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

Steve Hanneke STEVE.HANNEKE@GMAIL.COM
Toyota Technological Institute at Chicago

Abstract

This work initiates a general study of learning and generalization without the i.i.d. assumption, starting from first principles. While the traditional approach to statistical learning theory typically relies on standard assumptions from probability theory (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 necessary 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, the assumptions we have focused on in the literature tend to be imported from the probability theory literature, rather than being rooted in a principled approach to the learning problem itself. 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). While such assumptions are known to be sufficient for learning due to their convenient convergence properties (i.e., laws of large numbers), it is clear that they are not *necessary* for learning. 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 traditional probabilistic assumptions about the data, such as independence and stationarity (which will be recovered as special cases).

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 *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 under 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 $Y_t = f^\star(X_t)$,[1] 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., $t > n$).[2] 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 $f^\star$.[3]

1. Later we also discuss extensions to allow noisy responses Y_t.
2. 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.
3. 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. We omit this qualification and simply write "universally consistent" for brevity. The more-general case of random variable pairs (X_t, Y_t), where Y_t may be noisy, will be discussed in Section 9.

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 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) *test* 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 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. In principle, self-adaptive learning should be possible in many common learning scenarios where the test data are observed sequentially, such as in pattern recognition based on a data stream from a camera or other sensors. 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-closely related) 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; Hanneke and Kpotufe, 2019). 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 most-complete 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 *as-*

sumption, 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.

In addition to these general results for function learning, we also discuss extensions of the theory to allow *noisy responses* Y_t, in Section 9. Specifically, we consider the case of responses Y_t that are *conditionally independent* given X_t, with the further requirement that there is a time-invariant optimal function $f^\star$. We find that the results for inductive and self-adaptive learning indeed extend to these noisy scenarios, for certain families of losses: for instance, regression with the squared loss.

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 define $\bar{\ell} = \sup_{y,y' \in \mathcal{Y}} \ell(y, y')$. 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 *near-metric*: that is, $\forall y_1, y_2 \in \mathcal{Y}$, ℓ satisfies $\ell(y_1, y_2) = \ell(y_2, y_1)$, and $\ell(y_1, y_2) = 0$ if and only if $y_1 = y_2$, and also satisfies a relaxed triangle inequality, namely, there is a finite constant $c_\ell \geq 1$ such that $\forall y_1, y_2, y_3 \in \mathcal{Y}$, $\ell(y_1, y_2) \leq c_\ell(\ell(y_1, y_3) + \ell(y_3, y_2))$. We further suppose that $(\mathcal{Y}, \ell)$ is *separable*, in the usual sense that there exists a countable $\tilde{\mathcal{Y}} \subseteq \mathcal{Y}$ with $\sup_{y \in \mathcal{Y}} \inf_{\tilde{y} \in \tilde{\mathcal{Y}}} \ell(\tilde{y}, y) = 0$. For instance, these conditions are satisfied for discrete classification under the 0-1 loss ($\mathcal{Y}$ countable, $\ell(a, b) = \mathbb{1}[a \neq b]$), or bounded real-valued regression under the squared loss ($\mathcal{Y} = [-B, B]$, $\ell(a, b) = (a-b)^2$) or indeed any L_p loss ($\ell(a, b) = |a-b|^p$, $p > 0$), as well as many other losses. Most of the theory developed here also easily extends to any ℓ that is merely *dominated* by a separable near-metric ℓ_o, in the sense that $\forall y, y' \in \mathcal{Y}, \ell(y, y') \leq \chi(\ell_o(y, y'))$ for a continuous nondecreasing function $\chi : [0, \infty) \to [0, \infty)$ with $\chi(0) = 0$ and satisfying a non-triviality condition $\sup_{y_0, y_1} \inf_y \max\{\ell(y, y_0), \ell(y, y_1)\} > 0$. This then admits discrete classification with asymmetric misclassification costs, and many other interesting cases. We include a brief discussion of this generalization in Section 10.1.

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 $\mathcal{Y}$ generated by the open balls of ℓ, $\{\{y \in \mathcal{Y} : \ell(y, y_0) < r\} : y_0 \in \mathcal{Y}, r > 0\}$, 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 define $f(\mathbf{x}) = \{f(x_i)\}_{i=1}^{\infty}$ 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

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

In particular, we will often use this notation with $\mathbf{x} = \mathbb{X}$, for a process $\mathbb{X} = \{X_t\}_{t=1}^{\infty}$, in which case $\hat{\mu}_{\mathbb{X}}(g)$ is a random variable. 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): $\limsup_{i\to\infty} A_i = \bigcap_{k=1}^{\infty} \bigcup_{i=k}^{\infty} A_i$, $\liminf_{i\to\infty} A_i = \bigcup_{k=1}^{\infty} \bigcap_{i=k}^{\infty} A_i$, and $\lim_{i\to\infty} A_i$ exists and equals $\limsup_{i\to\infty} A_i$ if and only if $\limsup_{i\to\infty} A_i = \liminf_{i\to\infty} A_i$. As one final remark on notation, we note that we will generally interpret claims regarding conditional expectations to mean that there exist *versions* of the corresponding conditional expectations for which the claims hold, such as in $\mathbb{E}[Z|Y] \leq \mathbb{E}[Z|X] = \mathbb{E}[W|X]$.

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*. Note that, in the case of general processes $\mathbb{X}$, 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 n 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 implies the following simple equality.

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

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). For more-general 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; Hanneke and Kpotufe, 2019), as the additional samples $X_{(n+1):m}$ provide important information about how representative the training samples $X_{1:n}$ are for the purpose of estimating certain relevant long-run averages $\hat{\mu}_{\mathbb{X}}(g)$ (see Section 5.1 for details). In particular, for this purpose, it is important that these additional unlabeled samples are actual *test* samples, rather than (for instance) taken from an independent copy of the process, since general (non-ergodic) processes may have very different long-run behaviors in different sample paths.

