell')} = \mathrm{SUOL}_{(\mathbb{N}, \ell_{01})}$ for all separable metric spaces $(\mathcal{Y}', \ell')$ with $\sup\limits_{y,y' \in \mathcal{Y}'} \ell'(y, y') < \infty$ that are *not* totally bounded. Fix any such space $(\mathcal{Y}', \ell')$.

We again begin with the inclusion $\mathrm{SUOL}_{(\mathcal{Y}',\ell')} \subseteq \mathrm{SUOL}_{(\mathbb{N},\ell_{01})}$. For any $\mathbb{X} \in \mathrm{SUOL}_{(\mathcal{Y}',\ell')}$, letting $\hat{g}_n$ be an online learning rule that is strongly universally consistent under $\mathbb{X}$ (for the specification $(\mathcal{Y}, \ell) = (\mathcal{Y}', \ell')$), we can define an online learning rule $\hat{g}_n^{\mathbb{N}}$ for the specification $(\mathcal{Y}, \ell) = (\mathbb{N}, \ell_{01})$ as follows. Since $(\mathcal{Y}', \ell')$ is not totally bounded, $\exists \varepsilon > 0$ such that any $\mathcal{Y}'_\varepsilon \subseteq \mathcal{Y}'$ with $\sup\limits_{y \in \mathcal{Y}} \inf\limits_{y_\varepsilon \in \mathcal{Y}'_\varepsilon} \ell'(y_\varepsilon, y) \leq \varepsilon$ necessarily has $|\mathcal{Y}'_\varepsilon| = \infty$. In particular, this implies that for any finite sequence $z_1, \ldots, z_k \in \mathcal{Y}'$, $k \in \mathbb{N}$, there exists $z_{k+1} \in \mathcal{Y}'$ with $\inf\limits_{i \leq k} \ell'(z_i, z_{k+1}) > \varepsilon$. Thus, starting from any initial $z_1 \in \mathcal{Y}'$, we can inductively construct an infinite sequence $z_1, z_2, \ldots \in \mathcal{Y}'$ with $\inf\limits_{i,j \in \mathbb{N}: i \neq j} \ell'(z_i, z_j) \geq \varepsilon > 0$. For any $n \in \mathbb{N} \cup \{0\}$, and any sequences $x_{1:(n+1)}$ in $\mathcal{X}$ and $y_{1:n}$ in $\mathbb{N}$, define a sequence $y'_{1:n}$ with $y'_i = z_{y_i}$ for each $i \in \{1, \ldots, n\}$, and then define $\hat{g}_n^{\mathbb{N}}(x_{1:n}, y_{1:n}, x_{n+1})$ as the (unique) value $y \in \mathbb{N}$ with $\ell'(\hat{g}_n(x_{1:n}, y'_{1:n}, x_{n+1}), z_y) < \varepsilon/2$, if such a $y \in \mathbb{N}$ exists, and otherwise define it to be z_1. One can easily check that $\hat{g}_n^{\mathbb{N}}$ is a measurable function, due to measurability of $\hat{g}_n$. Then for any measurable $f : \mathcal{X} \to \mathbb{N}$, defining $f' : \mathcal{X} \to \mathcal{Y}'$ as $f'(x) = z_{f(x)}$ (which is clearly also measurable), we have (under the specification $(\mathcal{Y}, \ell) = (\mathbb{N}, \ell_{01})$)

$$\limsup_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}\left(\hat{g}_\cdot^{\mathbb{N}}, f; n\right) = \limsup_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \mathbb{1}\left[\hat{g}_t^{\mathbb{N}}(X_{1:t}, f(X_{1:t}), X_{t+1}) \neq f(X_{t+1})\right]$$

$$\leq \limsup_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \mathbb{1}\left[\ell'\big(\hat{g}_t(X_{1:t}, f'(X_{1:t}), X_{t+1}), f'(X_{t+1})\big) \geq \varepsilon/2\right]$$

$$\leq \frac{2}{\varepsilon} \limsup_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \ell'\big(\hat{g}_t(X_{1:t}, f'(X_{1:t}), X_{t+1}), f'(X_{t+1})\big) = 0 \text{ (a.s.)},$$

where the last equality (to which the "almost surely" qualifier applies) is due to strong universal consistency of $\hat{g}_n$ (and the fact that $\varepsilon > 0$). Since this argument holds for any choice of measurable $f : \mathcal{X} \to \mathbb{N}$, we conclude that $\hat{g}_n^{\mathbb{N}}$ is strongly universally consistent under $\mathbb{X}$ (for the specification $(\mathcal{Y}, \ell) = (\mathbb{N}, \ell_{01})$), so that $\mathbb{X} \in \mathrm{SUOL}_{(\mathbb{N},\ell_{01})}$. Since this argument holds for any $\mathbb{X} \in \mathrm{SUOL}_{(\mathcal{Y}',\ell')}$, we conclude that $\mathrm{SUOL}_{(\mathcal{Y}',\ell')} \subseteq \mathrm{SUOL}_{(\mathbb{N},\ell_{01})}$.

For the converse inclusion, fix any $\mathbb{X} \in \mathrm{SUOL}_{(\mathbb{N},\ell_{01})}$, and let $\hat{f}_n^{\mathbb{N}}$ be any online learning rule that is strongly universally consistent under $\mathbb{X}$ (for the specification $(\mathcal{Y}, \ell) = (\mathbb{N}, \ell_{01})$). We then define an online learning rule $\hat{f}'_n$ for the specification $(\mathcal{Y}, \ell) = (\mathcal{Y}', \ell')$ as follows. Let $\tilde{\mathcal{Y}}'$ be a countable subset of $\mathcal{Y}'$ such that $\sup\limits_{y \in \mathcal{Y}'} \inf\limits_{\tilde{y} \in \tilde{\mathcal{Y}}'} \ell'(\tilde{y}, y) = 0$; such a set $\tilde{\mathcal{Y}}'$ is guaranteed to exist by separability of $(\mathcal{Y}', \ell')$ (and furthermore, is necessarily infinite, due to $(\mathcal{Y}', \ell')$ not being totally bounded). Enumerate the elements of $\tilde{\mathcal{Y}}'$ as $\tilde{y}_1, \tilde{y}_2, \ldots$, and for each $\varepsilon > 0$ and each $y \in \mathcal{Y}'$, define $h_\varepsilon(y) = \min\{i \in \mathbb{N} : \ell'(\tilde{y}_i, y) \leq \varepsilon\}$. One can easily check that this is a measurable function $\mathcal{Y}' \to \mathbb{N}$.

For any $n \in \mathbb{N} \cup \{0\}$, and any $x_{1:n} \in \mathcal{X}^n$, $y_{1:n} \in (\mathcal{Y}')^n$, and $x \in \mathcal{X}$, define $\hat{f}_n^{(\varepsilon)}(x_{1:n}, y_{1:n}, x)$ $= \tilde{y}_i$ for $i = \hat{f}_n^{\mathbb{N}}(x_{1:n}, h_\varepsilon(y_{1:n}), x)$. That $\hat{f}_n^{(\varepsilon)}$ is a measurable function $\mathcal{X}^n \times (\mathcal{Y}')^n \times \mathcal{X} \to \mathcal{Y}'$ follows immediately from measurability of $\hat{f}_n^{\mathbb{N}}$ and h_ε. Thus, $\hat{f}_n^{(\varepsilon)}$ defines an online learning rule. Now, for any measurable function $f : \mathcal{X} \to \mathcal{Y}'$, the composed function $x \mapsto h_\varepsilon(f(x))$

is a measurable function $\mathcal{X} \to \mathbb{N}$, and therefore (by strong universal consistency of $\hat{f}_n^{\mathbb{N}}$)

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \ell_{01}\left(\hat{f}_t^{\mathbb{N}}(X_{1:t}, h_\varepsilon(f(X_{1:t})), X_{t+1}), h_\varepsilon(f(X_{t+1}))\right) = 0 \text{ (a.s.)}.$$

We therefore have that, under the specification $(\mathcal{Y}, \ell) = (\mathcal{Y}', \ell')$,

$$\limsup_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}\left(\hat{f}_\cdot^{(\varepsilon)}, f; n\right) = \limsup_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \ell\left(\hat{f}_t^{(\varepsilon)}(X_{1:t}, f(X_{1:t}), X_{t+1}), f(X_{t+1})\right)$$
$$\leq \limsup_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \left(\ell'(\tilde{y}_{h_\varepsilon(f(X_{t+1}))}, f(X_{t+1})) + \bar{\ell}\mathbb{1}\left[\hat{f}_t^{\mathbb{N}}(X_{1:t}, h_\varepsilon(f(X_{1:t})), X_{t+1}) \neq h_\varepsilon(f(X_{t+1}))\right]\right)$$
$$\leq \varepsilon + \bar{\ell} \limsup_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \ell_{01}\left(\hat{f}_t^{\mathbb{N}}(X_{1:t}, h_\varepsilon(f(X_{1:t})), X_{t+1}), h_\varepsilon(f(X_{t+1}))\right) = \varepsilon \text{ (a.s.)}.$$

The rest of this proof follows identically to the analogous part of the proof of Theorem 41. Briefly, for any measurable $f^\star : \mathcal{X} \to \mathcal{Y}'$, for each $i, n \in \mathbb{N}$, denoting $\beta_{i,n}^{f^\star} = \hat{\mathcal{L}}_{\mathbb{X}}\left(\hat{f}_\cdot^{(1/i)}, f^\star; n\right)$ (under the specification $(\mathcal{Y}, \ell) = (\mathcal{Y}', \ell')$), by the union bound, on an event of probability one, we have

$$\lim_{i\to\infty} \limsup_{n\to\infty} \beta_{i,n}^{f^\star} \leq \lim_{i\to\infty} 1/i = 0.$$

Therefore Lemma 33 (with $j_n = n$) and Lemma 32 imply that there exists an online learning rule $\hat{f}_n$ such that, for this process $\mathbb{X}$, for any measurable $f^\star : \mathcal{X} \to \mathcal{Y}'$, under the specification $(\mathcal{Y}, \ell) = (\mathcal{Y}', \ell')$, $\lim_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}\left(\hat{f}_\cdot, f^\star; n\right) = 0$ almost surely. This implies $\mathbb{X} \in \mathrm{SUOL}_{(\mathcal{Y}', \ell')}$. Since this argument holds for any $\mathbb{X} \in \mathrm{SUOL}_{(\mathbb{N}, \ell_{01})}$, we conclude $\mathrm{SUOL}_{(\mathbb{N}, \ell_{01})} \subseteq \mathrm{SUOL}_{(\mathcal{Y}', \ell')}$. Combining this with the first part, we have $\mathrm{SUOL}_{(\mathcal{Y}', \ell')} = \mathrm{SUOL}_{(\mathbb{N}, \ell_{01})}$, and since these arguments apply to any separable metric space $(\mathcal{Y}', \ell')$ with $\sup_{y, y' \in \mathcal{Y}'} \ell'(y, y') < \infty$ that is not totally bounded, this completes the proof. ■

Since the reductions used to construct the learning rules in the above two proofs do not explicitly depend on the distribution of the process $\mathbb{X}$, these proofs also establish another interesting property: namely, invariance to the specification of $(\mathcal{Y}, \ell)$ in the existence of optimistically universal learning rules. Specifically, the proof of Theorem 41 can also be used to establish that, for any given $\mathcal{X}$ space, there exists an optimistically universal online learning rule when $(\mathcal{Y}, \ell)$ is totally bounded if and only if there exists one for binary classification under the 0-1 loss. Similarly, the proof of Theorem 42 can be used to establish that, when $(\mathcal{Y}, \ell)$ is *not* totally bounded, there exists an optimistically universal online learning rule if and only if there exists one for multiclass classification with a countably infinite number of classes under the 0-1 loss.

The question of whether the two SUOL sets from the above two theorems are equivalent remains an interesting open problem.

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

In particular, in the notation of the above proofs, this problem is equivalent to the question of whether $\mathrm{SUOL}_{(\{0,1\},\ell_{01})} = \mathrm{SUOL}_{(\mathbb{N},\ell_{01})}$: that is, whether the set of processes that admit strong universal online learning is the same for *binary* classification as for *multiclass* classification with a *countably infinite* number of possible classes.

7. No Consistent Test for Existence of a Universally Consistent Learner

It is also interesting to ask to what extent admission of universal consistency is actually an *assumption*, rather than a testable hypothesis: that is, is there any way to *detect* whether or not a given data sequence $\mathbb{X}$ admits strong universal learning (in any of the above senses)? It turns out the answer is *no*.

In our present context, a *hypothesis test* is a sequence of (possibly random)[5] measurable functions $\hat{t}_n : \mathcal{X}^n \to \{0,1\}$, $n \in \mathbb{N} \cup \{0\}$. We say $\hat{t}_n$ is *consistent* for a set of processes $\mathcal{C}$ if, for every $\mathbb{X} \in \mathcal{C}$, $\hat{t}_n(X_{1:n}) \xrightarrow{P} 1$, and for every $\mathbb{X} \notin \mathcal{C}$, $\hat{t}_n(X_{1:n}) \xrightarrow{P} 0$. We have the following theorem.[6]

Theorem 43 *If $\mathcal{X}$ is infinite, there is no consistent hypothesis test for* SUIL, SUAL, *or* SUOL.

Proof Suppose $\mathcal{X}$ is infinite and fix any hypothesis test $\hat{t}_n$. Let $\{w_i\}_{i=0}^{\infty}$ be any sequence of distinct elements of $\mathcal{X}$. We construct a process $\mathbb{X}$ inductively, as follows. Let $n_0 = 0$. For the purpose of this inductive definition, suppose, for some $k \in \mathbb{N}$, that n_{k-1} is defined, and that X_t is defined for every $t \in \mathbb{N}$ with $t \le n_{k-1}$. Let $X_t^{(k)} = X_t$ for every $t \in \mathbb{N}$ with $t \le n_{k-1}$. If $(k+1)/2 \in \mathbb{N}$ (i.e., k is odd), then let $X_t^{(k)} = w_0$ for every $t \in \mathbb{N}$ with $t > n_{k-1}$. Otherwise, if $k/2 \in \mathbb{N}$ (i.e., k is even), then let $X_t^{(k)} = w_t$ for every $t \in \mathbb{N}$ with $t > n_{k-1}$. If $\exists n \in \mathbb{N}$ with $n > n_{k-1}$ such that

$$\mathbb{P}\Big(\hat{t}_n(X_{1:n}^{(k)}) = \mathbb{1}[(k+1)/2 \in \mathbb{N}]\Big) > 1/2, \tag{55}$$

then define $n_k = n$ for some such value of n, and define $X_t = X_t^{(k)}$ for every $t \in \{n_{k-1} + 1, \ldots, n_k\}$. Otherwise, if no such n exists, define $X_t = X_t^{(k)}$ for every $t \in \mathbb{N}$ with $t > n_{k-1}$, in which case the inductive definition is complete (upon reaching the smallest value of k for which no such n exists).

The above inductive definition specifies a deterministic process $\mathbb{X}$. Now consider two cases. First, suppose there is a maximum value k^* of $k \in \mathbb{N}$ for which n_{k-1} is defined. In this case, there is no $n > n_{k^*-1}$ satisfying (55) with $k = k^*$. Furthermore, by the definition of $X_t^{(k^*)}$ for every $t \le n_{k^*-1}$, and by our choice of X_t for every $t > n_{k^*-1}$, we have $\mathbb{X} = \{X_t^{(k^*)}\}_{t=1}^{\infty}$. Together, these imply that $\forall n \in \mathbb{N}$ with $n > n_{k^*-1}$,

$$\mathbb{P}\big(\hat{t}_n(X_{1:n}) = \mathbb{1}[(k^*+1)/2 \in \mathbb{N}]\big) \le 1/2. \tag{56}$$

5. In the case of random $\hat{t}_n$, we will suppose $\hat{t}_n$ is independent from $\mathbb{X}$.
6. There is actually a fairly simple proof of this theorem if $\mathcal{X}$ is uncountable and $(\mathcal{X}, \mathcal{T})$ is a Polish space. In that case, we can simply use the fact that no test can distinguish between an i.i.d. process with a given nonatomic marginal distribution vs a deterministic process chosen randomly among the sample paths of the i.i.d. process. However, the proof we present here has the advantage of applying also to countable $\mathcal{X}$, and indeed it remains valid even if we restrict to *deterministic* processes.