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 near-metric, $0 < \bar{\ell} < \infty$). 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). Though the analysis of learning techniques becomes significantly more challenging under these relaxations, in many cases the essential features of the i.i.d. setting useful for proving consistency are preserved (particularly, laws of large numbers). 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*. By definition, this is a *necessary* assumption for universal consistency. Thus, the important question is whether there is a learning rule for which it is also a *sufficient* assumption for establishing universal consistency. 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**. Hence 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. As mentioned, self-adaptive learning should be possible in many common learning scenarios where the unlabeled test data are observed sequentially, such as in pattern recognition based on a data stream from a camera or other sensors. In light of these results, it seems worthwhile to revisit the definitions of commonly-used learning methods with a view toward designing self-adaptive variants. In the self-adaptive method we propose in Section 5.1 below, the main utility of being self-adaptive is in a model selection component: for each hypothesis class $\mathcal{F}_i$ in a hierarchy of classes, we use $X_{1:n}$ and $X_{1:m}$ to produce two estimates of $\hat{\mu}_{\mathbb{X}}(\ell(f(\cdot), f'(\cdot)))$ for all $f, f' \in \mathcal{F}_i$, and select the largest class $\mathcal{F}_i$ in the hierarchy for which these two estimates are uniformly close. If $\mathcal{F}_i$ is sufficiently rich to approximate $f^\star$, this technique functions as an approximate test for whether a particular estimate of $\hat{\mu}_{\mathbb{X}}(\ell(f(\cdot), f^\star(\cdot)))$ based on $(X_{1:n}, f^\star(X_{1:n}))$ is close to the true value $\hat{\mu}_{\mathbb{X}}(\ell(f(\cdot), f^\star(\cdot)))$, for all $f \in \mathcal{F}_i$, so that minimizing the estimate over $f \in \mathcal{F}_i$ produces a function f with relatively small $\hat{\mu}_{\mathbb{X}}(\ell(f(\cdot), f^\star(\cdot)))$; see Section 5.1 for the details.

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 47 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.

In Section 9, we extend the theory to allow the Y_t response values to be *noisy*, subject to being conditionally independent. We discuss other extensions of the theory in Section 10, admitting 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 11 by summarizing several interesting open questions that arise from the theory developed below.

1.3 Related Work

There are several important works in the literature related to universal consistency under non-i.i.d. processes. Questions about consistency under general stationary ergodic processes were posed by Cover (1975) for the forecasting problem (i.e., predicting Y_{t+1} based on $Y_{1:t}$) and related settings. In particular, Cover's question of whether there is an estimator $\hat{m}(Y_{1:t})$ with $|\hat{m}(Y_{1:t}) - \mathbb{E}[Y_{t+1}|Y_{1:t}]| \to 0$ (a.s.) for all stationary ergodic $\mathbb{Y} = \{Y_t\}_{t=1}^{\infty}$ on $\{0,1\}$ was answered negatively by Bailey (1976) and Ryabko (1988). A related negative result was also established by Nobel (1999) for regression, showing there is no estimator $\hat{f}_t(X_{1:t}, Y_{1:t}, \cdot)$ with $\mathbb{E}|\hat{f}(X_{1:t}, Y_{1:t}, X) - \mathbb{E}[Y|X]| \to 0$ for all stationary ergodic processes $(\mathbb{X}, \mathbb{Y}) = \{(X_t, Y_t)\}_{t=1}^{\infty}$ on $[0,1]^2$, and where (X, Y) is an independent random variable of the same marginal distribution. In contrast, there is a substantial literature on estimators that are consistent (in various senses) under *mixing* conditions, which are stronger than ergodicity (e.g., Steinwart, Hush, and Scovel, 2009; Lozano, Kulkarni, and Schapire, 2006; Roussas, 1988; Collomb, 1984; Irle, 1997).

On the other hand, a series of works by Ornstein (1978), Algoet (1992, 1994, 1999), Morvai, Yakowitz, and Györfi (1996), Györfi, Lugosi, and Morvai (1999), Györfi and Lugosi (2002), and Nobel (2003) showed (among other results) that there *do* exist universally consistent forecasting rules under general (with bounded moment) stationary er-

godic processes $\mathbb{Y} = \{Y_t\}_{t\in\mathbb{Z}}$ on $\mathbb{R}$, if we are merely interested in the long-run *average* loss: that is, $\frac{1}{n}\sum_{t=0}^{n-1} |\hat{m}(Y_{1:t}) - \mathbb{E}[Y_{t+1}|Y_{-\infty:t}]| \to 0$ (a.s.). This is analogous to the *online* setting studied in the present work. This result was extended to classification and bounded regression settings by Morvai, Yakowitz, and Györfi (1996), Györfi, Lugosi, and Morvai (1999), and Györfi and Lugosi (2002), yielding an online learning rule $\hat{f}_t$ for which $\frac{1}{n}\sum_{t=0}^{n-1}(\hat{f}_t(X_{1:t}, Y_{1:t}, X_{t+1}) - \mathbb{E}[Y_{t+1}|X_{-\infty:(t+1)}, Y_{-\infty:t}])^2 \to 0$ (a.s.) for all stationary ergodic processes $(\mathbb{X}, \mathbb{Y}) = \{(X_t, Y_t)\}_{t\in\mathbb{Z}}$ on $\mathbb{R}^d \times \mathbb{R}$ with $|Y_t|$ bounded.

In contrast, as we discuss below (Section 3), an immediate implication of Theorems 5, 7, and 41 is that the ergodicity assumption is superfluous for the existence of such estimators (i.e., stationarity alone suffices), if we restrict to cases where $Y_t = f^\star(X_t)$ for arbitrary unknown functions $f^\star$, or more generally, cases where Y_t is conditionally independent of $\{(X_s, Y_s)\}_{s\neq t}$ given X_t. Indeed, universal consistency is even possible in the much weaker self-adaptive setting for such stationary processes. One interpretation of this is that, while the stationary ergodic assumption enables a learner to estimate and optimize its *expected* risk, stationarity alone already suffices if we are only interested in estimating and optimizing its average loss on the actual future samples in the individual sample path, so that information about the expected risk is unnecessary. We also remark that the results established here also hold for many non-stationary processes as well.