If $(k^*+1)/2 \in \mathbb{N}$, then $X_t = w_0$ for every $t \in \mathbb{N}$ with $t > n_{k^*-1}$. In this case, for any $A \in \mathcal{B}$, $\hat{\mu}_{\mathbb{X}}(A) = \mathbb{1}_A(w_0)$. Thus, for any monotone sequence $\{A_i\}_{i=1}^{\infty}$ of sets in $\mathcal{B}$ with $A_i \downarrow \emptyset$, $\lim_{i\to\infty} \mathbb{E}[\hat{\mu}_{\mathbb{X}}(A_i)] = \lim_{i\to\infty} \mathbb{1}_{A_i}(w_0) = \mathbb{1}_{\lim_{i\to\infty} A_i}(w_0) = \mathbb{1}_{\emptyset}(w_0) = 0$. Therefore, $\mathbb{X}$ satisfies Condition 1 (i.e., $\mathbb{X} \in \mathcal{C}_1$). Since Theorem 7 implies SUIL = SUAL = $\mathcal{C}_1$, we also have that $\mathbb{X} \in$ SUIL and $\mathbb{X} \in$ SUAL. Also, since Theorem 38 implies SUAL $\subset$ SUOL, we have $\mathbb{X} \in$ SUOL as well. However, (56) implies $\limsup_{n\to\infty} \mathbb{P}(\hat{t}_n(X_{1:n}) \neq 1) \geq 1/2$, so that $\hat{t}_n(X_{1:n})$ fails to converge in probability to 1, and hence $\hat{t}_n$ is not consistent for any of SUIL, SUAL, or SUOL.

On the other hand, if $(k^*+1)/2 \notin \mathbb{N}$, then $X_t = w_t$ for every $t \in \mathbb{N}$ with $t > n_{k^*-1}$. In this case, letting $A_i = \{w_i\} \in \mathcal{B}$ for each $i \in \mathbb{N}$, these A_i sets are disjoint, and for any $T \in \mathbb{N}$, $|\{i \in \mathbb{N} : X_{1:T} \cap A_i \neq \emptyset\}| \geq T - n_{k^*-1} \neq o(T)$, so that $\mathbb{X}$ fails to satisfy Condition 2: that is, $\mathbb{X} \notin \mathcal{C}_2$. Since Theorem 35 implies SUOL $\subseteq \mathcal{C}_2$, and Theorems 7 and 38 imply SUIL = SUAL $\subset$ SUOL, we also have that $\mathbb{X} \notin$ SUOL, $\mathbb{X} \notin$ SUAL, and $\mathbb{X} \notin$ SUIL. However, (56) implies $\limsup_{n\to\infty} \mathbb{P}(\hat{t}_n(X_{1:n}) \neq 0) \geq 1/2$, so that $\hat{t}_n(X_{1:n})$ fails to converge in probability to 0, and hence $\hat{t}_n$ is not consistent for any of SUIL, SUAL, or SUOL.

For the remaining case, suppose n_k is defined for all $k \in \mathbb{N} \cup \{0\}$, so that $\{n_k\}_{k=0}^{\infty}$ is an infinite strictly-increasing sequence of nonnegative integers. For each $k \in \mathbb{N}$, our choice of n_k guarantees that (55) is satisfied with $n = n_k$. Furthermore, for every $k \in \mathbb{N}$, our definition of $X_t^{(k)}$ for values $t \leq n_{k-1}$, and our choice of X_t for values $t \in \{n_{k-1}+1, \ldots, n_k\}$ imply that $X_{1:n_k} = X_{1:n_k}^{(k)}$. Thus, every $k \in \mathbb{N}$ satisfies $\mathbb{P}(\hat{t}_{n_k}(X_{1:n_k}) = \mathbb{1}[(k+1)/2 \in \mathbb{N}]) > 1/2$. In particular, this implies that

$$\limsup_{n\to\infty} \mathbb{P}(\hat{t}_n(X_{1:n}) \neq 1) \geq \limsup_{j\to\infty} \mathbb{P}(\hat{t}_{n_{2j}}(X_{1:n_{2j}}) = 0) \geq 1/2,$$

while

$$\limsup_{n\to\infty} \mathbb{P}(\hat{t}_n(X_{1:n}) \neq 0) \geq \limsup_{j\to\infty} \mathbb{P}(\hat{t}_{n_{2j+1}}(X_{1:n_{2j+1}}) = 1) \geq 1/2.$$

Thus, $\hat{t}_n(X_{1:n})$ fails to converge in probability to any value: that is, it neither converges in probability to 0 nor converges in probability to 1. Therefore, in this case as well, we find that $\hat{t}_n$ is not consistent for any of SUIL, SUAL, or SUOL.

Thus, regardless of which of these is the case, we have established that $\hat{t}_n$ is not a consistent test for SUIL, SUAL, or SUOL. ∎

Recall that, if $\mathcal{X}$ is *finite*, *every* $\mathbb{X}$ admits strong universal inductive learning: any sequence $A_k \downarrow \emptyset$ has $A_k = \emptyset$ for all sufficiently large k, so that every $\mathbb{X}$ has $\lim_{k\to\infty} \mathbb{E}[\hat{\mu}_{\mathbb{X}}(A_k)] = \hat{\mu}_{\mathbb{X}}(\emptyset) = 0$, and hence satisfies Condition 1, which implies $\mathbb{X} \in$ SUIL $\cap$ SUAL $\cap$ SUOL by Theorems 7 and 38. Therefore, the *constant* function $\hat{t}_n(\cdot) = 1$ is a consistent test for SUIL, SUAL, and SUOL in this case. Thus, we may conclude the following corollary.

Corollary 44 *There exist consistent hypothesis tests for each of* SUIL*,* SUAL*, and* SUOL *if and only if $\mathcal{X}$ is finite.*

Note that, since Theorem 7 implies SUIL = $\mathcal{C}_1$, this corollary also holds for consistent tests of $\mathcal{C}_1$. It is also easy to see that the proof above can further extend this corollary to consistent tests of $\mathcal{C}_2$ as well.

8. Unbounded Losses

In this section, we depart from the above discussion by considering the case of unbounded losses. Specifically, we retain the assumption that $(\mathcal{Y}, \ell)$ is a separable metric space, but now we replace the assumption that ℓ is bounded (i.e., $\bar{\ell} < \infty$) with the complementary assumption that $\bar{\ell} = \infty$. To be clear, we suppose $\ell(y_1, y_2)$ is finite for every $y_1, y_2 \in \mathcal{Y}$, but is *unbounded*, in that $\sup\limits_{y_1, y_2 \in \mathcal{Y}} \ell(y_1, y_2) = \infty$. All of the other restrictions from Section 1.1 (e.g., that $(\mathcal{Y}, \ell)$ is a separable metric space) remain unchanged. In this setting, we find that the condition necessary and sufficient for a process to admit universal learning becomes significantly stronger. Indeed, it was already known that not even all i.i.d. processes admit universal learning when $\bar{\ell} = \infty$. However, we are nevertheless able to establish results on the existence of optimistically universal learning rules and consistent tests. We again find that the set of processes admitting strong universal learning is invariant to ℓ (subject to $\bar{\ell} = \infty$), and specified by a simple condition. Specifically, consider the following condition.

Condition 3 *Every monotone sequence $\{A_k\}_{k=1}^{\infty}$ of sets in $\mathcal{B}$ with $A_k \downarrow \emptyset$ satisfies*

$$|\{k \in \mathbb{N} : \mathbb{X} \cap A_k \neq \emptyset\}| < \infty \text{ (a.s.).}$$

We denote by $\mathcal{C}_3$ the set of processes $\mathbb{X}$ satisfying Condition 3. It is straightforward to see that Condition 3 is equivalent to the condition that, for every *disjoint* sequence $\{B_k\}_{k=1}^{\infty}$ in $\mathcal{B}$, $|\{k \in \mathbb{N} : \mathbb{X} \cap B_k \neq \emptyset\}| < \infty$ (a.s.). To see this, note that, given any monotone sequence $A_k \downarrow \emptyset$, the sequence $B_k = A_k \setminus A_{k+1}$ is disjoint. Conversely, given any disjoint sequence $\{B_k\}_{k=1}^{\infty}$, the sequence $A_k = \bigcup\limits_{k'=k}^{\infty} B_{k'}$ is monotone with $A_k \downarrow \emptyset$. In either case, we have that

$$|\{k \in \mathbb{N} : \mathbb{X} \cap A_k \neq \emptyset\}| = \sup\{k \in \mathbb{N} : \mathbb{X} \cap A_k \neq \emptyset\} = \sup\{k \in \mathbb{N} : \mathbb{X} \cap B_k \neq \emptyset\},$$

and the rightmost expression is clearly finite if and only if $|\{k \in \mathbb{N} : \mathbb{X} \cap B_k \neq \emptyset\}| < \infty$.

It is straightforward to see that any process satisfying Condition 3 necessarily also satisfies Condition 1: i.e., $\mathcal{C}_3 \subseteq \mathcal{C}_1$. Specifically, for any $\mathbb{X} \in \mathcal{C}_3$, for any sequence $\{A_k\}_{k=1}^{\infty}$ in $\mathcal{B}$ with $A_k \downarrow \emptyset$, with probability one every sufficiently large k has $\mathbb{X} \cap A_k = \emptyset$, which implies $\lim\limits_{k \to \infty} \hat{\mu}_{\mathbb{X}}(A_k) = 0$; thus, $\mathbb{X} \in \mathcal{C}_1$ by Lemma 13.

Condition 3 will turn out to be the key condition for determining whether a given process admits strong universal learning (in *any* of the three protocols: inductive, self-adaptive, or online) when the loss is unbounded, analogous to the role of Condition 1 for the case of bounded losses in inductive and self-adaptive learning. This is stated formally in the following theorem.

Theorem 45 *When $\bar{\ell} = \infty$, the following statements are equivalent for any process $\mathbb{X}$.*

- $\mathbb{X}$ *satisfies Condition 3.*
- $\mathbb{X}$ *admits strong universal inductive learning.*
- $\mathbb{X}$ *admits strong universal self-adaptive learning.*
- $\mathbb{X}$ *admits strong universal online learning.*

Equivalently, when $\bar{\ell} = \infty$, SUOL = SUAL = SUIL = $\mathcal{C}_3$.

We present the proof of this result in Section 8.3 below. One remarkable consequence of this result is that, unlike Theorem 7 for bounded losses, this theorem includes *online* learning among the equivalences. This is noteworthy for two reasons. First, in the case of bounded losses, we found (in Theorem 38) that SUOL is typically *not* equivalent to SUAL and SUIL, instead forming a strict superset of these. This therefore creates an interesting distinction between bounded and unbounded losses regarding the relative strengths of these settings. A second interesting contrast to the above analysis of bounded losses is that, in the case of unbounded losses, Theorem 45 establishes a concise condition that is necessary and sufficient for a process to admit strong universal online learning; this contrasts with the analysis of online learning for bounded losses in Section 6, where we fell short of provably establishing a concise characterization of the processes admitting strong universal online learning (see Open Problem 2).

In addition to the above equivalence, we also find that in *all three* learning settings studied here, for unbounded losses, there exist optimistically universal learning rules. We have the following theorem, the proof of which is given in Section 8.3 below.

Theorem 46 *When* $\bar{\ell} = \infty$*, there exists an optimistically universal (inductive / self-adaptive / online) learning rule.*

Indeed, we find that effectively the *same* learning strategy, described in (66) below, suffices for optimistically universal learning in all three of these settings.

8.1 A Question Concerning the Number of Distinct Values

It is worth noting that Condition 3 is quite restrictive. In fact, it is even violated by many i.i.d. processes: namely, all those with the marginal distribution of X_t having infinite support. Clearly any process $\mathbb{X}$ such that the number of distinct points X_t is (almost surely) finite satisfies Condition 3. Indeed, for deterministic processes or for countable $\mathcal{X}$, one can easily show that this is *equivalent* to Condition 3. But in general, it is not presently known whether there exist processes $\mathbb{X}$ satisfying Condition 3 for which the number of distinct X_t values is *infinite* with nonzero probability. Thus we have the following open question.

Open Problem 4 *For some uncountable* $\mathcal{X}$*, does there exist* $\mathbb{X} \in \mathcal{C}_3$ *such that, with nonzero probability,* $|\{x \in \mathcal{X} : \mathbb{X} \cap \{x\} \neq \emptyset\}| = \infty$*?*

Either answer to this question would be interesting. If no such processes $\mathbb{X}$ exist, then the proof of Theorem 45 below could be dramatically simplified, since it would then be completely trivial to construct a strongly universally consistent learning rule (in any of the three settings) under $\mathbb{X} \in \mathcal{C}_3$, simply using memorization (once n is sufficiently large, all the distinct points will have been observed in the training sample). On the other hand, if there do exist such processes, then it would indicate that $\mathcal{C}_3$ is in fact a fairly rich family of processes, and that the learning problem is indeed nontrivial. It is straightforward to show that, if such processes do exist for $\mathcal{X} = [0, 1]$ (with the standard topology), then there would also exist processes of this type that are *convergent* (to a nondeterministic limit point) almost surely;[7] thus, in attempting to answer Open Problem 4 (in the case of $\mathcal{X} = [0,1]$), it suffices to focus on convergent processes.

8.2 An Equivalent Condition

Before getting into the discussion of consistency under processes in $\mathcal{C}_3$, we first note an elegant equivalent formulation of the condition, which may help to illuminate its relevance to the problem of learning with unbounded losses. Specifically, we have the following result.

Lemma 47 *A process $\mathbb{X}$ satisfies Condition 3 if and only if every measurable function $f : \mathcal{X} \to \mathbb{R}$ satisfies*

$$\sup_{t \in \mathbb{N}} f(X_t) < \infty \text{ (a.s.)}.$$

Proof First, suppose $\mathbb{X} \in \mathcal{C}_3$, and fix any measurable $f : \mathcal{X} \to \mathbb{R}$. For each $k \in \mathbb{N}$, define $A_k = f^{-1}([k-1, \infty))$. Since $f(x) < \infty$ for every $x \in \mathcal{X}$, we have $A_k \downarrow \emptyset$. Thus, by the definition of $\mathcal{C}_3$, with probability one $\exists k_0 \in \mathbb{N}$ such that $\mathbb{X} \cap A_{k_0+1} = \emptyset$; in other words, with probability one, $\exists k_0 \in \mathbb{N}$ such that every $t \in \mathbb{N}$ has $f(X_t) < k_0$, so that $\sup_{t \in \mathbb{N}} f(X_t) \le k_0 < \infty$.

For the other direction, suppose $\mathbb{X}$ is such that every measurable $f : \mathcal{X} \to \mathbb{R}$ satisfies $\sup_{t \in \mathbb{N}} f(X_t) < \infty$ (a.s.). Fix any monotone sequence $\{A_k\}_{k=1}^{\infty}$ of sets in $\mathcal{B}$ with $A_k \downarrow \emptyset$, and define a function $f : \mathcal{X} \to \mathbb{R}$ such that, $\forall x \in \mathcal{X}$, $f(x) = \sum_{k=1}^{\infty} \mathbb{1}_{A_k}(x) = |\{k \in \mathbb{N} : x \in A_k\}|$. Note that, since $A_k \downarrow \emptyset$, we indeed have $f(x) \in \mathbb{R}$ for every $x \in \mathcal{X}$. Furthermore, f is clearly measurable (being a limit of simple functions). Thus, $\sup_{t \in \mathbb{N}} f(X_t) < \infty$ (a.s.). Also note that monotonicity of the sequence $\{A_k\}_{k=1}^{\infty}$ implies that $\forall x \in \mathcal{X}$, $f(x) = \max(\{k \in \mathbb{N} : x \in A_k\} \cup \{0\})$. Thus, denoting $\hat{k} = \sup_{t \in \mathbb{N}} f(X_t)$, on the event (of probability one) that $\hat{k} < \infty$, every $k \in \mathbb{N}$ with $k > \hat{k}$ has $\mathbb{X} \cap A_k = \emptyset$, so that $|\{k \in \mathbb{N} : \mathbb{X} \cap A_k \neq \emptyset\}| \le \hat{k} < \infty$ (in fact, the first inequality holds with equality). Since this holds for any choice of monotone sequence $\{A_k\}_{k=1}^{\infty}$ in $\mathcal{B}$ with $A_k \downarrow \emptyset$, we have that $\mathbb{X} \in \mathcal{C}_3$. ■

8.3 Proofs of the Main Results for Unbounded Losses

This subsection presents the proofs of Theorems 45 and 46. As with Theorem 7, we prove Theorem 45 via a sequence of lemmas, corresponding to the implications among the various statements claimed to be equivalent. The first of these is analogous to Lemma 18, showing that processes admitting strong universal inductive learning also admit strong universal self-adaptive learning. The proof is identical to that of Lemma 18, and as such is omitted.

Lemma 48 *When $\bar{\ell} = \infty$,* SUIL $\subseteq$ SUAL*.*

Next, we have a result analogous to Lemma 19, showing that any process admitting strong universal self-adaptive or online learning necessarily satisfies Condition 3.

7. For instance, for $\{U_t\}_{t=0}^{\infty}$ i.i.d. Uniform$(0, 2/3)$, the process $X_t = U_0 + 2^{-t} U_t$ is convergent to the nondeterministic limit U_0.

Lemma 49 *When $\bar{\ell} = \infty$,* $\mathrm{SUAL} \cup \mathrm{SUOL} \subseteq \mathcal{C}_3$.

Proof Fix any $\mathbb{X}$ that fails to satisfy Condition 3. Then there exists a monotone sequence $\{B_k\}_{k=1}^{\infty}$ in $\mathcal{B}$ with $B_k \downarrow \emptyset$ such that, on a $\sigma(\mathbb{X})$-measurable event E of probability strictly greater than zero,

$$|\{k \in \mathbb{N} : \mathbb{X} \cap B_k \neq \emptyset\}| = \infty. \tag{57}$$

Furthermore, monotonicity of $B \mapsto \mathbb{X} \cap B$ implies that, without loss of generality, we may suppose $B_1 = \mathcal{X}$. Also, by monotonicity of $\{B_k\}_{k=1}^{\infty}$, on the event E, (57) implies that

$$\forall k \in \mathbb{N}, \mathbb{X} \cap B_k \neq \emptyset. \tag{58}$$

Now for each $i \in \mathbb{N}$, define $A_i = B_i \setminus B_{i+1}$. Note that, due to monotonicity of the $\{B_k\}_{k=1}^{\infty}$ sequence and the facts that $B_k \downarrow \emptyset$ and $B_1 = \mathcal{X}$, $\{A_i\}_{i=1}^{\infty}$ is a disjoint sequence of sets in $\mathcal{B}$ with $\bigcup_{i=1}^{\infty} A_i = \mathcal{X}$. Thus, for every $t \in \mathbb{N}$, there exists a unique $i_t \in \mathbb{N}$ with $X_t \in A_{i_t}$. Also note that every $j \in \mathbb{N}$ has $B_j = \bigcup_{i \geq j} A_i$, again due to monotonicity of $\{B_k\}_{k=1}^{\infty}$ and the fact that $B_k \downarrow \emptyset$.

For each $j \in \mathbb{N}$, define a random variable

$$\tau_j = \begin{cases} \min\{t \in \mathbb{N} : X_t \in B_j\}, & \text{if } \mathbb{X} \cap B_j \neq \emptyset \\ 0, & \text{otherwise} \end{cases}.$$

Note that, on the event E, (58) implies that we have $\tau_j = \min\{t \in \mathbb{N} : X_t \in B_j\}$ for every $j \in \mathbb{N}$ (and that this minimum exists and is well-defined). Let $\{T_j\}_{j=1}^{\infty}$ be a nondecreasing sequence of (nonrandom) values in $\mathbb{N} \cup \{0\}$ such that, for each $j \in \mathbb{N}$,

$$\mathbb{P}(\tau_j > T_j) < 2^{-j}.$$

Such a sequence must exist, since τ_j always has a finite value, so that $\lim_{t \to \infty} \mathbb{P}(\tau_j > t) = 0$ (e.g., Schervish, 1995, Theorem A.19). Since $\sum_{j=1}^{\infty} \mathbb{P}(\tau_j > T_j) < \sum_{j=1}^{\infty} 2^{-j} = 1 < \infty$, the Borel-Cantelli Lemma implies that, on a $\sigma(\mathbb{X})$-measurable event E' of probability one, $\exists \iota_0 \in \mathbb{N}$ such that $\forall j \in \mathbb{N}$ with $j \geq \iota_0$, $\tau_j \leq T_j$. For each $i \in \mathbb{N}$, let $y_{i,0}, y_{i,1} \in \mathcal{Y}$ be such that $\ell(y_{i,0}, y_{i,1}) > T_i$. For every $\kappa \in [0,1)$ and $i \in \mathbb{N}$, denote $\kappa_i = \lfloor 2^i \kappa \rfloor - 2\lfloor 2^{i-1}\kappa \rfloor$: the i^{th} bit of the binary representation of κ. Then for each $\kappa \in [0,1)$, $i \in \mathbb{N}$, and $x \in A_i$, define $f_\kappa^\star(x) = y_{i,\kappa_i}$. Note that $f_\kappa^\star$ is clearly a measurable function for every $\kappa \in [0,1)$ (as it is constant within each of the A_i sets, and $\mathcal{X}$ is the union of these sets),

So that we may treat both self-adaptive and online learning simultaneously, for any $n, m \in \mathbb{N} \cup \{0\}$, let $f_{n,m}$ denote any measurable function $\mathcal{X}^m \times \mathcal{Y}^m \times \mathcal{X} \to \mathcal{Y}$. We will see below that any online learning rule can be expressed as such a function by simply disregarding the n index, while any self-adaptive learning rule can be expressed as such a function by disregarding the $\mathcal{Y}$-valued arguments beyond the first n (when $m \geq n$). Additionally, for every $x \in \mathcal{X}$, $n, m \in \mathbb{N} \cup \{0\}$, and every $\kappa \in [0,1)$, for brevity we denote

$f^{\kappa}_{n,m}(x) = f_{n,m}(X_{1:m}, \{y_{i_j,\kappa_{i_j}}\}_{j=1}^{m}, x)$. We generally have

$$\begin{aligned}
&\sup_{\kappa \in [0,1)} \mathbb{E}\left[\limsup_{n\to\infty} \limsup_{t\to\infty} \frac{1}{t} \sum_{m=0}^{t-1} \ell\Big(f^{\kappa}_{n,m}(X_{m+1}), y_{i_{m+1},\kappa_{i_{m+1}}} \Big) \right] \\
&\geq \int_0^1 \mathbb{E}\left[\limsup_{n\to\infty} \limsup_{t\to\infty} \frac{1}{t} \sum_{m=0}^{t-1} \ell\Big(f^{\kappa}_{n,m}(X_{m+1}), y_{i_{m+1},\kappa_{i_{m+1}}} \Big) \right] \mathrm{d}\kappa. \qquad (59)
\end{aligned}$$

We therefore aim to establish that this last expression is strictly greater than 0.

Since ℓ is nonnegative, Tonelli's theorem implies that the last expression in (59) equals

$$\begin{aligned}
&\mathbb{E}\left[\int_0^1 \limsup_{n\to\infty} \limsup_{t\to\infty} \frac{1}{t} \sum_{m=0}^{t-1} \ell\Big(f^{\kappa}_{n,m}(X_{m+1}), y_{i_{m+1},\kappa_{i_{m+1}}} \Big) \mathrm{d}\kappa \right] \\
&\geq \mathbb{E}\left[\mathbb{1}_{E\cap E'} \int_0^1 \limsup_{n\to\infty} \limsup_{t\to\infty} \frac{1}{t} \sum_{m=0}^{t-1} \ell\Big(f^{\kappa}_{n,m}(X_{m+1}), y_{i_{m+1},\kappa_{i_{m+1}}} \Big) \mathrm{d}\kappa \right]. \qquad (60)
\end{aligned}$$

Since $B_k \downarrow \emptyset$, for any $t \in \mathbb{N}$ there exists $k_t \in \mathbb{N}$ with $X_{1:t} \cap B_{k_t} = \emptyset$, which (by monotonicity of $\{B_j\}_{j=1}^{\infty}$) implies that on the event E (so that (58) holds), every integer $j \geq k_t$ has $\tau_j > t$. Thus, on E, $\tau_j \to \infty$ as $j \to \infty$. Therefore, the expression in (60) is at least as large as

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}_{E\cap E'} \int_0^1 \limsup_{n\to\infty} \limsup_{j\to\infty} \frac{1}{\tau_j} \sum_{m=0}^{\tau_j-1} \ell\Big(f^{\kappa}_{n,m}(X_{m+1}), y_{i_{m+1},\kappa_{i_{m+1}}} \Big) \mathrm{d}\kappa \right] \\
&\geq \mathbb{E}\left[\mathbb{1}_{E\cap E'} \int_0^1 \limsup_{n\to\infty} \limsup_{j\to\infty} \frac{1}{\tau_j} \left(\ell\Big(f^{\kappa}_{n,\tau_j-1}(X_{\tau_j}), y_{i_{\tau_j},\kappa_{i_{\tau_j}}} \Big) \wedge \tau_j \right) \mathrm{d}\kappa \right]. \qquad (61)
\end{aligned}$$

In particular, since $\forall n, j \in \mathbb{N}$ with $\tau_j > 0$, we have $\frac{1}{\tau_j}\left(\ell\Big(f^{\kappa}_{n,\tau_j-1}(X_{\tau_j}), y_{i_{\tau_j},\kappa_{i_{\tau_j}}} \Big) \wedge \tau_j \right) \leq 1$, Fatou's lemma (applied twice) implies that (61) is at least as large as

$$\mathbb{E}\left[\mathbb{1}_{E\cap E'} \limsup_{n\to\infty} \limsup_{j\to\infty} \frac{1}{\tau_j} \int_0^1 \left(\ell\Big(f^{\kappa}_{n,\tau_j-1}(X_{\tau_j}), y_{i_{\tau_j},\kappa_{i_{\tau_j}}} \Big) \wedge \tau_j \right) \mathrm{d}\kappa \right]. \qquad (62)$$

Now note that on the event E, for every $j \in \mathbb{N}$, minimality of τ_j implies that every $t \in \mathbb{N}$ with $t < \tau_j$ has $X_t \notin B_j$, and since $B_j = \bigcup_{i\geq j} A_i$, this implies $i_t < j$. Furthermore, on E, by definition of τ_j we have $X_{\tau_j} \in B_j = \bigcup_{i\geq j} A_i$, so that $i_{\tau_j} \geq j$ for every $j \in \mathbb{N}$. Together these facts imply that on E, every $j \in \mathbb{N}$ has $i_{\tau_j} \notin \{i_1, \ldots, i_{\tau_j-1}\}$, so that $f^{\kappa}_{n,\tau_j-1}(X_{\tau_j})$ is functionally independent of $\kappa_{i_{\tau_j}}$. Therefore, for $K \sim \mathrm{Uniform}([0,1))$ independent of $\mathbb{X}$ and f_{n,τ_j-1}, it holds that $f^{K}_{n,\tau_j-1}(X_{\tau_j})$ is conditionally independent of $K_{i_{\tau_j}}$ given $K_{i_1}, \ldots, K_{i_{\tau_j-1}}$, $\mathbb{X}$, and f_{n,τ_j-1}, on the event E. Furthermore, on this event, $K_{i_{\tau_j}}$ is conditionally independent of $K_{i_1}, \ldots, K_{i_{\tau_j-1}}$ given $\mathbb{X}$, and f_{n,τ_j-1}, and the conditional distribution of $K_{i_{\tau_j}}$ is

Bernoulli(1/2), given $\mathbb{X}$ and f_{n,τ_j-1}, on this event. Therefore, on the event E,

$$\int_0^1 \left(\ell\left(f^{\kappa}_{n,\tau_j-1}(X_{\tau_j}), y_{i_{\tau_j},\kappa_{i_{\tau_j}}}\right) \wedge \tau_j\right) \mathrm{d}\kappa = \mathbb{E}\left[\left(\ell\left(f^{K}_{n,\tau_j-1}(X_{\tau_j}), y_{i_{\tau_j},K_{i_{\tau_j}}}\right) \wedge \tau_j\right) \middle| \mathbb{X}, f_{n,\tau_j-1}\right]$$
$$= \mathbb{E}\left[\mathbb{E}\left[\left(\ell\left(f_{n,\tau_j-1}\left(X_{1:(\tau_j-1)}, \{y_{i_s,K_{i_s}}\}_{s=1}^{\tau_j-1}, X_{\tau_j}\right), y_{i_{\tau_j},K_{i_{\tau_j}}}\right) \wedge \tau_j\right) \middle| \mathbb{X}, \{K_{i_s}\}_{s=1}^{\tau_j-1}, f_{n,\tau_j-1}\right] \middle| \mathbb{X}, f_{n,\tau_j-1}\right]$$
$$= \mathbb{E}\left[\sum_{b\in\{0,1\}} \frac{1}{2}\left(\ell\left(f_{n,\tau_j-1}\left(X_{1:(\tau_j-1)}, \{y_{i_s,K_{i_s}}\}_{s=1}^{\tau_j-1}, X_{\tau_j}\right), y_{i_{\tau_j},b}\right) \wedge \tau_j\right) \middle| \mathbb{X}, f_{n,\tau_j-1}\right]. \tag{63}$$

Since $\tau_j \geq 0$, one can easily verify that $\ell(\cdot,\cdot) \wedge \tau_j$ is a pseudo-metric. Thus, by the triangle inequality,

$$\sum_{b\in\{0,1\}} \left(\ell\left(f_{n,\tau_j-1}\left(X_{1:(\tau_j-1)}, \{y_{i_s,K_{i_s}}\}_{s=1}^{\tau_j-1}, X_{\tau_j}\right), y_{i_{\tau_j},b}\right) \wedge \tau_j\right)$$
$$\geq \ell\left(y_{i_{\tau_j},0}, y_{i_{\tau_j},1}\right) \wedge \tau_j \geq T_{i_{\tau_j}} \wedge \tau_j. \tag{64}$$

As established above, on the event E, every $j \in \mathbb{N}$ has $i_{\tau_j} \geq j$. Since $\{T_i\}_{i=1}^{\infty}$ is nondecreasing, this implies that, on E, $T_{i_{\tau_j}} \geq T_j$. Furthermore, on the event E', every $j \geq \iota_0$ has $T_j \geq \tau_j$. Combining this with (63) and (64) yields that, on the event $E \cap E'$, $\forall n, j \in \mathbb{N}$ with $j \geq \iota_0$,

$$\int_0^1 \left(\ell\left(f^{\kappa}_{n,\tau_j-1}(X_{\tau_j}), y_{i_{\tau_j},\kappa_{i_{\tau_j}}}\right) \wedge \tau_j\right) \mathrm{d}\kappa \geq \mathbb{E}\left[\frac{1}{2}\tau_j \middle| \mathbb{X}, f_{n,\tau_j-1}\right] = \frac{1}{2}\tau_j,$$

where the rightmost equality follows from $\sigma(\mathbb{X})$-measurability of τ_j. Therefore, the expression in (62) is at least as large as

$$\mathbb{E}\left[\mathbb{1}_{E\cap E'} \limsup_{n\to\infty} \limsup_{j\to\infty} \frac{1}{\tau_j}\left(\frac{1}{2}\tau_j\right)\right] = \frac{1}{2}\mathbb{P}\left(E \cap E'\right) \geq \frac{1}{2}\left(\mathbb{P}(E) - \mathbb{P}((E')^c)\right) = \frac{1}{2}\mathbb{P}(E),$$

where the rightmost equality is due to the fact that $\mathbb{P}(E') = 1$. In particular, recall that $\mathbb{P}(E) > 0$, so that the above is strictly greater than zero.

Altogether, we have established that the last expression in (59) is strictly greater than 0. By the inequality in (59) this implies $\exists \kappa \in [0,1)$ such that

$$\mathbb{E}\left[\limsup_{n\to\infty} \limsup_{t\to\infty} \frac{1}{t}\sum_{m=0}^{t-1} \ell\left(f^{\kappa}_{n,m}(X_{m+1}), y_{i_{m+1},\kappa_{i_{m+1}}}\right)\right] > 0,$$

which further implies (see e.g., Theorem 1.6.5 of Ash and Doléans-Dade, 2000) that, with probability strictly greater than zero,

$$\limsup_{n\to\infty} \limsup_{t\to\infty} \frac{1}{t}\sum_{m=0}^{t-1} \ell\left(f^{\kappa}_{n,m}(X_{m+1}), y_{i_{m+1},\kappa_{i_{m+1}}}\right) > 0.$$

This argument applies to any measurable functions $f_{n,m} : \mathcal{X}^m \times \mathcal{Y}^m \times \mathcal{X} \to \mathcal{Y}$. In particular, for any online learning rule h_n, we can define a function $f_{n,m}(x_{1:m}, y_{1:m}, x) =$

$h_m(x_{1:m}, y_{1:m}, x)$ (for every $n, m \in \mathbb{N} \cup \{0\}$ and $x_{1:m} \in \mathcal{X}^m$, $y_{1:m} \in \mathcal{Y}^m$, $x \in \mathcal{X}$), in which case any $\kappa \in [0,1)$ has

$$\limsup_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(h_\cdot, f^\star_\kappa; n) = \limsup_{n\to\infty} \limsup_{t\to\infty} \frac{1}{t} \sum_{m=0}^{t-1} \ell\Big(f^\kappa_{n,m}(X_{m+1}), y_{i_{m+1},\kappa_{i_{m+1}}}\Big).$$

Therefore, the above argument implies that $\exists \kappa \in [0,1)$ for which, with probability strictly greater than 0, $\limsup_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(h_\cdot, f^\star_\kappa; n) > 0$, so that h_n is not strongly universally consistent under $\mathbb{X}$. Since this argument applies to any online learning rule h_n, this implies $\mathbb{X} \notin$ SUOL, and since the argument applies to any process $\mathbb{X}$ failing to satisfy Condition 3, we conclude that SUOL $\subseteq \mathcal{C}_3$.