Other works have considered learning under various *non-stationary* processes. A mild form of non-stationarity was discussed by Irle (1997), who constructs consistent regression estimators under mixing processes that have vanishing average total variation distance of the marginals to a fixed distribution the risk is defined under. Steinwart, Hush, and Scovel (2009) generalize this to the family of all processes for which a *law of large numbers* holds, which includes all *asymptotically mean stationary* ergodic processes (Gray and Kieffer, 1980; Gray, 2009). However, the learning rule of Steinwart, Hush, and Scovel (2009) has a dependence on the distribution of $(\mathbb{X}, \mathbb{Y})$, which is in fact necessary (due to the negative result of Nobel, 1999; see also the proof of Theorem 6 below). However, Steinwart, Hush, and Scovel (2009) show that this dependence can be removed if we further restrict to processes $(\mathbb{X}, \mathbb{Y})$ satisfying a mixing condition (constrained weak α-bi-mixing rate), in which case an $(\mathbb{X}, \mathbb{Y})$-independent choice of the parameter sequence yields weak consistency. This relaxes the requirements of an earlier result of Lozano, Kulkarni, and Schapire (2006) establishing strong consistency (for a different learning rule) for stationary processes satisfying a stronger type of mixing condition (constrained β-mixing rate). Morvai, Kulkarni, and Nobel (1999) also studied consistency under general processes satisfying a law of large numbers, in a bounded regression setting on $\mathbb{R}$. The results there even hold for deterministic processes, as long as the frequencies converge to a probability measure in the limit. They specifically show that a particular learning rule is consistent as long as the regression function of the limit distribution satisfies a known constraint on its total variation in each bounded interval.

Other works have considered learning with families of non-stationary processes not even satisfying a law of large numbers. Kulkarni, Posner, and Sandilya (2002) established a very general result, showing that for the regression setting (generalized to Hilbert spaces $\mathcal{Y}$), if we only require the learner to be consistent for *continuous* target functions $f^\star$, then there is an online learning rule that is strongly consistent under every process $\mathbb{X} = \{X_t\}_{t=1}^\infty$ such that the set $\{X_t : t \in \mathbb{N}\}$ is almost surely totally bounded. For instance, if $\mathcal{X}$ is totally bounded, then this is the case for all processes $\mathbb{X}$ on $\mathcal{X}$. They in fact establish a more general

result that also allows Y_t to be corrupted by conditionally independent noise (subject to $f^\star(X_t) = \mathbb{E}[Y_t|X_t]$), a topic we discuss in Section 9 below. The results in the present work reveal that, if we seek truly *universal* learners, consistent for *all* possible target functions $f^\star$, including discontinuous functions, then even in totally bounded spaces $\mathcal{X}$, there exist processes $\mathbb{X}$ where no such universal learners exist. Thus, the best we can aim for is a learning rule that is universally consistent under every process $\mathbb{X}$ that *admits* universal learning: i.e., an *optimistically* universal learner.

Ryabko (2006) introduced another type of non-stationary process for the classification setting with finite $\mathcal{Y}$: namely, processes $(\mathbb{X}, \mathbb{Y})$ where the process $\mathbb{Y} = \{Y_t\}_{t=1}^\infty = \{f^\star(X_t)\}_{t=1}^\infty$ is arbitrary (subject to each $y \in \mathcal{Y}$ occurring with non-vanishing liminf frequency), and the X_t sequence is "conditionally i.i.d.", meaning that the X_t variables are conditionally independent given their respective $Y_t = f^\star(X_t)$ values, with time-invariant conditional distribution. This family of processes captures many interesting scenarios beyond the i.i.d. assumption, including many non-stationary processes. Under this condition, Ryabko (2006) shows that certain learning rules known to be universally consistent under the i.i.d. assumption remain (weakly) consistent under this more-general family of processes (in the online setting). In particular, this is true of the nearest neighbor rule. He also shows strong universal consistency for learning rules based on empirical risk minimization for a sequence of hypothesis classes becoming rich as $n \to \infty$. The consistency result of Ryabko (2006) is stated in a stronger form: $\mathbb{P}(h_n(X_{1:n}, Y_{1:n}, X_{n+1}) \neq Y_{n+1}|X_{1:n}, Y_{1:n}) \to 0$ (a.s.). However, under the conditions considered by Ryabko (2006), this would actually be satisfied by any inductive learning rule h_n satisfying $\hat{\mathcal{L}}_{\mathbb{X}}(h_n, f^\star; n) \to 0$ (a.s.), and indeed the learning rules considered by Ryabko (2006) are of this type.

The nature of the conditional i.i.d. assumption of Ryabko (2006) is somewhat different from the conditions studied in the present work, in that it is a condition on the joint process $(\mathbb{X}, \mathbb{Y})$ rather than $\mathbb{X}$ alone. Nevertheless, it is straightforward to verify that for finite $\mathcal{Y}$, for any $(\mathbb{X}, \mathbb{Y})$ satisfying the conditional i.i.d. condition, $\mathbb{X}$ satisfies Condition 1, and hence by Theorems 7 and 41 it admits strong universal learning (in any of the three settings studied here); this is true even without restrictions on the frequencies of each $y \in \mathcal{Y}$. Note that this also implies consistency even for target functions $f^\star$ for which $(\mathbb{X}, f^\star(\mathbb{X}))$ is not conditionally i.i.d.

Ryabko (2006) also asks a question of how to extend beyond the setting of deterministic responses $Y_t = f^\star(X_t)$ to allow more-general distributions of Y_t given X_t. The results in Section 9.3 on the topic of handling noisy labels are relevant to this question, in particular studying the case where the noise is conditionally independent: that is, the optimal prediction $f^\star(X_t)$ remains a t-invariant function of X_t, but the observed response Y_t may be stochastic. This condition can be combined with Ryabko's conditional i.i.d. assumption by taking the X_t sequence to be conditionally i.i.d. given an (arbitrary) $f^\star(X_t)$ sequence, and then taking the Y_t responses to be conditionally independent given the respective X_t values (subject to the requirement that the conditional distribution of Y_t given X_t is a t-invariant function of X_t, and $y = f^\star(X_t)$ minimizes $\mathbb{E}[\ell(y, Y_t)|X_t]$). The results in Section 9.3 then imply that there is a learning rule that is strongly consistent for every process $(\mathbb{X}, \mathbb{Y})$ of this

type (in the self-adaptive or online setting).[4] Moreover, the results in Section 9 also imply consistency under a much broader family of processes.

In the broader subject of learning theories beyond i.i.d. processes, the topic of finding a function with near-minimal risk within a fixed restricted hypothesis class $\mathcal{F}$ has received considerably more attention. There, the goal is consistency relative to $\mathcal{F}$: that is, finding a function $\hat{f} \in \mathcal{F}$ with risk converging to at most the best risk achievable by functions in $\mathcal{F}$. Most of this work has studied learning under stationary processes satisfying various *mixing* conditions. Of particular relevance in this context is the set $\mathcal{F}_\ell = \{(x, y) \mapsto \ell(f(x), y) : f \in \mathcal{F}\}$. When $\mathcal{F}_\ell$ is a VC class, or generally has bounded covering numbers, methods based on variants of empirical risk minimization have been shown to be consistent relative to $\mathcal{F}$ (see e.g., Yu, 1994; Karandikar and Vidyasagar, 2002, 2004; Vidyasagar, 2005; Zou, Li, and Xu, 2009). Moving beyond mixing assumptions, when $\mathcal{F}_\ell$ is a VC class, Adams and Nobel (2010a,b, 2012) showed that empirical risk minimization is consistent relative to $\mathcal{F}$ under every stationary ergodic process. Later, van Handel (2013) extended this to all classes such that $\mathcal{F}_\ell$ is a universal Glivenko-Cantelli class. Kuznetsov and Mohri (2014) and Hanneke and Yang (2019) have also considered extensions of some of these results to some restricted families of *non-stationary* mixing processes, constrained by the rate of change of the single-index marginal distribution.

A significant change in the present work compared to the above is that much of the prior work on statistical learning without the i.i.d. assumption essentially studies the same learning methods developed for i.i.d. processes, such as local averaging estimators or empirical risk minimization. In contrast, we argue below that such methods will fail in certain non-stationary scenarios, in which other methods would be consistent. As such, the techniques we develop in this work necessarily differ significantly from those designed with i.i.d. processes in mind.

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 proving some basic properties of the $\hat{\mu}_{\mathbf{x}}$ functional that will be indispensable in the main proofs below.

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

4. One can show this result also holds in the inductive setting for this special case, though we will not discuss this extension, for the sake of brevity.

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

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

■

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* $A_k \downarrow \emptyset$, $\lim_{k\to\infty}\nu(A_k) = 0$.

Note that we have defined "submeasure" to only require *finite* subadditivity. However, it immediately follows that any *continuous* submeasure would also be *countably* subadditive (see Fremlin, 2002, Chapter 39, Lemma 392H). The relevance of this definition to our present discussion is via the set function $\mathbb{E}[\hat{\mu}_{\mathbb{X}}(\cdot)]$, which we can easily show is always a submeasure, as follows.

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 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), $\lim_{k\to\infty} \hat{\mu}_{\mathbb{X}}(A_k)$ always exists and is nonnegative. Therefore, since the set function $\hat{\mu}_{\mathbb{X}}$ is bounded in $[0, 1]$, the dominated convergence theorem implies $\mathbb{E}\left[\lim_{k\to\infty} \hat{\mu}_{\mathbb{X}}(A_k)\right] = \lim_{k\to\infty} \mathbb{E}\left[\hat{\mu}_{\mathbb{X}}(A_k)\right] = 0$, where the last equality is due to Condition 1. Combined with the fact that $\lim_{k\to\infty} \hat{\mu}_{\mathbb{X}}(A_k) \geq 0$, it follows that $\lim_{k\to\infty} \hat{\mu}_{\mathbb{X}}(A_k) = 0$ (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 $B_k = A_k \setminus \bigcup_{j>k} A_j$. Note that $\{B_k\}_{k=1}^{\infty}$ is a sequence of disjoint measurable sets. In particular, this implies $\bigcup_{k\geq i} B_k \downarrow \emptyset$, so that (since $\mathbb{X}$ satisfies the first condition) $\lim_{i\to\infty} \hat{\mu}_{\mathbb{X}}\left(\bigcup_{k\geq i} B_k\right) = 0$ (a.s.). Furthermore, for any $i \in \mathbb{N}$, we have $\bigcup_{k\geq i} A_k = \left(\limsup_{j\to\infty} A_j\right) \cup \bigcup_{k\geq i} B_k$. 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 $\limsup_{j\to\infty} A_j \subseteq \bigcup_{k\geq i} A_k$ for every $i \in \mathbb{N}$, monotonicity of $\hat{\mu}_{\mathbb{X}}$ (Lemma 8) implies $\hat{\mu}_{\mathbb{X}}\left(\bigcup_{k\geq i} A_k\right) \geq \hat{\mu}_{\mathbb{X}}\left(\limsup_{j\to\infty} A_j\right)$, which implies $\lim_{i\to\infty} \hat{\mu}_{\mathbb{X}}\left(\bigcup_{k\geq i} A_k\right) \geq \hat{\mu}_{\mathbb{X}}\left(\limsup_{j\to\infty} A_j\right)$. 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 $\limsup_{k\to\infty} A_k = \emptyset$, 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 $B_k = A_k \setminus \bigcup_{j>k} A_j$. 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 $\forall k \in \mathbb{N}$, $A_k = \left(\limsup_{j\to\infty} A_j\right) \cup \bigcup_{i\geq k} B_i$; furthermore, $A_k \downarrow \emptyset$ implies $\limsup_{j\to\infty} A_j = \emptyset$, so that $A_k = \bigcup_{i\geq k} B_i$. 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 (3) will be a key component (via Corollary 15) in the proof below (in Lemma 20) 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.*

$$\lim_{k\to\infty} \hat{\mu}_{\mathbf{x}}\left(\bigcup_{i\geq k} A_i\right) = 0. \tag{2}$$

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

$$\lim_{m\to\infty} \lim_{n\to\infty} \hat{\mu}_{\mathbf{x}}\Big(\bigcup\{A_i : |x_{1:n} \cap A_i| < m\}\Big) = 0. \tag{4}$$

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 (3) 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

$$\lim_{k\to\infty} \hat{\mu}_{\mathbf{x}}\left(\bigcup_{i\geq k} A_i\right) \leq \lim_{k\to\infty} \hat{\mu}_{\mathbf{x}}\Big(\bigcup\{A_i : x_{1:n_k} \cap A_i = \emptyset\}\Big). \tag{5}$$

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 (2) when combined with (5). Otherwise, suppose $n_k < \infty$ for all $k \in \mathbb{N}$. In this case, we will argue that $n_k \to \infty$. 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 that for $k' = i(x_{n_k+1}) + 1$ we have $x_{1:(n_k+1)} \cap \bigcup_{i\geq k'} A_i = \emptyset$, and therefore $n_{k'} \geq n_k + 1$. Together with monotonicity of n_k in k, this implies $n_k \to \infty$. Combined with (3) and monotonicity of $\hat{\mu}_{\mathbf{x}}(\bigcup\{A_i : x_{1:n} \cap A_i = \emptyset\})$ in n, 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 (2) when combined with (5) and nonnegativity of $\hat{\mu}_{\mathbf{x}}$ (Lemma 9).