Similarly, for any self-adaptive learning rule $g_{n,m}$, for every $n, m \in \mathbb{N} \cup \{0\}$ with $m \geq n$, we can define a function $f_{n,m}(x_{1:m}, y_{1:m}, x) = g_{n,m}(x_{1:m}, y_{1:n}, x)$ (for every $x_{1:m} \in \mathcal{X}^m$, $y_{1:m} \in \mathcal{Y}^m$, $x \in \mathcal{X}$). For $n, m \in \mathbb{N} \cup \{0\}$ with $m < n$, we can simply define $f_{n,m}(x_{1:m}, y_{1:m}, x)$ as an arbitrary fixed $y \in \mathcal{Y}$ (invariant to the arguments $x_{1:m} \in \mathcal{X}^m$, $y_{1:m} \in \mathcal{Y}^m$, $x \in \mathcal{X}$). Then for any $\kappa \in [0,1)$, we have

$$\begin{aligned}\limsup_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(g_{n,\cdot}, f^\star_\kappa; n) &= \limsup_{n\to\infty} \limsup_{t\to\infty} \frac{1}{t+1} \sum_{m=n}^{n+t} \ell\Big(f^\kappa_{n,m}(X_{m+1}), y_{i_{m+1},\kappa_{i_{m+1}}}\Big) \\ &= \limsup_{n\to\infty} \limsup_{t\to\infty} \frac{1}{t} \sum_{m=0}^{t-1} \ell\Big(f^\kappa_{n,m}(X_{m+1}), y_{i_{m+1},\kappa_{i_{m+1}}}\Big).\end{aligned}$$

Therefore, the above argument implies that $\exists \kappa \in [0,1)$ for which, with probability strictly greater than 0, $\limsup_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(g_{n,\cdot}, f^\star_\kappa; n) > 0$, so that $g_{n,m}$ is not strongly universally consistent under $\mathbb{X}$. Since this argument applies to any self-adaptive learning rule $g_{n,m}$, this implies $\mathbb{X} \notin$ SUAL, and since the argument applies to any process $\mathbb{X}$ failing to satisfy Condition 3, we conclude that SUAL $\subseteq \mathcal{C}_3$, which completes the proof. ∎

To argue sufficiency of $\mathcal{C}_3$ for strong universal inductive learning, we propose a new type of learning rule, suitable for learning with unbounded losses under processes in $\mathcal{C}_3$. Specifically, let $\varepsilon_0 = \infty$, and for each $k \in \mathbb{N}$, let $\varepsilon_k = 2^{-k}$. Given a sequence $\{\tilde{f}_i\}_{i=1}^\infty$ of measurable functions $\mathcal{X} \to \mathcal{Y}$ (described below), any sequence $\{i_n\}_{n=1}^\infty$ in $\mathbb{N}$ with $i_n \to \infty$, and any $n \in \mathbb{N}$, $x_{1:n} \in \mathcal{X}^n$, and $y_{1:n} \in \mathcal{Y}^n$, define $\hat{i}_{n,0}(x_{1:n}, y_{1:n}) = 1$, and for each $k \in \mathbb{N}$, inductively define

$$\begin{aligned}\hat{i}_{n,k}(x_{1:n}, y_{1:n}) = \min \Bigg\{ i \in \{1, \ldots, i_n\} : \max_{1\leq t\leq n} \ell\Big(\tilde{f}_i(x_t), y_t\Big) \leq \varepsilon_k, \text{ and} \\ \sup_{x\in\mathcal{X}} \ell\Big(\tilde{f}_i(x), \tilde{f}_{\hat{i}_{n,k-1}(x_{1:n},y_{1:n})}(x)\Big) \leq \varepsilon_{k-1} + \varepsilon_k \Bigg\}, \quad (65)\end{aligned}$$

if it exists. For completeness, if the set on the right hand side of (65) is empty for a given $k \in \mathbb{N}$, let us define $\hat{i}_{n,k}(x_{1:n}, y_{1:n}) = \hat{i}_{n,k-1}(x_{1:n}, y_{1:n})$. Then, for any sequence $\{k_n\}_{n=1}^\infty$ in $\mathbb{N}$ with $k_n \to \infty$, for any $n \in \mathbb{N}$, and any $x_{1:n} \in \mathcal{X}^n$, $y_{1:n} \in \mathcal{Y}^n$, and $x \in \mathcal{X}$, define

$$\hat{f}_n(x_{1:n}, y_{1:n}, x) = \tilde{f}_{\hat{i}_{n,k_n}(x_{1:n},y_{1:n})}(x). \quad (66)$$

This defines an inductive learning rule (it is straightforward to verify that $\hat{f}_n$ is a measurable function). We will see below that, for an appropriate choice of the sequence $\{\tilde{f}_i\}_{i=1}^{\infty}$, this inductive learning rule is strongly universally consistent under every $\mathbb{X} \in \mathcal{C}_3$, even for unbounded losses. To specify an appropriate sequence $\{\tilde{f}_i\}_{i=1}^{\infty}$, and to study the performance of the resulting learning rule, we first prove modified versions of Lemmas 22 and 23, under the restriction of $\mathbb{X}$ to $\mathcal{C}_3$.

Lemma 50 *There exists a countable set $\mathcal{T}_1 \subseteq \mathcal{B}$ such that, $\forall \mathbb{X} \in \mathcal{C}_3$, $\forall A \in \mathcal{B}$, with probability one, $\exists \hat{A} \in \mathcal{T}_1$ s.t. $\mathbb{X} \cap \hat{A} = \mathbb{X} \cap A$.*

Proof This proof follows along similar lines to the proof of Lemma 22, and indeed the set $\mathcal{T}_1$ will be the same as defined in that proof. Let $\mathcal{T}_0$ be as in the proof of Lemma 22. As in the proof of Lemma 22, there is an immediate proof based on the monotone class theorem (Ash and Doléans-Dade, 2000, Theorem 1.3.9), by taking $\mathcal{T}_1$ as the algebra generated by $\mathcal{T}_0$ (which, one can show, is a countable set), and then showing that the collection of sets A for which the claim holds forms a monotone class (straightforwardly using Condition 3 for this part). However, as was the case for Lemma 22, we will instead establish the claim with a *smaller* set $\mathcal{T}_1$. Unlike Lemma 22, in this case this smaller $\mathcal{T}_1$ will not actually help to simplify the learning rule itself (as we will end up using the algebra generated by $\mathcal{T}_1$ anyway); but establishing Lemma 50 with this smaller $\mathcal{T}_1$ may be of independent interest. Specifically, as in the proof of Lemma 22, take $\mathcal{T}_1 = \{\bigcup \mathcal{A} : \mathcal{A} \subseteq \mathcal{T}_0, |\mathcal{A}| < \infty\}$, which (as discussed in that proof) is a countable set. Fix any $\mathbb{X} \in \mathcal{C}_3$, and let

$$\Lambda = \left\{ A \in \mathcal{B} : \mathbb{P}\left(\exists \hat{A} \in \mathcal{T}_1 \text{ s.t. } \mathbb{X} \cap \hat{A} = \mathbb{X} \cap A \right) = 1 \right\}.$$

For any $A \in \mathcal{T}$, as mentioned in the proof of Lemma 22, $\exists \{B_i\}_{i=1}^{\infty}$ in $\mathcal{T}_0$ such that $A = \bigcup_{i=1}^{\infty} B_i$. Then letting $A_k = \bigcup_{i=1}^{k} B_i$ for each $k \in \mathbb{N}$, we have $A_k \triangle A = A \setminus A_k \downarrow \emptyset$ (monotonically), and $A_k \in \mathcal{T}_1$ for each $k \in \mathbb{N}$. Therefore, by Condition 3, with probability one, $\exists k \in \mathbb{N}$ such that $\mathbb{X} \cap (A_k \triangle A) = \emptyset$, which implies $\mathbb{X} \cap A_k = \mathbb{X} \cap A$. Thus, $A \in \Lambda$. Since this holds for any $A \in \mathcal{T}$, we have $\mathcal{T} \subseteq \Lambda$.

Next, we argue that Λ is a σ-algebra, beginning with the property of being closed under complements. First, consider any $A \in \mathcal{T}_1$. Since $\mathcal{T}_1 \subseteq \mathcal{T}$, it follows that $\mathcal{X} \setminus A$ is a closed set. Since $(\mathcal{X}, \mathcal{T})$ is metrizable, this implies $\exists \{B_i\}_{i=1}^{\infty}$ in $\mathcal{T}$ such that $\mathcal{X} \setminus A = \bigcap_{i=1}^{\infty} B_i$ (Kechris, 1995, Proposition 3.7). Denoting $C_k = \bigcap_{i=1}^{k} B_i$ for each $k \in \mathbb{N}$, we have that $C_k \triangle (\mathcal{X} \setminus A) = C_k \setminus (\mathcal{X} \setminus A) \downarrow \emptyset$ (monotonically), and $C_k \in \mathcal{T}$ for each $k \in \mathbb{N}$. In particular, by Condition 3, this implies that on an event $E_0^{(A)}$ of probability one, there exists $k_0 \in \mathbb{N}$ such that $\mathbb{X} \cap (C_{k_0} \triangle (\mathcal{X} \setminus A)) = \emptyset$, which implies $\mathbb{X} \cap C_{k_0} = \mathbb{X} \cap (\mathcal{X} \setminus A)$. Furthermore, for each $k \in \mathbb{N}$, since $C_k \in \mathcal{T} \subseteq \Lambda$, there is an event $E_k^{(A)}$ of probability one, on which $\exists \hat{A}_k \in \mathcal{T}_1$ with $\mathbb{X} \cap \hat{A}_k = \mathbb{X} \cap C_k$. Altogether, on the event $\bigcap_{k=0}^{\infty} E_k^{(A)}$ (which has probability one, by the union bound), $\mathbb{X} \cap \hat{A}_{k_0} = \mathbb{X} \cap (\mathcal{X} \setminus A)$. Now denote $E^{(\mathcal{T}_1)} = \bigcap_{A \in \mathcal{T}_1} \bigcap_{k=0}^{\infty} E_k^{(A)}$, which has probability one by the union bound (since $\mathcal{T}_1$ is countable).

Next, consider any $A \in \Lambda$, and suppose the event E' (of probability one) holds that $\exists \hat{A} \in \mathcal{T}_1$ s.t. $\mathbb{X} \cap \hat{A} = \mathbb{X} \cap A$, which also implies $\mathbb{X} \cap (\mathcal{X} \setminus \hat{A}) = \mathbb{X} \cap (\mathcal{X} \setminus A)$. Since $\hat{A} \in \mathcal{T}_1$, on the event $E^{(\mathcal{T}_1)}$ we have that $\exists \hat{A}' \in \mathcal{T}_1$ with $\mathbb{X} \cap \hat{A}' = \mathbb{X} \cap (\mathcal{X} \setminus \hat{A})$. Thus, on the event $E' \cap E^{(\mathcal{T}_1)}$, we have $\mathbb{X} \cap \hat{A}' = \mathbb{X} \cap (\mathcal{X} \setminus A)$. Since $E' \cap E^{(\mathcal{T}_1)}$ has probability one (by the union bound), we have that $\mathcal{X} \setminus A \in \Lambda$. Since this argument holds for any $A \in \Lambda$, we have that Λ is closed under complements.

Next, we show that Λ is closed under countable unions. Let $\{A_i\}_{i=1}^{\infty}$ be a sequence in Λ, and denote $A = \bigcup_{i=1}^{\infty} A_i$. Since each $A_i \in \Lambda$, by the union bound there is an event E of probability one, on which there exists a sequence $\{\hat{A}_i\}_{i=1}^{\infty}$ in $\mathcal{T}_1$ such that $\forall i \in \mathbb{N}$, $\mathbb{X} \cap A_i = \mathbb{X} \cap \hat{A}_i$. Furthermore, since $A \triangle \bigcup_{i=1}^{k} A_i = A \setminus \bigcup_{i=1}^{k} A_i \downarrow \emptyset$ (monotonically), Condition 3 implies that, on an event E'' of probability one, $\exists k \in \mathbb{N}$ such that $\mathbb{X} \cap \left(A \triangle \bigcup_{i=1}^{k} A_i \right) = \emptyset$, which implies $\mathbb{X} \cap \bigcup_{i=1}^{k} A_i = \mathbb{X} \cap A$. Since, for any $k \in \mathbb{N}$, $\mathbb{X} \cap \bigcup_{i=1}^{k} A_i$ is simply the subsequence of $\mathbb{X}$ consisting of all entries appearing in any of the $\mathbb{X} \cap A_i$ subsequences with $i \leq k$, and (on E) each $\mathbb{X} \cap A_i = \mathbb{X} \cap \hat{A}_i$, together we have that on the event $E \cap E''$ (which has probability one, by the union bound), $\exists k \in \mathbb{N}$ such that $\mathbb{X} \cap \bigcup_{i=1}^{k} \hat{A}_i = \mathbb{X} \cap \bigcup_{i=1}^{k} A_i = \mathbb{X} \cap A$. Since it follows immediately from its definition that the set $\mathcal{T}_1$ is closed under finite unions, we have that $\bigcup_{i=1}^{k} \hat{A}_i \in \mathcal{T}_1$. Therefore, $A \in \Lambda$. Since this holds for any choice of the sequence $\{A_i\}_{i=1}^{\infty}$ in Λ, we have that Λ is closed under countable unions.

Finally, recalling that $\mathcal{T}$ is a topology, we have $\mathcal{X} \in \mathcal{T}$, and since $\mathcal{T} \subseteq \Lambda$, this implies $\mathcal{X} \in \Lambda$. Altogether, we have established that Λ is a σ-algebra. Therefore, since $\mathcal{B}$ is the σ-algebra generated by $\mathcal{T}$, and $\mathcal{T} \subseteq \Lambda$, it immediately follows that $\mathcal{B} \subseteq \Lambda$ (which also implies $\Lambda = \mathcal{B}$). Since this argument holds for any choice of $\mathbb{X} \in \mathcal{C}_3$, the lemma immediately follows. ■

Lemma 51 *There exists a sequence $\{\tilde{f}_i\}_{i=1}^{\infty}$ of measurable functions $\mathcal{X} \to \mathcal{Y}$ such that, for every $\mathbb{X} \in \mathcal{C}_3$, for every measurable $f : \mathcal{X} \to \mathcal{Y}$, with probability one, $\forall \varepsilon > 0$, $\forall i \in \mathbb{N}$, $\exists j \in \mathbb{N}$ with*

$$\sup_{x \in \mathcal{X}} \ell\Big(\tilde{f}_j(x), \tilde{f}_i(x)\Big) \leq \sup_{t \in \mathbb{N}} \ell\Big(\tilde{f}_i(X_t), f(X_t)\Big) + \varepsilon$$

$$\text{and} \quad \sup_{t \in \mathbb{N}} \ell\Big(\tilde{f}_j(X_t), f(X_t)\Big) \leq \varepsilon.$$

Proof Let $\mathcal{T}_1$ be as in Lemma 50, and let $\tilde{y}_i$ and $B_{\varepsilon,i}$ be defined as in the proof of Lemma 23, for each $i \in \mathbb{N}$ and $\varepsilon > 0$. Also fix an arbitrary value $\tilde{y}_0 \in \mathcal{Y}$. Let $\mathcal{T}_2$ denote the algebra generated by $\mathcal{T}_1$. Since $\mathcal{T}_1$ is countable, one can easily verify that $\mathcal{T}_2$ is countable as well (see e.g., Bogachev, 2007, page 5), and by definition, has $\mathcal{T}_1 \subseteq \mathcal{T}_2$. Furthermore, since $\mathcal{T}_1 \subseteq \mathcal{B}$ and $\mathcal{B}$ is an algebra, minimality of the algebra $\mathcal{T}_2$ implies $\mathcal{T}_2 \subseteq \mathcal{B}$ (see e.g., Dudley, 2003, page 86). Now for each $k_0 \in \mathbb{N}$ and disjoint sets $A_1, \ldots, A_{k_0} \in \mathcal{B}$, let $A_0 = \mathcal{X} \setminus \bigcup_{k=1}^{k_0} A_k$, and for any $x \in \mathcal{X}$, define $\tilde{f}(x; \{A_k\}_{k=1}^{k_0}) = \tilde{y}_k$ for the unique value $k \in \{0, \ldots, k_0\}$ with $x \in A_k$. One can easily verify that $\tilde{f}(\cdot; \{A_k\}_{k=1}^{k_0})$ is a measurable function. Now define $\tilde{\mathcal{F}}$ as the set of all functions $\tilde{f}(\cdot; \{A_k\}_{k=1}^{k_0})$ with $k_0 \in \mathbb{N}$ and $A_1, \ldots, A_{k_0}$ disjoint elements of $\mathcal{T}_2$. Note that, given an indexing of $\mathcal{T}_2$ by $\mathbb{N}$, we can index $\tilde{\mathcal{F}}$ by finite tuples of integers (the indices of the corresponding A_i sets within $\mathcal{T}_2$), of which there are countably many, so that $\tilde{\mathcal{F}}$ is countable. We may therefore enumerate the elements of $\tilde{\mathcal{F}}$ as $\tilde{f}_1, \tilde{f}_2, \ldots$. For simplicity, we will suppose this sequence is infinite (which can always be achieved by repetition, if necessary).