Next, suppose (2) 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 $n, k \in \mathbb{N}$,

$$\hat{\mu}_{\mathbf{x}}\Big(\bigcup\{A_i : |x_{1:n} \cap A_i| < m\}\Big) \leq \hat{\mu}_{\mathbf{x}}\Big(\bigcup\{A_i : |x_{1:n} \cap A_i| < m, i \geq k\}\Big) + \sum_{\substack{i\in\{1,\ldots,k-1\}:\\ |x_{1:n}\cap A_i|<m}} \hat{\mu}_{\mathbf{x}}(A_i). \tag{6}$$

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, for every $n \in \mathbb{N}$, defining

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

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

$$\sum_{\substack{i \in \{1,\ldots,k_n-1\}: \\ |x_{1:n} \cap A_i| < m}} \hat{\mu}_{\mathbf{x}}(A_i) = 0. \tag{7}$$

Next we argue that $k_n \to \infty$. To see this, note that 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 (6) and (7), 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 (2) 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 (4).

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 (4) is satisfied, then (3) must also hold. ■

This further implies the following corollary relating Condition 1 to a requirement of having vanishing *missing mass* in any countable discretization of $\mathcal{X}$.

Corollary 15 *A process $\mathbb{X}$ satisfies Condition 1 if and only if every disjoint sequence $\{A_i\}_{i=1}^{\infty}$ in $\mathcal{B}$ with $\bigcup_{i=1}^{\infty} A_i = \mathcal{X}$ (i.e., every countable measurable partition) satisfies*

$$\lim_{n\to\infty} \hat{\mu}_{\mathbb{X}}\Big(\bigcup\{A_i : X_{1:n} \cap A_i = \emptyset\}\Big) = 0 \text{ (a.s.).} \tag{8}$$

Proof If $\mathbb{X}$ satisfies Condition 1, then for any disjoint sequence $\{A_i\}_{i=1}^{\infty}$ in $\mathcal{B}$, Lemma 13 implies that, on an event of probability one, $\lim_{k\to\infty} \hat{\mu}_{\mathbb{X}}\left(\bigcup_{i\geq k} A_i\right) = 0$. The equivalence of (2) and (3) in Lemma 14 then implies that, on this same event, $\lim_{n\to\infty} \hat{\mu}_{\mathbb{X}}(\bigcup\{A_i : X_{1:n} \cap A_i = \emptyset\}) = 0$, so that (8) holds.

On the other hand, if $\mathbb{X}$ does not satisfy Condition 1, then Lemma 13 implies that there exists a disjoint sequence $\{A'_i\}_{i=1}^{\infty}$ in $\mathcal{B}$ such that, on an event of nonzero probability, $\lim_{k\to\infty} \hat{\mu}_{\mathbb{X}}\left(\bigcup_{i\geq k} A'_i\right) > 0$. Since this claim only involves the tail of the A'_i sequence, if we define $A_1 = \mathcal{X} \setminus \bigcup_{i=1}^{\infty} A'_i$ and $A_{i+1} = A'_i$ for $i \in \mathbb{N}$ (so that $\{A_i\}_{i=1}^{\infty}$ is a countable measurable partition of $\mathcal{X}$), then on this same event we have $\lim_{k\to\infty} \hat{\mu}_{\mathbb{X}}\left(\bigcup_{i\geq k} A_i\right) > 0$. The equivalence of (2) and (3) in Lemma 14 then implies that, on this same event, $\lim_{n\to\infty} \hat{\mu}_{\mathbb{X}}(\bigcup\{A_i : X_{1:n} \cap A_i = \emptyset\}) > 0$, so that (8) does not 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 deterministic process $\mathbb{X}$ on $\mathbb{N}$ with $\forall i \in \mathbb{N}, \hat{\mu}_{\mathbb{X}}(\{i\}) = 1$ has countably subadditive $\hat{\mu}_{\mathbb{X}}$ and yet $\mathbb{X} \notin \mathcal{C}_1$).

Lemma 16 *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 (2) 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 $k \in \mathbb{N}$,

$$\hat{\mu}_{\mathbf{x}}\left(\bigcup_{i=1}^{\infty} A_i\right) \leq \hat{\mu}_{\mathbf{x}}\left(\bigcup_{i\geq k} A_i\right) + \sum_{i=1}^{k-1} \hat{\mu}_{\mathbf{x}}(A_i). \tag{9}$$

If (2) 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 (9) yields the claimed inequality, completing the proof. ■

Lemma 17 *If $\mathbb{X}$ satisfies Condition 1, then for any sequence $\{A_i\}_{i=1}^{\infty}$ in $\mathcal{B}$,*

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

Proof Let $B_1 = A_1$, and for each $i \in \mathbb{N}\backslash\{1\}$, let $B_i = A_i \backslash \bigcup_{j=1}^{i-1} A_j$. 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 16, 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. In particular, in Section 3.2 we provide an example of a process satisfying Condition 1 under which the nearest neighbor predictor fails to be universally consistent.

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

$$\lim_{m\to\infty} \frac{1}{m} \sum_{t=1}^{m} \mathbb{1}_A(X_t) \text{ exists (a.s.).} \tag{10}$$

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* (Gray, 2009, Theorem 8.1): that is, $\forall A \in \mathcal{B}^\infty$, $\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathbb{P}(X_{t:\infty} \in A)$ exists.

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 18 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 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 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 $\mathbb{X} \notin$ CRF. ■

3.2 Inconsistency of the Nearest Neighbor Rule

The separation $\mathcal{C}_1 \setminus$ 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 $\mathcal{X} \subseteq \mathbb{R}^d$ (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 $i_n = \operatorname*{argmin}_{i \in \{1,\ldots,n\}} \|x - x_i\|$ (with an appropriate policy for breaking ties). For classification and regression in $\mathbb{R}^d$ 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 $\mathbb{X} \in \mathcal{C}_1$ for $\mathcal{X} = [0,1]$, under which the nearest neighbor inductive learning rule is *not* universally consistent for binary classification.[5] This also provides a second proof that $\mathcal{C}_1 \backslash \text{CRF} \neq \emptyset$ 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 the fraction of the data in $[0, 1/2)$ is relatively small compared to $[1/2, 1]$, and other segments of the sequence in which the fraction of the data in $[1/2, 1]$ is relatively small 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 in 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 $I = \{1\} \cup \{n_{k-1} + (i-1)k + 1 : k \in \mathbb{N} \backslash \{1\}, i \in \{1, \ldots, n_{k-1}^2\}\}$. Note that, for any $k \in \mathbb{N} \backslash \{1, 2\}$ 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

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

5. 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.

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, noting that every $t \in \mathbb{N}\setminus I$ has $X_t \in \left\{W_t, \frac{1}{2}+W_t\right\}$, we have

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

and the strong law of large numbers implies 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 $A_k \downarrow \emptyset$, $\lim_{k\to\infty} \mathbb{E}[\hat{\mu}_{\mathbb{X}}(A_k)] \leq \lim_{k\to\infty} 2\lambda(A_k) = 0$ 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

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

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 countable. Note that we have defined $f^\star$ so that every $t \in I$ has $f^\star(X_t) = y_0$, and with probability one 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 for the nearest neighbor inductive learning rule f_n, letting $N_k = \{x \in [0,1] : f_{n_k}(X_{1:n_k}, f^\star(X_{1:n_k}), x) = y_0\}$, it holds that $\lambda(N_k \cap (0,1/2)) \geq \frac{1}{2} - \frac{n_{k-1}}{2n_{k-1}^2} = \frac{1}{2}\left(1-\frac{1}{n_{k-1}}\right)$. In particular, this implies that a Uniform$(0,1/2)$ random variable (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

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

with probability at least $1 - \exp\left\{-\frac{1}{2(2k'-1)^2}\left(1-\frac{1}{n_{k-1}}\right)(2k'-1)n_{2k'-1}^2\right\} > 1 - e^{-(2k'-1)/4}$ (since $n_{2k'-1} \geq 2k'-1$ and $n_{k-1} > 2$). Noting that $\sum_{k'=2}^{\infty} e^{-(2k'-1)/4} < \infty$, the Borel-Cantelli

lemma implies that with probability one this event occurs for all sufficiently large k'. Thus, by the union bound, we have that with probability one,

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

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,

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

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) $\mathrm{SUIL} = \mathrm{SUAL} = \mathcal{C}_1$.

The proof is by way of three lemmas: Lemma 20, representing necessity of Condition 1 for strong universal self-adaptive learning, Lemma 27, representing sufficiency of Condition 1 for strong universal inductive learning, and Lemma 19, 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 19 $\mathrm{SUIL} \subseteq \mathrm{SUAL}$.

Proof Let $\mathbb{X} \in \text{SUIL}$, and let f_n be an inductive learning rule such that, 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.). 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 20 $\text{SUAL} \subseteq \mathcal{C}_1$.

Proof We prove this result in the contrapositive. Suppose $\mathbb{X} \notin \mathcal{C}_1$. By Corollary 15, there exists a disjoint sequence $\{A_i\}_{i=1}^\infty$ in $\mathcal{B}$ with $\bigcup_{i=1}^\infty A_i = \mathcal{X}$ such that, 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 $\{f^\star_\kappa : \kappa \in [0,1)\}$ 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 $f^\star_\kappa(x) = y_{\kappa_i}$. Note that $(x, \kappa) \mapsto f^\star_\kappa(x)$ is measurable in the product σ-algebra (under $\mathcal{B}$ for the x argument, and the usual Borel σ-algebra on $[0,1)$ for the κ argument), since the inverse image of $\{y_1\}$ is $\bigcup_{i=1}^\infty (A_i \times \{\kappa : \kappa_i = 1\})$ (a countable union of measurable rectangle sets) and the inverse image of $\{y_0\}$ is the complement of this set.

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 $f^\kappa_{n,m} : \mathcal{X} \to \mathcal{Y}$ as $f^\kappa_{n,m}(\cdot) = g_{n,m}(X_{1:m}, f^\star_\kappa(X_{1:n}), \cdot)$ (a composition of measurable functions, and therefore measurable). Then we have

$$\sup_{\kappa\in[0,1)} \mathbb{E}\left[\limsup_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(g_{n,\cdot}, f^\star_\kappa; n)\right] \geq \int_0^1 \mathbb{E}\left[\limsup_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(g_{n,\cdot}, f^\star_\kappa; n)\right] \mathrm{d}\kappa$$

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

By Fubini's theorem, this last expression is equal

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

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

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

and linearity of integration implies this equals

$$\mathbb{E}\left[\limsup_{n\to\infty}\limsup_{t\to\infty}\frac{1}{t+1}\sum_{m=n}^{n+t}\mathbb{1}_{\bar{\mathcal{A}}(X_{1:n})}(X_{m+1})\int_0^1 \ell\big(f^\kappa_{n,m}(X_{m+1}), f^\star_\kappa(X_{m+1})\big)\,\mathrm{d}\kappa\right]. \tag{11}$$

Note that, for any m, the value of $f^\kappa_{n,m}(X_{m+1})$ 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 $f^\kappa_{n,m}(X_{m+1})$ 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

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

By the relaxed triangle inequality, this last line is no smaller than $\mathbb{E}\Big[\frac{1}{2c_\ell}\ell(y_0,y_1)\Big|\mathbb{X}, g_{n,m}\Big] = \frac{1}{2c_\ell}\ell(y_0,y_1)$, so that (11) is at least as large as