Fix any $\mathbb{X} \in \mathcal{C}_3$, any measurable $f : \mathcal{X} \to \mathcal{Y}$, and any $\varepsilon > 0$. For each $k \in \mathbb{N}$, define $C_k = f^{-1}(B_{\varepsilon,k})$. Since $\lim_{k_1 \to \infty} \bigcup_{k=k_1}^{\infty} C_k = \emptyset$ and $\mathbb{X} \in \mathcal{C}_3$, on an event $E_{\varepsilon,1}$ of probability one, $\exists k_1 \in \mathbb{N}$ s.t. $\mathbb{X} \cap \bigcup_{k=k_1+1}^{\infty} C_k = \emptyset$. Furthermore, by the union bound and the defining property of $\mathcal{T}_1$ from Lemma 50, on an event $E_{\varepsilon,2}$ of probability one, $\forall k \in \mathbb{N}$, $\exists A'_k \in \mathcal{T}_1$ with $\mathbb{X} \cap A'_k = \mathbb{X} \cap C_k$. Now note that for any $k, k' \in \mathbb{N}$, $\mathbb{X} \cap (A'_k \cap A'_{k'})$ is simply the subsequence of $\mathbb{X}$ consisting of entries X_t appearing in *both* subsequences $\mathbb{X} \cap A'_k$ and $\mathbb{X} \cap A'_{k'}$. Thus, since the sets $\{C_k\}_{k=1}^{\infty}$ are disjoint, and on the event $E_{\varepsilon,2}$ every $k \in \mathbb{N}$ has $\mathbb{X} \cap A'_k = \mathbb{X} \cap C_k$, we have that on $E_{\varepsilon,2}$ every $k, k' \in \mathbb{N}$ with $k \neq k'$ satisfy $\mathbb{X} \cap (A'_k \cap A'_{k'}) = \mathbb{X} \cap (C_k \cap C_{k'}) = \emptyset$, so that any $k \in \mathbb{N} \setminus \{1\}$ has $\mathbb{X} \cap \bigcup_{k'=1}^{k-1} \left(A'_k \cap A'_{k'}\right) = \emptyset$. Therefore, on the event $E_{\varepsilon,2}$, defining $A_1 = A'_1$, and $A_k = A'_k \setminus \bigcup_{k'=1}^{k-1} A'_{k'} = A'_k \setminus \bigcup_{k'=1}^{k-1} \left(A'_k \cap A'_{k'}\right)$ for every $k \in \mathbb{N} \setminus \{1\}$, we have that $\forall k \in \mathbb{N}$, $\mathbb{X} \cap A_k = \mathbb{X} \cap A'_k = \mathbb{X} \cap C_k$. Note that the sets $\{A_k\}_{k=1}^{\infty}$ are disjoint elements of $\mathcal{T}_2$.

Now fix any $i \in \mathbb{N}$, and let $k_2 \in \mathbb{N}$ and $\{A''_k\}_{k=1}^{k_2} \in \mathcal{T}_2^{k_2}$ (disjoint) be such that $\tilde{f}_i(\cdot) = \tilde{f}\left(\cdot; \{A''_k\}_{k=1}^{k_2}\right)$. For simplicity, denote $k_0 = \max\{k_1, k_2\}$, and (if $k_0 > k_2$) for any $k \in \{k_2 + 1, \ldots, k_0\}$ define $A''_k = \emptyset$; in particular, $\tilde{f}_i(\cdot) = \tilde{f}\left(\cdot; \{A''_k\}_{k=1}^{k_0}\right)$ as well. Also define $A''_0 = \mathcal{X} \setminus \bigcup_{k=1}^{k_0} A''_k$ and $A_0 = \mathcal{X} \setminus \bigcup_{k=1}^{k_0} A_k$. Now for each $k \in \{1, \ldots, k_0\}$, define

$$\begin{aligned}\tilde{A}_k = & \bigcup \left\{ A_k \cap A''_{k'} : \mathbb{X} \cap \left(A_k \cap A''_{k'}\right) \neq \emptyset, k' \in \{0, \ldots, k_0\} \right\} \cup \\ & \bigcup \left\{ A_{k'} \cap A''_k : \mathbb{X} \cap \left(A_{k'} \cap A''_k\right) = \emptyset, k' \in \{0, \ldots, k_0\} \right\}.\end{aligned}$$

Note that $\tilde{A}_1, \ldots, \tilde{A}_{k_0}$ are elements of $\mathcal{T}_2$. Furthermore, disjointness of the sets $\{A_k\}_{k=0}^{k_0}$, and disjointness of the sets $\{A''_k\}_{k=0}^{k_0}$, together imply that the sets $\{A_k \cap A''_{k'}, k, k' \in \{0, \ldots, k_0\}\}$ are disjoint. From this and the definition of the $\tilde{A}_k$ sets, it easily follows that the sets $\{\tilde{A}_k\}_{k=1}^{k_0}$ are disjoint. Thus, on the event $E_{\varepsilon,1} \cap E_{\varepsilon,2}$, $\exists j \in \mathbb{N}$ such that $\tilde{f}_j(\cdot) = \tilde{f}\left(\cdot; \{\tilde{A}_k\}_{k=1}^{k_0}\right)$.

Now suppose the event $E_{\varepsilon,1} \cap E_{\varepsilon,2}$ holds and that we have constructed this function $\tilde{f}_j$ as above. Then for every $k \in \{1, \ldots, k_0\}$, since

$$\mathbb{X} \cap \bigcup \left\{ A_{k'} \cap A''_k : \mathbb{X} \cap \left(A_{k'} \cap A''_k\right) = \emptyset, k' \in \{0, \ldots, k_0\} \right\} = \emptyset,$$

we have

$$\mathbb{X} \cap \tilde{A}_k = \mathbb{X} \cap \bigcup \left\{ A_k \cap A''_{k'} : \mathbb{X} \cap \left(A_k \cap A''_{k'} \right) \neq \emptyset, k' \in \{0, \ldots, k_0\} \right\},$$

and since

$$\mathbb{X} \cap \bigcup \left\{ A_k \cap A''_{k'} : \mathbb{X} \cap \left(A_k \cap A''_{k'} \right) = \emptyset, k' \in \{0, \ldots, k_0\} \right\} = \emptyset,$$

this also implies

$$\mathbb{X} \cap \tilde{A}_k = \mathbb{X} \cap \bigcup \left\{ A_k \cap A''_{k'} : k' \in \{0, \ldots, k_0\} \right\} = \mathbb{X} \cap \left(A_k \cap \bigcup_{k'=0}^{k_0} A''_{k'} \right) = \mathbb{X} \cap A_k = \mathbb{X} \cap C_k. \tag{67}$$

In particular, since $k_0 \geq k_1$, this implies that every $t \in \mathbb{N}$ has X_t contained in exactly one set $\tilde{A}_k$ with $k \in \{1, \ldots, k_0\}$. Therefore,

$$\sup_{t \in \mathbb{N}} \ell\Big(\tilde{f}_j(X_t), f(X_t) \Big) = \sup_{t \in \mathbb{N}} \sum_{k=1}^{k_0} \mathbb{1}_{\tilde{A}_k}(X_t) \ell\Big(\tilde{f}_j(X_t), f(X_t) \Big).$$

By the definition of $\tilde{f}_j$, every $X_t \in \tilde{A}_k$ has $\tilde{f}_j(X_t) = \tilde{y}_k$, so that the above equals

$$\sup_{t \in \mathbb{N}} \sum_{k=1}^{k_0} \mathbb{1}_{\tilde{A}_k}(X_t) \ell(\tilde{y}_k, f(X_t)).$$

Furthermore, (67) implies $\mathbb{1}_{\tilde{A}_k}(X_t) = \mathbb{1}_{C_k}(X_t)$ for every $t \in \mathbb{N}$. Therefore, the above expression equals

$$\sup_{t \in \mathbb{N}} \sum_{k=1}^{k_0} \mathbb{1}_{C_k}(X_t) \ell(\tilde{y}_k, f(X_t)).$$

By the definition of C_k, every $X_t \in C_k$ has $\ell(\tilde{y}_k, f(X_t)) \leq \varepsilon$, so that the above is at most

$$\sup_{t \in \mathbb{N}} \sum_{k=1}^{k_0} \mathbb{1}_{C_k}(X_t) \varepsilon = \varepsilon.$$

Altogether, we have that

$$\sup_{t \in \mathbb{N}} \ell\Big(\tilde{f}_j(X_t), f(X_t) \Big) \leq \varepsilon.$$

Next, continuing to suppose $E_{\varepsilon,1} \cap E_{\varepsilon,2}$ holds and that $\tilde{f}_j$ is as above, fix any $x \in \mathcal{X}$. First, consider the case of $x \notin \bigcup_{k=1}^{k_0} \tilde{A}_k$. By the definition of $\tilde{f}_j$, we have $\tilde{f}_j(x) = \tilde{y}_0$. Also note that every $k, k' \in \{1, \ldots, k_0\}$ have $A_k \cap A''_{k'} \subseteq \tilde{A}_k \cup \tilde{A}_{k'}$, so that $x \notin A_k \cap A''_{k'}$ for every such k, k'. Furthermore, since $k_0 \geq k_1$ and $\mathbb{X} \cap A_k = \mathbb{X} \cap C_k$ for every $k \in \mathbb{N}$, we have that $\mathbb{X} \cap A_0 = \emptyset$. It follows (from the definition of $\tilde{A}_k$) that $A_0 \cap A''_k \subseteq \tilde{A}_k$ for every $k \in \{1, \ldots, k_0\}$, so that $x \notin A_0 \cap A''_k$ for every such k. Since $\bigcup_{k,k' \in \{0,\ldots,k_0\}} A_k \cap A''_{k'} = \mathcal{X}$, the only remaining possibility is that $\exists k \in \{0, \ldots, k_0\}$ with $x \in A_k \cap A''_0$. In particular, since this implies $x \in A''_0$, the definition of $\tilde{f}_i$ implies $\tilde{f}_i(x) = \tilde{y}_0$, so that $\ell\Big(\tilde{f}_j(x), \tilde{f}_i(x) \Big) = \ell(\tilde{y}_0, \tilde{y}_0) = 0$.

Next, consider the remaining case, in which $\exists k \in \{1, \ldots, k_0\}$ with $x \in \tilde{A}_k$. Now there are two subcases to consider. In the first subcase, $\exists k' \in \{0, \ldots, k_0\}$ such that $x \in A_{k'} \cap A''_k$ and $\mathbb{X} \cap (A_{k'} \cap A''_k) = \emptyset$. In this case, since $x \in \tilde{A}_k$, we have $\tilde{f}_j(x) = \tilde{y}_k$, and since $x \in A''_k$, we have $\tilde{f}_i(x) = \tilde{y}_k$ as well. Therefore, $\ell\Big(\tilde{f}_j(x), \tilde{f}_i(x)\Big) = \ell(\tilde{y}_k, \tilde{y}_k) = 0$. In the second (and only remaining) subcase, we have that $\exists k' \in \{0, \ldots, k_0\}$ such that $x \in A_k \cap A''_{k'}$ and $\mathbb{X} \cap \big(A_k \cap A''_{k'}\big) \neq \emptyset$. In this case, by the definitions of $\tilde{f}_i$ and $\tilde{f}_j$, we have that $\tilde{f}_j(x) = \tilde{y}_k$ (due to $x \in \tilde{A}_k$) and $\tilde{f}_i(x) = \tilde{y}_{k'}$ (due to $x \in A''_{k'}$). Also, since $\mathbb{X} \cap \big(A_k \cap A''_{k'}\big) \neq \emptyset$, we have that $\exists t_k \in \mathbb{N}$ with $X_{t_k} \in A_k \cap A''_{k'} \subseteq \tilde{A}_k$; in particular, this implies $\tilde{f}_i(X_{t_k}) = \tilde{y}_{k'}$. Furthermore, (67) implies $X_{t_k} \in C_k$, so that $\ell(f(X_{t_k}), \tilde{y}_k) \leq \varepsilon$. Together with the triangle inequality, we have that

$$\begin{aligned}\ell\Big(\tilde{f}_j(x), \tilde{f}_i(x)\Big) &= \ell(\tilde{y}_k, \tilde{y}_{k'}) \leq \ell(f(X_{t_k}), \tilde{y}_k) + \ell(\tilde{y}_{k'}, f(X_{t_k})) \\ &\leq \varepsilon + \ell\Big(\tilde{f}_i(X_{t_k}), f(X_{t_k})\Big) \leq \varepsilon + \sup_{t \in \mathbb{N}} \ell\Big(\tilde{f}_i(X_t), f(X_t)\Big).\end{aligned}$$

Since the above arguments together cover *every* $x \in \mathcal{X}$, we have that, on $E_{\varepsilon,1} \cap E_{\varepsilon,2}$,

$$\sup_{x \in \mathcal{X}} \ell\Big(\tilde{f}_j(x), \tilde{f}_i(x)\Big) \leq \varepsilon + \sup_{t \in \mathbb{N}} \ell\Big(\tilde{f}_i(X_t), f(X_t)\Big).$$

The above results hold for any fixed $\varepsilon > 0$. Now letting $\varepsilon_k = 2^{-k}$ for each $k \in \mathbb{N}$, we have that on the event $\bigcap_{k=1}^{\infty} (E_{\varepsilon_k,1} \cap E_{\varepsilon_k,2})$, for any $i \in \mathbb{N}$ and any $\varepsilon > 0$, letting $k = \lceil \log_2((1/\varepsilon) \vee 2) \rceil$, we have that $\exists j \in \mathbb{N}$ with

$$\sup_{t \in \mathbb{N}} \ell(\tilde{f}_j(X_t), f(X_t)) \leq \varepsilon_k \leq \varepsilon,$$

and

$$\sup_{x \in \mathcal{X}} \ell(\tilde{f}_j(x), \tilde{f}_i(x)) \leq \varepsilon_k + \sup_{t \in \mathbb{N}} \ell(\tilde{f}_i(X_t), f(X_t)) \leq \varepsilon + \sup_{t \in \mathbb{N}} \ell(\tilde{f}_i(X_t), f(X_t)).$$

Noting that the event $\bigcap_{k=1}^{\infty} (E_{\varepsilon_k,1} \cap E_{\varepsilon_k,2})$ has probability one (by the union bound) completes the proof. ■

We are now ready to present a result analogous to Lemma 25, showing that any process satisfying Condition 3 necessarily admits strong universal inductive learning. For clarity, we make explicit the fact that this result holds for $\bar{\ell} = \infty$, though it clearly also holds for $\bar{\ell} < \infty$ (since $\mathcal{C}_3 \subseteq \mathcal{C}_1$).

Lemma 52 *When* $\bar{\ell} = \infty$, $\mathcal{C}_3 \subseteq \mathrm{SUIL} \cap \mathrm{SUOL}$.

Proof We begin by showing that $\mathcal{C}_3 \subseteq \mathrm{SUIL}$. Let $\hat{f}_n$ be the inductive learning rule specified by (66), where the sequence $\{\tilde{f}_i\}_{i=1}^{\infty}$ is chosen as in Lemma 51, and $\{i_n\}_{n=1}^{\infty}$ and $\{k_n\}_{n=1}^{\infty}$ are arbitrary (nonrandom) sequences in $\mathbb{N}$ with $i_n \to \infty$ and $k_n \to \infty$. We establish the stated result by arguing that $\hat{f}_n$ is strongly universally consistent for every $\mathbb{X} \in \mathcal{C}_3$, which thereby establishes that every $\mathbb{X} \in \mathcal{C}_3$ admits strong universal inductive learning.

Toward this end, fix any $\mathbb{X} \in \mathcal{C}_3$ and any measurable function $f^\star : \mathcal{X} \to \mathcal{Y}$. To simplify the notation, let us abbreviate $\hat{i}_{n,k} = \hat{i}_{n,k}(X_{1:n}, f^\star(X_{1:n}))$ for every $n \in \mathbb{N}$ and $k \in \mathbb{N} \cup \{0\}$. Let E denote the event of probability one guaranteed by Lemma 51, for the process $\mathbb{X}$ and the function $f = f^\star$: that is, on E, $\forall \varepsilon > 0$, $\forall i \in \mathbb{N}$, $\exists j \in \mathbb{N}$ with

$$\sup_{x \in \mathcal{X}} \ell\Big(\tilde{f}_j(x), \tilde{f}_i(x)\Big) \leq \sup_{t \in \mathbb{N}} \ell\Big(\tilde{f}_i(X_t), f^\star(X_t)\Big) + \varepsilon \tag{68}$$

$$\text{and} \quad \sup_{t \in \mathbb{N}} \ell\Big(\tilde{f}_j(X_t), f^\star(X_t)\Big) \leq \varepsilon. \tag{69}$$

Let us suppose this event E occurs.