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

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 $\lim_{n\to\infty}\hat{\mu}_{\mathbb{X}}(\bigcup\{A_i : X_{1:n}\cap A_i=\emptyset\}) > 0$ with probability strictly greater than 0, and the left hand side of this inequality is nonnegative, we have that $\mathbb{E}\Big[\lim_{n\to\infty}\hat{\mu}_{\mathbb{X}}(\bigcup\{A_i : X_{1:n}\cap A_i=\emptyset\})\Big] > 0$. Furthermore, since ℓ is a near-metric, we also have $\ell(y_0,y_1)>0$. Altogether we have that

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

In particular, this implies $\exists\kappa\in[0,1)$ such that $\mathbb{E}\left[\limsup_{n\to\infty}\hat{\mathcal{L}}_{\mathbb{X}}(g_{n,\cdot}, f^\star_\kappa; n)\right] > 0$. Since any random variable equal 0 (a.s.) necessarily has expected value 0, and since we have $\limsup_{n\to\infty}\hat{\mathcal{L}}_{\mathbb{X}}(g_{n,\cdot}, f^\star_\kappa; n)\geq 0$, it must be that $\limsup_{n\to\infty}\hat{\mathcal{L}}_{\mathbb{X}}(g_{n,\cdot}, f^\star_\kappa; n) > 0$ with probability strictly greater than 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, $\mathbb{X} \notin \text{SUAL}$. 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

$$\operatorname*{argmin}_{f \in \mathcal{F}_n} \max_{\hat{m}_n \le m \le n} \frac{1}{m} \sum_{t=1}^{m} \ell(f(x_t), y_t), \tag{12}$$

where $\mathcal{F}_n$ is a well-chosen finite class of functions $\mathcal{X} \to \mathcal{Y}$, and $\hat{m}_n$ is a well-chosen integer. Suppose ties in the argmin are broken based on a fixed preference ordering of $\mathcal{F}_n$. In particular, this makes $\hat{f}_n$ a measurable function, and hence (12) defines a valid inductive learning rule. For our purposes, $\hat{f}_0(\{\}, \{\}, \cdot)$ can be defined as an arbitrary measurable function $\mathcal{X} \to \mathcal{Y}$.

The sequence of classes $\mathcal{F}_n$ and values $\hat{m}_n$, and the guarantees they provide, originate in the following several lemmas. The general strategy is to define $\mathcal{F}_n$ so that, for large n, $\mathcal{F}_n$ is rich enough to contain a function f with small $\hat{\mu}_{\mathbb{X}}(\ell(f(\cdot), f^\star(\cdot)))$, while at the same time not too rich, so that (for an appropriate choice of $\hat{m}_n$) $\max_{\hat{m}_n \le m \le n} \frac{1}{m} \sum_{t=1}^{m} \ell(f(X_t), f^\star(X_t))$ is a reasonable estimate of $\hat{\mu}_{\mathbb{X}}(\ell(f(\cdot), f^\star(\cdot)))$ for all $f \in \mathcal{F}_n$. The fact that these two properties can exist simultaneously, and for all possible $f^\star$, is enabled by $\mathbb{X}$ satisfying Condition 1. We now proceed with the details.

Lemma 21 *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 $\{s_n\}_{n=1}^{\infty}$ in $\mathbb{N}$ with $s_n \le n$ and $s_n \to \infty$ such that*

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

Proof The proof of this lemma proceeds in three steps. First, we note that the result would follow almost immediately from the definition of limsup, if the sequence $\mathbb{X}$ were *deterministic* and we were merely interested in the case of a *single* function g. Second, we extend this observation to any finite *set* $\mathcal{G}$ of functions by taking the s_n sequence as the minimum of the corresponding s_n values for each individual g. The final component is to extend the result to hold for non-deterministic processes $\mathbb{X}$, by replacing the s_n sequence corresponding to each sample path of $\mathbb{X}$ with an appropriate confidence bound on its value. This last step requires us to introduce some notation in the first two steps to explicitly define these s_n sequences for the sample paths, so that together they are a measurable function of $\mathbb{X}$, and hence confidence bounds are well-define. We now turn to the details of the 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,

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

In particular, for each $s \in \mathbb{N}$, since $\max_{s\le m\le n} \frac{1}{m}\sum_{t=1}^{m} g(x_t)$ is nondecreasing in $n \ge s$, and g is bounded, $\lim_{n\to\infty} \max_{s\le m\le n} \frac{1}{m}\sum_{t=1}^{m} g(x_t)$ exists and is finite. This implies that, for each $s \in \mathbb{N}$, $\exists n_s^g(\mathbf{x}) \in \mathbb{N}$ s.t. $n_s^g(\mathbf{x}) \ge s$ and every $n' \ge n_s^g(\mathbf{x})$ has

$$\max_{s\le m\le n'} \frac{1}{m}\sum_{t=1}^{m} g(x_t) \le \sup_{s\le m<\infty} \frac{1}{m}\sum_{t=1}^{m} g(x_t) \le 2^{-s} + \max_{s\le m\le n'} \frac{1}{m}\sum_{t=1}^{m} g(x_t). \tag{13}$$

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 (13) is trivially satisfied for every $s, n' \in \mathbb{N}$ with $n' \ge s$. Moreover, for any $n', s \in \mathbb{N}$ with $s \ge 2$ and $n' \ge n_s^g(\mathbf{x})$, either $\sup_{s-1\le m<\infty} \frac{1}{m}\sum_{t=1}^{m} g(x_t) = \frac{1}{s-1}\sum_{t=1}^{s-1} g(x_t)$, in which case it is clearly less than $2^{-(s-1)} + \max_{s-1\le m\le n'} \frac{1}{m}\sum_{t=1}^{m} g(x_t)$, or else $\sup_{s-1\le m<\infty} \frac{1}{m}\sum_{t=1}^{m} g(x_t) = \sup_{s\le m<\infty} \frac{1}{m}\sum_{t=1}^{m} g(x_t)$, in which case (since $n' \ge n_s^g(\mathbf{x})$) it is at most $2^{-s} + \max_{s\le m\le n'} \frac{1}{m}\sum_{t=1}^{m} g(x_t) \le 2^{-(s-1)} + \max_{s-1\le m\le n'} \frac{1}{m}\sum_{t=1}^{m} g(x_t)$. Furthermore, we have $n_s^g(\mathbf{x}) \ge s \ge s-1$. Altogether, we have $n_{s-1}^g(\mathbf{x}) \le n_s^g(\mathbf{x})$, so that $n_s^g(\mathbf{x})$ is indeed nondecreasing in s.