We now argue by induction that, $\forall k \in \mathbb{N} \cup \{0\}$, $\exists i_k^*, n_k^* \in \mathbb{N}$ such that, $\forall n \geq n_k^*$, $\hat{i}_{n,k} = i_k^*$ and $\sup_{t \in \mathbb{N}} \ell\Big(\tilde{f}_{i_k^*}(X_t), f^\star(X_t)\Big) \leq \varepsilon_k$. As a base case, the result is trivially satisfied for $k = 0$, since taking $i_0^* = 1$, we have defined $\hat{i}_{n,0} = i_0^*$ for every $n \in \mathbb{N}$, so that we may take $n_0^* = 1$; moreover, $\varepsilon_0 = \infty$, so that we trivially have $\sup_{t \in \mathbb{N}} \ell\Big(\tilde{f}_{i_0^*}(X_t), f^\star(X_t)\Big) \leq \varepsilon_0$.

Now take as an inductive hypothesis that, for some $k \in \mathbb{N}$, $\exists i_{k-1}^*, n_{k-1}^* \in \mathbb{N}$ such that, $\forall n \geq n_{k-1}^*$, $\hat{i}_{n,k-1} = i_{k-1}^*$ and $\sup_{t \in \mathbb{N}} \ell\Big(\tilde{f}_{i_{k-1}^*}(X_t), f^\star(X_t)\Big) \leq \varepsilon_{k-1}$. Then define

$$i_k^* = \min\left\{ j \in \mathbb{N} : \sup_{t \in \mathbb{N}} \ell\Big(\tilde{f}_j(X_t), f^\star(X_t)\Big) \leq \varepsilon_k \text{ and } \sup_{x \in \mathcal{X}} \ell\Big(\tilde{f}_j(x), \tilde{f}_{i_{k-1}^*}(x)\Big) \leq \varepsilon_{k-1} + \varepsilon_k \right\}.$$

Note that, taking $\varepsilon = \varepsilon_k$ and $i = i_{k-1}^*$ in (68) and (69), and combining with the fact (from the inductive hypothesis) that $\sup_{t \in \mathbb{N}} \ell\Big(\tilde{f}_{i_{k-1}^*}(X_t), f^\star(X_t)\Big) \leq \varepsilon_{k-1}$, we can conclude that the set of values j on the right hand side of the definition of i_k^* is nonempty, so that i_k^* is a well-defined element of $\mathbb{N}$. In particular, by definition, we have $\sup_{t \in \mathbb{N}} \ell\Big(\tilde{f}_{i_k^*}(X_t), f^\star(X_t)\Big) \leq \varepsilon_k$. Next note that, by minimality of i_k^*, for every $j \in \mathbb{N}$ with $j < i_k^*$ and $\sup_{x \in \mathcal{X}} \ell\Big(\tilde{f}_j(x), \tilde{f}_{i_{k-1}^*}(x)\Big) \leq \varepsilon_{k-1} + \varepsilon_k$ (if any such j exists), we have $\sup_{t \in \mathbb{N}} \ell\Big(\tilde{f}_j(X_t), f^\star(X_t)\Big) > \varepsilon_k$, so that $\exists t_{j,k} \in \mathbb{N}$ such that $\ell\Big(\tilde{f}_j(X_{t_{j,k}}), f^\star(X_{t_{j,k}})\Big) > \varepsilon_k$. Furthermore, since $i_n \to \infty$, $\exists n_k' \in \mathbb{N}$ such that $\min_{n \geq n_k'} i_n \geq i_k^*$. Now define

$$n_k^* = \max\left(\left\{ t_{j,k} : j \in \{1, \ldots, i_k^* - 1\}, \sup_{x \in \mathcal{X}} \ell\Big(\tilde{f}_j(x), \tilde{f}_{i_{k-1}^*}(x)\Big) \leq \varepsilon_{k-1} + \varepsilon_k \right\} \cup \left\{ n_k', n_{k-1}^* \right\}\right),$$

which (being a maximum of a finite subset of $\mathbb{N}$) is a finite positive integer. In particular, note that (since $n_k^* \geq n_{k-1}^*$) for any $n \geq n_k^*$, the inductive hypothesis implies $\hat{i}_{n,k-1} = i_{k-1}^*$. Additionally, for any $n \geq n_k^*$, every $j \in \mathbb{N}$ with $j < i_k^*$ and $\sup_{x \in \mathcal{X}} \ell\Big(\tilde{f}_j(x), \tilde{f}_{i_{k-1}^*}(x)\Big) \leq \varepsilon_{k-1} + \varepsilon_k$ has $\max_{1 \leq t \leq n} \ell\Big(\tilde{f}_j(X_t), f^\star(X_t)\Big) \geq \ell\Big(\tilde{f}_j(X_{t_{j,k}}), f^\star(X_{t_{j,k}})\Big) > \varepsilon_k$. In particular, this means that any such j is not included in the set on the right hand side of (65) (when $x_{1:n} = X_{1:n}$ and $y_{1:n} = f^\star(X_{1:n})$). Furthermore, for $n \geq n_k^*$, every $j \in \mathbb{N}$ with $j < i_k^*$ and

$\sup_{x \in \mathcal{X}} \ell\Big(\tilde{f}_j(x), \tilde{f}_{i^*_{k-1}}(x)\Big) > \varepsilon_{k-1} + \varepsilon_k$ is clearly also not included in the set on the right hand side of (65) in this case (again, since $\hat{i}_{n,k-1} = i^*_{k-1}$). On the other hand, by definition we have $\sup_{x \in \mathcal{X}} \ell\Big(\tilde{f}_{i^*_k}(x), \tilde{f}_{i^*_{k-1}}(x)\Big) \leq \varepsilon_{k-1} + \varepsilon_k$, and

$$\max_{1 \leq t \leq n} \ell\Big(\tilde{f}_{i^*_k}(X_t), f^\star(X_t)\Big) \leq \sup_{t \in \mathbb{N}} \ell\Big(\tilde{f}_{i^*_k}(X_t), f^\star(X_t)\Big) \leq \varepsilon_k,$$

so that, since any $n \geq n^*_k$ has $i_n \geq i^*_k$ (due to $n^*_k \geq n'_k$) and $\hat{i}_{n,k-1} = i^*_{k-1}$ (as argued above), we have that i^*_k *is* included in the set on the right hand side of (65) (again with $x_{1:n} = X_{1:n}$ and $y_{1:n} = f^\star(X_{1:n})$). Together with the definition of $\hat{i}_{n,k}$, these observations imply that, for any $n \geq n^*_k$, $\hat{i}_{n,k} = i^*_k$.

By the principle of induction, we have established the existence of a sequence $\{n^*_k\}_{k=0}^\infty$ in $\mathbb{N}$ such that, $\forall k \in \mathbb{N} \cup \{0\}$, $\forall n \in \mathbb{N}$ with $n \geq n^*_k$, we have $\sup_{t \in \mathbb{N}} \ell\Big(\tilde{f}_{\hat{i}_{n,k}}(X_t), f^\star(X_t)\Big) \leq \varepsilon_k$. Now for any $n \in \mathbb{N}$, let $k^*_n = \max\{k \in \{0, \ldots, k_n\} : n \geq n^*_k\}$ (recalling that we defined $n^*_0 = 1$ above, so that k^*_n always exists). Note that, by the above guarantee,

$$\sup_{t \in \mathbb{N}} \ell\Big(\tilde{f}_{\hat{i}_{n,k^*_n}}(X_t), f^\star(X_t)\Big) \leq \varepsilon_{k^*_n}. \tag{70}$$

Furthermore, since $k_n \to \infty$, and each n^*_k is finite, we have that $k^*_n \to \infty$.

Note that, by definition, for each $k \in \{1, \ldots, k_n\}$, we have $\sup_{x \in \mathcal{X}} \ell\Big(\tilde{f}_{\hat{i}_{n,k}}(x), \tilde{f}_{\hat{i}_{n,k-1}}(x)\Big) \leq \varepsilon_{k-1} + \varepsilon_k$ (noting that this is true even when the set on the right hand side of (65) is empty, by our choice to define $\hat{i}_{n,k} = \hat{i}_{n,k-1}$ in that case). Combining this with an inductive application of the triangle inequality and subadditivity of the supremum, and noting that $k^*_n \leq k_n$ (by definition), this implies

$$\begin{aligned}
\sup_{x \in \mathcal{X}} \ell\Big(\tilde{f}_{\hat{i}_{n,k_n}}(x), \tilde{f}_{\hat{i}_{n,k^*_n}}(x)\Big) &\leq \sup_{x \in \mathcal{X}} \sum_{k=k^*_n+1}^{k_n} \ell\Big(\tilde{f}_{\hat{i}_{n,k}}(x), \tilde{f}_{\hat{i}_{n,k-1}}(x)\Big) \\
\leq \sum_{k=k^*_n+1}^{k_n} \sup_{x \in \mathcal{X}} \ell\Big(\tilde{f}_{\hat{i}_{n,k}}(x), \tilde{f}_{\hat{i}_{n,k-1}}(x)\Big) &\leq \sum_{k=k^*_n+1}^{k_n} (\varepsilon_{k-1} + \varepsilon_k) \leq \sum_{k=k^*_n+1}^{\infty} (\varepsilon_{k-1} + \varepsilon_k).
\end{aligned}$$

If $k^*_n = 0$, this rightmost expression is $\infty = 3\varepsilon_0$; on the other hand, if $k^*_n \geq 1$, then by our choice of $\varepsilon_k = 2^{-k}$ for every $k \in \mathbb{N}$, the rightmost expression above equals $3 \cdot 2^{-k^*_n} = 3\varepsilon_{k^*_n}$. Thus, either way, we have

$$\sup_{x \in \mathcal{X}} \ell\Big(\tilde{f}_{\hat{i}_{n,k_n}}(x), \tilde{f}_{\hat{i}_{n,k^*_n}}(x)\Big) \leq 3\varepsilon_{k^*_n}. \tag{71}$$

Therefore, by the triangle inequality, $\forall n \in \mathbb{N}$,

$$\begin{aligned}
\sup_{t \in \mathbb{N}} \ell\Big(\tilde{f}_{\hat{i}_{n,k_n}}(X_t), f^\star(X_t)\Big) &\leq \sup_{t \in \mathbb{N}} \Big(\ell\Big(\tilde{f}_{\hat{i}_{n,k^*_n}}(X_t), f^\star(X_t)\Big) + \ell\Big(\tilde{f}_{\hat{i}_{n,k_n}}(X_t), \tilde{f}_{\hat{i}_{n,k^*_n}}(X_t)\Big)\Big) \\
&\leq \sup_{t \in \mathbb{N}} \ell\Big(\tilde{f}_{\hat{i}_{n,k^*_n}}(X_t), f^\star(X_t)\Big) + \sup_{x \in \mathcal{X}} \ell\Big(\tilde{f}_{\hat{i}_{n,k_n}}(x), \tilde{f}_{\hat{i}_{n,k^*_n}}(x)\Big) \leq 4\varepsilon_{k^*_n},
\end{aligned}$$

where the last inequality is due to (70) and (71). Since $k_n^* \to \infty$ and $\varepsilon_k \to 0$, and since $\hat{f}_n(X_{1:n}, f^\star(X_{1:n}), \cdot) = \tilde{f}_{\hat{i}_{n,k_n}}(\cdot)$ by its definition in (66), we may conclude that

$$\sup_{t \in \mathbb{N}} \ell\Big(\hat{f}_n(X_{1:n}, f^\star(X_{1:n}), X_t), f^\star(X_t)\Big) \to 0.$$

Since all of the above claims hold on the event E, which has probability one, and since the above argument holds for *any* choice of measurable function $f^\star : \mathcal{X} \to \mathcal{Y}$, we may conclude that, for any measurable $f^\star : \mathcal{X} \to \mathcal{Y}$,

$$\sup_{t \in \mathbb{N}} \ell\Big(\hat{f}_n(X_{1:n}, f^\star(X_{1:n}), X_t), f^\star(X_t)\Big) \to 0 \text{ (a.s.)}. \tag{72}$$

This further implies that, for any measurable $f^\star : \mathcal{X} \to \mathcal{Y}$,

$$\begin{aligned}
\lim_{n \to \infty} \hat{\mathcal{L}}_{\mathbb{X}}\Big(\hat{f}_n, f^\star; n\Big) &= \lim_{n \to \infty} \limsup_{m \to \infty} \frac{1}{m} \sum_{t=n+1}^{n+m} \ell\Big(\hat{f}_n(X_{1:n}, f^\star(X_{1:n}), X_t), f^\star(X_t)\Big) \\
&\leq \lim_{n \to \infty} \sup_{t \in \mathbb{N}} \ell\Big(\hat{f}_n(X_{1:n}, f^\star(X_{1:n}), X_t), f^\star(X_t)\Big) = 0 \text{ (a.s.)}.
\end{aligned}$$

Thus, the inductive learning rule $\hat{f}_n$ is strongly universally consistent under $\mathbb{X}$. In particular, this implies that $\mathbb{X}$ *admits* strong universal inductive learning: that is, $\mathbb{X} \in$ SUIL.

The above argument can also be used to show that $\mathbb{X} \in$ SUOL. Specifically, consider this same $\hat{f}_n$ function defined above, but now interpreted as an *online* learning rule. We then have, for any measurable $f^\star : \mathcal{X} \to \mathcal{Y}$,

$$\begin{aligned}
\lim_{n \to \infty} \hat{\mathcal{L}}_{\mathbb{X}}\Big(\hat{f}_\cdot, f^\star; n\Big) &= \lim_{n \to \infty} \frac{1}{n} \sum_{t=0}^{n-1} \ell\Big(\hat{f}_t(X_{1:t}, f^\star(X_{1:t}), X_{t+1}), f^\star(X_{t+1})\Big) \\
&\leq \lim_{n \to \infty} \frac{1}{n} \sum_{t=0}^{n-1} \sup_{m \in \mathbb{N}} \ell\Big(\hat{f}_t(X_{1:t}, f^\star(X_{1:t}), X_m), f^\star(X_m)\Big). \qquad (73)
\end{aligned}$$

The convergence in (72) implies $\sup_{m \in \mathbb{N}} \ell\Big(\hat{f}_t(X_{1:t}, f^\star(X_{1:t}), X_m), f^\star(X_m)\Big) \to 0$ (a.s.) as $t \to \infty$. Thus, since the arithmetic mean of the first n elements in any convergent sequence in $\mathbb{R}$ is also convergent (as $n \to \infty$) with the same limit value, this immediately implies that the last expression in (73) equals 0 almost surely. Since this holds for any measurable $f^\star : \mathcal{X} \to \mathcal{Y}$, we have that $\hat{f}_n$ is also a strongly universally consistent *online* learning rule under $\mathbb{X}$. In particular, this implies that $\mathbb{X}$ admits strong universal online learning: that is, $\mathbb{X} \in$ SUOL.

Finally, since the above arguments hold for *any* choice of $\mathbb{X} \in \mathcal{C}_3$, we may conclude that $\mathcal{C}_3 \subseteq$ SUIL $\cap$ SUOL, which completes the proof. ∎

Combining the above lemmas immediately provides the following proof of Theorem 45.

Proof of Theorem 45 Taking Lemmas 48, 49, and 52 together, we have that SUIL $\cup$ SUOL $\subseteq$ SUAL $\cup$ SUOL $\subseteq \mathcal{C}_3 \subseteq$ SUIL $\cap$ SUOL $\subseteq$ SUAL $\cap$ SUOL. This further implies that

SUAL $\triangle$ SUOL = (SUAL $\cup$ SUOL) $\setminus$ (SUAL $\cap$ SUOL) = $\emptyset$, and similarly SUIL $\triangle$ SUOL = (SUIL $\cup$ SUOL) $\setminus$ (SUIL $\cap$ SUOL) = $\emptyset$, so that SUIL = SUOL = SUAL. Combining this with Lemmas 49 and 52, we obtain SUOL = SUAL$\cup$SUOL $\subseteq \mathcal{C}_3 \subseteq$ SUIL$\cap$SUOL = SUOL, so that SUOL = $\mathcal{C}_3$. Hence SUIL = SUAL = SUOL = $\mathcal{C}_3$, which completes the proof. ■

Remark: Interestingly, the proof of Lemma 52 in fact establishes a much stronger kind of convergence for $\hat{f}_n$ under any $\mathbb{X} \in \mathcal{C}_3$: for any measurable $f^\star : \mathcal{X} \to \mathcal{Y}$,

$$\sup_{t \in \mathbb{N}} \ell\Big(\hat{f}_n(X_{1:n}, f^\star(X_{1:n}), X_t), f^\star(X_t)\Big) \to 0 \text{ (a.s.)}. \tag{74}$$

Denoting by SUIL$^{\text{sup}}$ the set of processes $\mathbb{X}$ that admit the existence of an inductive learning rule $\hat{f}_n$ satisfying (74) for every measurable $f^\star : \mathcal{X} \to \mathcal{Y}$, we have thus established that $\mathcal{C}_3 \subseteq$ SUIL$^{\text{sup}}$ when $\bar{\ell} = \infty$. Furthermore, as shown in the proof of Lemma 52, this type of convergence itself implies strong universal consistency of $\hat{f}_n$ in the original sense of Definition 1, so that SUIL$^{\text{sup}} \subseteq$ SUIL. Thus, since SUIL = $\mathcal{C}_3$ when $\bar{\ell} = \infty$ (from Theorem 45, just established), we have established that, when $\bar{\ell} = \infty$, SUIL$^{\text{sup}}$ = SUIL: that is, the set of processes $\mathbb{X}$ admitting this stronger type of universal consistency is in fact the *same* as those admitting strong universal inductive learning in the usual sense of Definition 1. It is clear that this is *not* the case when $\bar{\ell} < \infty$ if $\mathcal{X}$ is infinite. Indeed, combining the proof of Lemma 52 with a straightforward variation on the proof of Lemma 49, one can show that *even when* $\bar{\ell} < \infty$, Condition 3 remains a necessary and sufficient condition for a process $\mathbb{X}$ to admit the existence of an inductive learning rule satisfying (74) for all measurable functions $f^\star : \mathcal{X} \to \mathcal{Y}$: that is, SUIL$^{\text{sup}} = \mathcal{C}_3$. For these same reasons, the same is true of the analogous guarantee for self-adaptive or online learning: that is, regardless of whether $\bar{\ell} = \infty$ or $\bar{\ell} < \infty$, Condition 3 is necessary and sufficient for there to exist a self-adaptive learning rule $\hat{g}_{n,m}$ such that, for all measurable $f^\star : \mathcal{X} \to \mathcal{Y}$,

$$\sup_{t \in \mathbb{N}: t \geq n} \ell(\hat{g}_{n,t}(X_{1:t}, f^\star(X_{1:n}), X_{t+1}), f^\star(X_{t+1})) \to 0 \text{ (a.s.)},$$

and Condition 3 is also necessary and sufficient for there to exist an online learning rule $\hat{h}_n$ such that, for all measurable $f^\star : \mathcal{X} \to \mathcal{Y}$,

$$\ell\Big(\hat{h}_n(X_{1:n}, f^\star(X_{1:n}), X_{n+1}), f^\star(X_{n+1})\Big) \to 0 \text{ (a.s.)}.$$

We may also note that the proof of Lemma 52 specifically establishes that the inductive learning rule $\hat{f}_n$ specified in (66) (with $\{\tilde{f}_i\}_{i=1}^\infty$ from Lemma 51) is strongly universally consistent for every $\mathbb{X} \in \mathcal{C}_3$, and therefore by Theorem 45 (just established), for every $\mathbb{X} \in$ SUIL when $\bar{\ell} = \infty$. Since the definition of $\hat{f}_n$ has no direct dependence on the distribution of $\mathbb{X}$, this implies $\hat{f}_n$ is an optimistically universal inductive learning rule when $\bar{\ell} = \infty$. This is particularly interesting, as it contrasts with the fact, established in Theorem 6 above, that for *bounded* losses, no optimistically universal inductive learning rule exists (if $\mathcal{X}$ is an uncountable Polish space). Furthermore, this also means we can easily define an optimistically universal *self-adaptive* learning rule when $\bar{\ell} = \infty$, simply defining

$$\hat{g}_{n,m}(x_{1:m}, y_{1:n}, x) = \hat{f}_n(x_{1:n}, y_{1:n}, x) \tag{75}$$

for every $n, m \in \mathbb{N} \cup \{0\}$ with $m \geq n$, and every $x_{1:m} \in \mathcal{X}^n$, $y_{1:n} \in \mathcal{Y}^n$, and $x \in \mathcal{X}$. In particular, it is clear that $\hat{\mathcal{L}}_{\mathbb{X}}(\hat{g}_{n,\cdot}, f^{\star}; n) = \hat{\mathcal{L}}_{\mathbb{X}}(\hat{f}_n, f^{\star}; n)$ for this definition of $\hat{g}_{n,m}$. Thus, since $\hat{f}_n$ is strongly universally consistent under every $\mathbb{X} \in \mathcal{C}_3$ by Lemma 52, it immediately follows that $\hat{g}_{n,m}$ also has this property, and the fact that it is an optimistically universal self-adaptive learning rule (when $\bar{\ell} = \infty$) then follows from SUAL $= \mathcal{C}_3$ (from Theorem 45, just established). The proof of Lemma 52 also establishes strong universal consistency of $\hat{f}_n$ under any $\mathbb{X} \in \mathcal{C}_3$ when $\hat{f}_n$ is interpreted as an *online* learning rule, so that (since $\mathcal{C}_3 =$ SUOL when $\bar{\ell} = \infty$, again by Theorem 45) $\hat{f}_n$ is also an optimistically universal *online* learning rule when $\bar{\ell} = \infty$. We summarize these findings in the following theorem.

Theorem 53 *When $\bar{\ell} = \infty$, with $\{\tilde{f}_i\}_{i=1}^{\infty}$ as in Lemma 51, the learning rule $\hat{f}_n$ from* (66) *is an optimistically universal inductive learning rule, and an optimistically universal online learning rule. Moreover, defining $\hat{g}_{n,m}$ as in* (75)*, when $\bar{\ell} = \infty$, $\hat{g}_{n,m}$ is an optimistically universal self-adaptive learning rule.*

In particular, this implies that for unbounded losses, there *exist* optimistically universal (inductive/self-adaptive/online) learning rules, so that Theorem 46 immediately follows.

8.4 No Consistent Test for Existence of a Universally Consistent Learner

As we did in Section 7 in the case of bounded losses, it is also natural to ask whether there exist consistent hypothesis tests for whether or not a given data process $\mathbb{X}$ admits strong universal learning, in this case when $\bar{\ell} = \infty$. As was true for bounded losses, we again find that the answer is generally *no*. Formally, we have the following theorem.

Theorem 54 *When $\bar{\ell} = \infty$ and $\mathcal{X}$ is infinite, there is no consistent hypothesis test for* SUIL*,* SUAL*, or* SUOL*.*

Proof Suppose $\mathcal{X}$ is infinite. Since Theorem 45 implies SUIL $=$ SUAL $=$ SUOL $= \mathcal{C}_3$ when $\bar{\ell} = \infty$, it suffices to prove that there is no consistent hypothesis test for $\mathcal{C}_3$. Fix any hypothesis test $\hat{t}_n$. Fix $\mathbb{X}$ to be that specific process constructed in the proof of Theorem 43, relative to this hypothesis test $\hat{t}_n$. The proof of Theorem 43 (combined with Theorem 7) establishes that, for this specific process $\mathbb{X}$, if $\mathbb{X} \in \mathcal{C}_1$, then $\hat{t}_n(X_{1:n})$ fails to converge in probability to 1, and if $\mathbb{X} \notin \mathcal{C}_1$, then $\hat{t}_n(X_{1:n})$ fails to converge in probability to 0.

Recall that $\mathcal{C}_3 \subseteq \mathcal{C}_1$, so that if $\mathbb{X} \notin \mathcal{C}_1$, then $\mathbb{X} \notin \mathcal{C}_3$ as well. But, as mentioned above, $\hat{t}_n(X_{1:n})$ fails to converge in probability to 0 in this case. Thus, in the case that this process $\mathbb{X} \notin \mathcal{C}_1$, we have established that $\hat{t}_n$ is not a consistent test for $\mathcal{C}_3$.

On the other hand, in the case that the constructed process $\mathbb{X}$ *is* in $\mathcal{C}_1$, there are two subcases to consider. First, recalling the construction of $\mathbb{X}$, if there exists a largest $k \in \mathbb{N}$ for which n_{k-1} is defined, then for $\mathbb{X}$ to be in $\mathcal{C}_1$ we necessarily have $(k+1)/2 \in \mathbb{N}$ (i.e., k is odd). In this case, every $t > n_{k-1}$ has $X_t = w_0 = X_{n_{k-1}+1}$, so that for any sequence $\{A_i\}_{i=1}^{\infty}$ in $\mathcal{B}$ with $A_i \downarrow \emptyset$,

$$|\{i \in \mathbb{N} : \mathbb{X} \cap A_i \neq \emptyset\}| = \left|\left\{i \in \mathbb{N} : X_{1:(n_{k-1}+1)} \cap A_i \neq \emptyset\right\}\right|.$$

Since $A_i \downarrow \emptyset$, $\exists i_0 \in \mathbb{N}$ such that $\forall i > i_0$, $X_{1:(n_{k-1}+1)} \cap A_i = \emptyset$. Therefore,

$$|\{i \in \mathbb{N} : \mathbb{X} \cap A_i \neq \emptyset\}| \leq i_0 < \infty,$$

so that $\mathbb{X} \in \mathcal{C}_3$ as well. But, as mentioned above, in the case that this constructed process $\mathbb{X} \in \mathcal{C}_1$, $\hat{t}_n(X_{1:n})$ fails to converge in probability to 1, so that if $\mathbb{X} \in \mathcal{C}_1$ and there is a largest $k \in \mathbb{N}$ with n_{k-1} defined, this establishes that $\hat{t}_n$ is not a consistent test for $\mathcal{C}_3$. Finally, the only remaining case is where $\mathbb{X} \in \mathcal{C}_1$ and n_{k-1} is defined for every $k \in \mathbb{N}$. In this case, as established in the proof of Theorem 43, $\hat{t}_n(X_{1:n})$ fails to converge in probability *at all* (i.e., *neither* converges in probability to 0 *nor* converges in probability to 1), which trivially establishes that $\hat{t}_n$ is not a consistent test for $\mathcal{C}_3$ in this case. ■

Since it is trivially true that *every* $\mathbb{X}$ is in $\mathcal{C}_3$ when $\mathcal{X}$ is *finite* (and hence also in SUIL, SUAL, and SUOL when $\bar{\ell} = \infty$, by Theorem 45), we have the following immediate corollary.

Corollary 55 *When $\bar{\ell} = \infty$, there exist consistent hypothesis tests for each of* SUIL, SUAL, *and* SUOL *if and only if $\mathcal{X}$ is finite.*

9. Extensions

Here we briefly mention two simple extensions of the above theory: namely, extension beyond metric losses ℓ, and extension of the results to also hold for *weak* universal consistency.

9.1 More-General Loss Functions

For simplicity, we have chosen to restrict the loss function ℓ to be a *metric* in the above results. However, as mentioned in Section 1.1, most of the theory developed above extends to a much broader family of loss functions, including all functions $\ell : \mathcal{Y}^2 \to [0, \infty)$ that are merely *dominated* by a separable metric ℓ_o, in the sense that $\forall y, y' \in \mathcal{Y}, \ell(y, y') \leq \phi(\ell_o(y, y'))$ for some continuous nondecreasing function $\phi : [0, \infty) \to [0, \infty)$ with $\phi(0) = 0$ and $\lim\limits_{x \to \infty} \phi(x) = \infty$, and that also satisfy a non-triviality condition: $\sup\limits_{y_0, y_1 \in \mathcal{Y}} \inf\limits_{y \in \mathcal{Y}} \max\{\ell(y, y_0), \ell(y, y_1)\} > 0$. The measurable sets $\mathcal{B}_y$ are then defined as the Borel σ-algebra generated by the topology induced by ℓ_o, and we also require that ℓ be a measurable function with respect to this. For instance, this extension admits such common settings as ℓ the squared loss for $\mathcal{Y}$ as $\mathbb{R}$ or $[-1, 1]$, taking $\ell_o(y, y') = |y - y'|$ and $\phi(x) = x^2$.

Here we briefly elaborate on the (minor) changes to the above theory yielding this generalization. For any $z \in [0, \infty)$, denote $\phi^{-1}(z) = \inf\{x \in [0, \infty) : \phi(x) \geq z\}$; this always exists since the conditions on ϕ guarantee that its range is $[0, \infty)$, and moreover by continuity of ϕ we have $\phi(\phi^{-1}(z)) = z$. Still defining $\bar{\ell} = \sup\limits_{y, y' \in \mathcal{Y}} \ell(y, y')$, in the case of bounded losses ($\bar{\ell} < \infty$), note that we can suppose ℓ_o is also bounded without loss of generality, and in fact that it is bounded by $\phi^{-1}(\bar{\ell})$, since the metric $(y, y') \mapsto \ell_o(y, y') \wedge \phi^{-1}(\bar{\ell})$ still satisfies the requirement $\ell(y, y') \leq \phi(\ell_o(y, y') \wedge \phi^{-1}(\bar{\ell}))$. Then we can simply replace ℓ with ℓ_o in the learning rules proposed in (10) and (29), and the resulting performance guarantees in terms of the loss ℓ_o then imply universal consistency under ℓ under the same conditions. To see this, note that for any $\hat{y}, y^\star \in \mathcal{Y}$, for any $\varepsilon > 0$, we have

$$\ell(\hat{y}, y^\star) \leq \phi(\ell_o(\hat{y}, y^\star)) \leq \varepsilon + \bar{\ell} \mathbb{1}\left[\ell_o(\hat{y}, y^\star) > \phi^{-1}(\varepsilon)\right] \leq \varepsilon + \frac{\bar{\ell}}{\phi^{-1}(\varepsilon)} \ell_o(\hat{y}, y^\star),$$

noting that $\phi^{-1}(\varepsilon) > 0$. Plugging this inequality into the three $\hat{\mathcal{L}}_{\mathbb{X}}$ definitions, and noting that it holds for all $\varepsilon > 0$, it easily follows that, in any of the three learning settings discussed above, strong universal consistency under the loss ℓ_o implies strong universal consistency under the loss ℓ.

Furthermore, in the results where it is needed to argue inconsistency of a given learning rule (Lemma 19, Theorems 6 and 35), the only property of ℓ used in those arguments is the stated non-triviality condition; more specifically, this condition is represented there by the fact that, for ℓ a metric, any distinct $y_0, y_1 \in \mathcal{Y}$ have $\inf_{y \in \mathcal{Y}} \frac{1}{2}(\ell(y, y_0) + \ell(y, y_1)) \geq \ell(y_0, y_1)/2 > 0$, but the arguments would hold just as well for these more-general losses ℓ by replacing $\ell(y_0, y_1)/2$ with $\inf_{y \in \mathcal{Y}} \max\{\ell(y, y_0), \ell(y, y_1)\}/2$ and choosing $y_0, y_1 \in \mathcal{Y}$ specifically to make this latter quantity nonzero.

These generalizations can be applied to all of the results involving a loss function in Sections 1 through 6.3. Section 6.4 is the only place (involving bounded losses) where somewhat-nontrivial modifications are necessary to extend the results to these more-general losses, simply due to needing an appropriate generalization of the notion of "total boundedness" for the arguments to remain valid.

The results on unbounded losses in Section 8 can also be generalized. In this case, the same trick of using ℓ_o in place of ℓ in the definition of the learning rule (66) again works for establishing universal consistency with ℓ under $\mathbb{X} \in \mathcal{C}_3$ in Lemma 52, but in this case it follows from the stronger guarantee (74) for ℓ_o (together with continuity and monotonicity of ϕ, and $\phi(0) = 0$) rather than from directly relating $\hat{\mathcal{L}}_{\mathbb{X}}$ for the losses ℓ_o and ℓ: that is, the learning rule defined in terms of ℓ_o satisfies the convergence in (74) for the loss ℓ_o under $\mathbb{X} \in \mathcal{C}_3$, and the properties of ϕ imply that it remains true for $\phi(\ell_o(\cdot, \cdot))$, and hence also for the loss ℓ. However, the complementary result in Lemma 49 requires an additional restriction to ℓ for the argument there to generalize: namely, that $\sup_{y_0, y_1 \in \mathcal{Y}} \inf_{y \in \mathcal{Y}} \max\{\ell(y, y_0), \ell(y, y_1)\} = \infty$, a property satisfied by most unbounded losses studied in the literature anyway. Using this to replace the values $\ell(y_{i,0}, y_{i,1})$ appearing in the proof of Lemma 49 with values $\inf_{y \in \mathcal{Y}} \max\{\ell(y, y_{i,0}), \ell(y, y_{i,1})\}$ (both in the definition of $y_{i,0}, y_{i,1}$, and in (64)), the result is then extended to these more-general loss functions. Together, these modifications allow us to extend all of the results in Section 8 to these more-general loss functions ℓ.

9.2 Weak Universal Consistency

It is straightforward to extend the above results on inductive and self-adaptive learning (Sections 4 and 5) to *weak* universal consistency as well, where the definition of weakly universally consistent learning is as above except replacing the *almost sure* convergence of $\hat{\mathcal{L}}_{\mathbb{X}}$ to 0 with convergence *in probability*. The proof of *necessity* of Condition 1 for inductive learning and self-adaptive learning (from Lemmas 18 and 19) can easily be modified to show necessity of Condition 1 for *weak* universal consistency by inductive or self-adaptive learning rules as well. Specifically, the proof of Lemma 19 in this case would follow the same argument, but starting from $\sup_{\kappa \in [0,1)} \limsup_{n \to \infty} \mathbb{E}\left[\hat{\mathcal{L}}_{\mathbb{X}}(g_{n,\cdot}, f^{\star}_{\kappa}; n)\right]$ instead of $\sup_{\kappa \in [0,1)} \mathbb{E}\left[\limsup_{n \to \infty} \hat{\mathcal{L}}_{\mathbb{X}}(g_{n,\cdot}, f^{\star}_{\kappa}; n)\right]$. After relaxing $\sup_{\kappa \in [0,1)}$ to an integral over $\kappa \in$

$[0,1)$ (as in the present proof) and applying Fatou's lemma to exchange the integral operator with the $\limsup_{n\to\infty}$, the proof proceeds identically as before, and the final conclusion follows by noting that if $\lim_{n\to\infty} \hat{\mu}_{\mathbb{X}}(\bigcup\{A_i : X_{1:n} \cap A_i = \emptyset\}) > 0$ with nonzero probability, then (by the monotone convergence theorem) $\lim_{n\to\infty} \mathbb{E}[\hat{\mu}_{\mathbb{X}}(\bigcup\{A_i : X_{1:n} \cap A_i = \emptyset\})] = \mathbb{E}\Big[\lim_{n\to\infty} \hat{\mu}_{\mathbb{X}}(\bigcup\{A_i : X_{1:n} \cap A_i = \emptyset\})\Big] > 0$. For brevity, we leave the details of the proof as an exercise for the interested reader. Since strong universal consistency implies weak universal consistency, the sufficiency of Condition 1 for universal consistency of inductive or self-adaptive learning (from Lemmas 25 and 18), as well as the result on optimistically universal self-adaptive learning (Lemma 27), continue to hold for the *weak* universal consistency criterion in place of *strong* universal consistency. In particular, this means that the set of processes admitting weak universal inductive or self-adaptive learning is equal to SUIL or SUAL, both of which are equal $\mathcal{C}_1$ by Theorem 7. Additionally, it follows from statements made in the proof of Theorem 6 that Theorem 6 remains valid for weak universal consistency as well. Again, the details are left as an exercise for the interested reader.

Interestingly, the extension to weak consistency in the *online* learning setting (with $\bar{\ell} < \infty$) is substantially more involved, and indeed the set of processes that admit *weak* universal online learning is in fact a *strict* superset of SUOL (if $\mathcal{X}$ is infinite). That it is a superset easily follows from the fact that almost sure convergence implies convergence in probability, so the interesting detail here is that there exist processes $\mathbb{X}$ that admit weak universal online learning but *not* strong universal online learning. To see this, consider the following construction of a process $\mathbb{X}$. Let $\{z_i\}_{i=0}^{\infty}$ be distinct elements of $\mathcal{X}$ (supposing $\mathcal{X}$ is infinite), and let $\{B_k\}_{k=1}^{\infty}$ be independent random variables with $B_k \sim \text{Bernoulli}(1/k)$. Then for each $k \in \mathbb{N}$ and each $t \in \{2^{k-1}, \ldots, 2^k - 1\}$, if $B_k = 1$, then set $X_t = z_t$, and if $B_k = 0$, then set $X_t = z_0$. Since $\sum_{k=1}^{\infty} 1/k = \infty$, the second Borel-Cantelli lemma implies that, with probability one, there exists an infinite strictly-increasing sequence $\{k_i\}_{i=1}^{\infty}$ in $\mathbb{N}$ with $B_{k_i} = 1$ for every $i \in \mathbb{N}$. On this event, every $k \in \{k_i : i \in \mathbb{N}\}$ has $|\{j \in \mathbb{N} : X_{1:(2^k-1)} \cap \{z_j\} \neq \emptyset\}| \geq 2^{k-1}$, so that $|\{j \in \mathbb{N} : X_{1:T} \cap \{z_j\} \neq \emptyset\}| \neq o(T)$ (a.s.). Thus, $\mathbb{X} \notin \mathcal{C}_2$, and hence by Theorem 35, $\mathbb{X} \notin$ SUOL. However, if we take f_n as the simple memorization-based online learning rule (from the proof of Theorem 36), then for any $n \in \mathbb{N}$ and measurable $f^\star : \mathcal{X} \to \mathcal{Y}$, we have $\mathbb{E}\Big[\hat{\mathcal{L}}_{\mathbb{X}}(f_\cdot, f^\star; n)\Big] \leq \frac{\bar{\ell}}{n}\mathbb{E}\left[|\{j \in \mathbb{N} \cup \{0\} : X_{1:n} \cap \{z_j\} \neq \emptyset\}|\right] \leq \frac{\bar{\ell}}{n}\left(1 + \sum_{k=1}^{\lfloor \log_2(2n) \rfloor} 2^{k-1}(1/k)\right) \leq \frac{\bar{\ell}}{n}\left(1 + \int_1^{\lfloor \log_2(4n) \rfloor} 2^{x-1}(1/x)\mathrm{d}x\right)$. Since $\int_1^t 2^{x-1}(1/x)\mathrm{d}x = o\left(\int_1^t 2^x \mathrm{d}x\right)$ as $t \to \infty$ (by L'Hôpital's rule and the fundamental theorem of calculus), and $\int_1^t 2^x \mathrm{d}x = \frac{1}{\ln(2)}2^t$, we conclude that $\mathbb{E}\Big[\hat{\mathcal{L}}_{\mathbb{X}}(f_\cdot, f^\star; n)\Big] = o(1)$, which implies $\hat{\mathcal{L}}_{\mathbb{X}}(f_\cdot, f^\star; n) \xrightarrow{P} 0$ by Markov's inequality. Thus, $\mathbb{X}$ admits weak universal online learning.

Following arguments analogous to the proof of Theorem 35, one can show that a *necessary* condition for a process $\mathbb{X}$ to admit weak universal online learning is that every disjoint sequence $\{A_i\}_{i=1}^{\infty}$ in $\mathcal{B}$ satisfies $\mathbb{E}[|\{i \in \mathbb{N} : X_{1:T} \cap A_i \neq \emptyset\}|] = o(T)$. This represents a sort of *weak* form of Condition 2. Furthermore, following similar arguments to the proof of

Theorem 36, one can show that in the special case of *countable* $\mathcal{X}$, this condition is both necessary *and sufficient* for $\mathbb{X}$ to admit weak universal online learning. However, as was the case for Condition 2 and strong universal consistency (Open Problem 2), in the general case (allowing uncountable $\mathcal{X}$) it remains an open problem to determine whether this weaker form of Condition 2 is equivalent to the condition that $\mathbb{X}$ admits weak universal online learning. Likewise, it also remains an open problem to determine whether there generally exist optimistically universal online learning rules under this weak consistency criterion instead of the strong consistency criterion.

In the case of unbounded losses, one can show that Theorems 45 and 46 extend to weak universal consistency without modification. Since almost sure convergence implies convergence in probability, Theorem 45 immediately implies sufficiency of Condition 3 for a process to admit weak universal learning (in all three settings). Furthermore, the same construction used in the proof of Lemma 49 can be used to show that Condition 3 is also necessary for weak universal learning (again in all three settings) when $\bar{\ell} = \infty$. Specifically, for any $\mathbb{X} \notin \mathcal{C}_3$, in the notation defined in the proof of Lemma 49, we would have that for any online learning rule h_n, every $j \in \mathbb{N}$ has $\mathbb{P}\Big(\hat{\mathcal{L}}_{\mathbb{X}}(h_{\cdot}, f^{\star}_{K}; T_j) > \frac{1}{2}\Big) \geq \frac{1}{2}\mathbb{P}(0 < \tau_j \leq T_j) \geq \frac{1}{2}(\mathbb{P}(E) - 2^{-j})$, which is bounded away from 0 for all sufficiently large j. Since one can also show that $T_j \to \infty$, it follows that $\exists \kappa \in [0,1)$ such that $\limsup_{n\to\infty} \mathbb{P}\Big(\hat{\mathcal{L}}_{\mathbb{X}}(h_{\cdot}, f^{\star}_{\kappa}; n) > \frac{1}{2}\Big) > 0$, so that h_n is not weakly universally consistent under $\mathbb{X}$. Similarly, for any self-adaptive learning rule $g_{n,m}$, we would have that for any $n \in \mathbb{N}$, $\mathbb{P}\Big(\hat{\mathcal{L}}_{\mathbb{X}}(g_{n,\cdot}, f^{\star}_{K}; n) \geq \frac{1}{2}\Big) \geq \mathbb{P}(E \cap E') > 0$, which implies $\exists \kappa \in [0,1)$ such that $\limsup_{n\to\infty} \mathbb{P}\Big(\hat{\mathcal{L}}_{\mathbb{X}}(g_{n,\cdot}, f^{\star}_{\kappa}; n) \geq \frac{1}{2}\Big) > 0$, so that $g_{n,m}$ is not weakly universally consistent under $\mathbb{X}$. The same argument holds for any inductive learning rule f_n as well. The details of these arguments are left as an exercise for the interested reader. Together with Theorems 45 and 46 and the fact that almost sure convergence implies convergence in probability, this also implies that there exists an optimistically universal learning rule (in all three settings) under this weak consistency criterion as well.

10. Open Problems

For convenience, we conclude the paper by briefly gathering in summary form the main open problems posed in the sections above, along with additional general directions for future study. The statements involving ℓ regard the case $\bar{\ell} < \infty$.

- Open Problem 1: Does there exist an optimistically universal online learning rule?
- Open Problem 2: Is SUOL $= \mathcal{C}_2$?
- Open Problem 3: Is the set SUOL invariant to the specification of $(\mathcal{Y}, \ell)$, subject to being separable with $0 < \bar{\ell} < \infty$?
- Open Problem 4: For some uncountable $\mathcal{X}$, do there exist processes $\mathbb{X} \in \mathcal{C}_3$ such that, with nonzero probability, the number of distinct $x \in \mathcal{X}$ appearing in $\mathbb{X}$ is infinite?

One additional general direction for future study is to introduce the possibility of stochastic Y_t values given X_t, rather than simply supposing $Y_t = f^{\star}(X_t)$ as above. Such an extension would necessarily re-introduce some degree of arbitrariness in the assumptions. One simple starting place would be to suppose Y_t is conditionally independent of $\{Y_{t'}\}_{t' \neq t}$ given X_t, and that $\mathbb{E}[\ell(f^\star(X_t), Y_t)|X_t] = \min_{y \in \mathcal{Y}} \mathbb{E}[\ell(y, Y_t)|X_t]$, and then we would be interested in the average loss achieved by a learning rule f_n relative to $f^\star$: that is (in the case of inductive learning, for instance), $\limsup_{m \to \infty} \frac{1}{m} \sum_{t=n+1}^{n+m} (\ell(f_n(X_{1:n}, f^\star(X_{1:n}), X_t), Y_t) - \ell(f^\star(X_t), Y_t))$. The question is then whether the conditions for achieving universal consistency in this excess loss remain the same as for the theory developed above, or generally what the conditions would be, and whether optimistically universal learning rules exist (under this conditional independence assumption). However, it is conceivable that other assumptions may be important for such a theory, such as (as an extreme example) supposing that the conditional distribution of Y_t given X_t is invariant to t; this would still be a strict generalization of the theory developed in the present work.

References

T. M. Adams and A. B. Nobel. On density estimation from ergodic processes. *The Annals of Probability*, 26(2):794–804, 1998. 5.2

R. B. Ash and C. A. Doléans-Dade. *Probability & Measure Theory*. Academic Press, second edition, 2000. 1.1, 2.3, 4, 4, 5.2, 6.2, 8.3, 8.3

S. Ben-David, D. Pál, and S. Shalev-Shwartz. Agnostic online learning. In *Proceedings of the* 22^{nd} *Conference on Learning Theory*, 2009. 1

S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1):151–175, 2010. 1, 1.1

V. I. Bogachev. *Measure Theory*, volume 1. Springer-Verlag, 2007. 2.2, 8.3

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006. 1, 1.1

N. Cesa-Bianchi, Y. Freund, D. Haussler, D. P. Helmbold, R. E. Schapire, and M. K. Warmuth. How to use expert advice. *Journal of the Association for Computing Machinery*, 44(3):427–485, 1997. 6.1

O. Chapelle, B. Schölkopf, and A. Zien. *Semi-supervised Learning*. MIT Press, 2010. 1, 1.1

G. Choquet. Theory of capacities. *Annales de l'institut Fourier*, 5:131–295, 1954. 2.2

D. L. Cohn. *Measure Theory*. Birkhäuser, 1980. 5.2

C. Cortes, M. Mohri, M. Riley, and A. Rostamizadeh. Sample selection bias correction theory. In *Proceedings of the* 19^{th} *International Conference on Algorithmic Learning Theory*, 2008. 1, 1.1

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag New York, 1996. 1, 3.2

I. Dobrakov. *On Submeasures I*, volume 112 of *Dissertationes Mathematicae*. Państwowe Wydawnictwo Naukowe, 1974. 2.2

I. Dobrakov. On extension of submeasures. *Mathematica Slovaca*, 34(3):265–271, 1984. 2.2

R. M. Dudley. *Real Analysis and Probability*. Cambridge University Press, second edition, 2003. 8.3

R. M. Gray. *Probability, Random Processes, and Ergodic Properties*. Springer, second edition, 2009. 3, 3.1

L. Györfi and G. Lugosi. Strategies for sequential prediction of stationary time series. In M. Dror, P. L'Ecuyer, and F. Szidarovszky, editors, *Modeling Uncertainty: An Examination of Stochastic Theory, Methods, and Applications*, pages 225–248. Kluwer Academic Publishers, 2002. 6.1

L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer-Verlag New York, 2002. 1.1, 1.1

D. Haussler, N. Littlestone, and M. Warmuth. Predicting $\{0,1\}$-functions on randomly drawn points. *Information and Computation*, 115(2):248–292, 1994. 1, 1.1

J. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, and B. Schölkopf. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems 19*, 2007. 1, 1.1

A. S. Kechris. *Classical Descriptive Set Theory*. Springer-Verlag New York, 1995. 4, 8.3

J. Kivinen and M. K. Warmuth. Averaging expert predictions. In *Proceedings of the* 4^{th} *European Conference on Computational Learning Theory*, 1999. 6.1

A. N. Kolmogorov and S. V. Fomin. *Introductory Real Analysis*. Dover, 1975. 5.2

N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2(4):285–318, 1988. 1, 1.1

N. Littlestone and M. K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994. 6.1, 6.1

D. Maharam. An algebraic characterization of measure algebras. *Annals of Mathematics*, 48(1):154–167, 1947. 2.2

A. B. Nobel. Limits to classification and regression estimation from ergodic processes. *The Annals of Statistics*, 27(1):262–273, 1999. 5.2

G. L. O'Brien and W. Vervaat. How subadditive are subadditive capacities? *Commentationes Mathematicae Universitatis Carolinae*, 35(2):311–324, 1994. 2.2

K. R. Parthasarathy. *Probability Measures on Metric Spaces*. Academic Press, 1967. 5.2

A. Rakhlin, K. Sridharan, and A. Tewari. Online learning via sequential complexities. *Journal of Machine Learning Research*, 16(2):155–186, 2015. 1, 1.1

M. J. Schervish. *Theory of Statistics*. Springer-Verlag New York, 1995. 3.1, 3.2, 4, 5.2, 8.3

A. Singer and M. Feder. Universal linear prediction by model order weighting. *IEEE Transactions on Signal Processing*, 47(10):2685–2699, 1999. 6.1

S. M. Srivastava. *A Course on Borel Sets*. Springer-Verlag New York, 1998. 4, 5.2

I. Steinwart, D. Hush, and C. Scovel. Learning from dependent observations. *Journal of Multivariate Analysis*, 100(1):175–194, 2009. 1.1

M. Talagrand. Maharam's problem. *Annals of Mathematics*, 168(3):981–1009, 2008. 2.2

V. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag New York, 1982. 1

V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998. 1

V. Vovk. Aggregating strategies. In *Proceedings of the* 3^{rd} *Annual Workshop on Computational Learning Theory*, 1990. 6.1

V. Vovk. Universal forecasting algorithms. *Information and Computation*, 96(2):245–277, 1992. 6.1