For each $n \in \mathbb{N}$ with $n \ge n_1^g(\mathbf{x})$, let $s_n^g(\mathbf{x}) = \max\{s \in \{1,\dots,n\} : n \ge n_s^g(\mathbf{x})\}$; 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 $s_n^{\mathcal{G}}(\mathbf{x}) = \min_{g\in\mathcal{G}} s_n^g(\mathbf{x}) = \max\left(\left\{s \in \{1,\dots,n\} : n \ge \max_{g\in\mathcal{G}} n_s^g(\mathbf{x})\right\} \cup \{0\}\right)$. Since $n_s^g(\mathbf{x})$ is nondecreasing in s, we have for any $n, n' \in \mathbb{N}$ with $n' \ge n$, for $1 \le s \le s_n^{\mathcal{G}}(\mathbf{x})$, every $g \in \mathcal{G}$ has $n' \ge n_s^g(\mathbf{x})$, so that (13) is satisfied for every $g \in \mathcal{G}$, which implies

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

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

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

Furthermore, for any such sequence s_n, for every $g \in \mathcal{G}$, by definition

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

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

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

Altogether, the triangle inequality implies

$$\begin{aligned}
&\lim_{n\to\infty} \sup_{n'\ge n} \max_{g\in\mathcal{G}} \left| \hat{\mu}_{\mathbf{x}}(g) \;-\; \max_{s_n \le m \le n'} \frac{1}{m}\sum_{t=1}^{m} g(x_t) \right| \\
&\le \lim_{n\to\infty} \sup_{n'\ge n} \max_{g\in\mathcal{G}} \left| \sup_{s_n \le m < \infty} \frac{1}{m}\sum_{t=1}^{m} g(x_t) \;-\; \max_{s_n \le m \le n'} \frac{1}{m}\sum_{t=1}^{m} g(x_t) \right| \\
&\qquad + \lim_{n\to\infty} \max_{g\in\mathcal{G}} \left| \hat{\mu}_{\mathbf{x}}(g) \;-\; \sup_{s_n \le m < \infty} \frac{1}{m}\sum_{t=1}^{m} g(x_t) \right| = 0. \qquad (14)
\end{aligned}$$

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 (13) 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 $\mathbb{X}$. At this point, it may be tempting to use $s_n^{\mathcal{G}}(\mathbb{X})$ to complete the proof. However, recall that the lemma requires a *nonrandom* sequence s_n, whereas $s_n^{\mathcal{G}}(\mathbb{X})$ is a random variable. To address this, we will replace $s_n^{\mathcal{G}}(\mathbb{X})$ with an appropriate sequence of *confidence bounds* on its value, as follows. For any $n \in \mathbb{N}$ and $\delta \in (0, 1]$ define

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

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 $s \in \mathbb{N}$ and $g \in \mathcal{G}$ have $n_s^g(\mathbf{x}) < \infty$, and $\mathcal{G}$ is a finite set, we have $s_n^{\mathcal{G}}(\mathbf{x}) \to \infty$ for any sequence $\mathbf{x}$; thus, 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

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

for each $n \in \mathbb{N}$, we have that s_n is nondecreasing, and $s_n \to \infty$. Furthermore, by definition, we have $\mathbb{P}(s_n^{\mathcal{G}}(\mathbb{X}) \ge s_n) \ge 1 - 2^{-s_n}$, and since any $\delta \in (0, 1]$ has $s_n^{\mathcal{G}}(\delta) \le n$, the definition of s_n also implies $s_n \le n$. 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 $s_n \le n < \infty$, 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

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

Therefore, the Borel-Cantelli Lemma implies that, with probability one, for all sufficiently large k, we have $s^{\mathcal{G}}_{n_k}(\mathbb{X}) \geq s_{n_k}$. Furthermore, since $s^{\mathcal{G}}_n(\mathbb{X})$ 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 $s^{\mathcal{G}}_{n_k}(\mathbb{X}) \geq s_{n_k}$ for a given $k \in \mathbb{N}$, then $s^{\mathcal{G}}_n(\mathbb{X}) \geq s_n$ for every $n \in \{n_k, \ldots, n_{k+1}-1\}$. Combining this again with the fact that $s_n \to \infty$, we may conclude that, with probability one, for all sufficiently large $n \in \mathbb{N}$, we have $1 \leq s_n \leq s^{\mathcal{G}}_n(\mathbb{X})$. Thus, s_n almost surely satisfies the requirements for (14) to hold for $\mathbf{x} = \mathbb{X}$, which therefore implies

$$\lim_{n\to\infty} \sup_{n' \geq n} \max_{g \in \mathcal{G}} \left| \hat{\mu}_{\mathbb{X}}(g) \ - \max_{s_n \leq m \leq n'} \frac{1}{m} \sum_{t=1}^{m} g(X_t) \right| = 0 \text{ (a.s.).} \tag{15}$$

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 (15) with the dominated convergence theorem implies

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

■

Lemma 22 *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 $\forall n \in \mathbb{N}$, $m_{i_n} \leq n$ and*

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

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

$$\lim_{n\to\infty} \mathbb{E}\left[\sup_{n' \geq n} \max_{g \in \mathcal{G}_i} \left| \hat{\mu}_{\mathbb{X}}(g) \ - \max_{m_{i,n} \leq m \leq n'} \frac{1}{m} \sum_{t=1}^{m} g(X_t) \right| \right] = 0. \tag{16}$$

Such a sequence is guaranteed to exist by Lemma 21. From here, it would be straightforward to produce a sequence i_n satisfying $\mathbb{E}\left[\max_{g \in \mathcal{G}_{i_n}} \left| \hat{\mu}_{\mathbb{X}}(g) \ - \max_{m_{i_n,n} \leq m \leq n} \frac{1}{m} \sum_{t=1}^{m} g(X_t) \right| \right] \leq \gamma_{i_n}$, simply letting i_n grow sufficiently slowly. However, comparing this to the claim in the lemma, we require slightly more than this: namely, replacing $m_{i_n,n}$ with a single-index sequence

m_{i_n}. The existence of such a sequence m_{i_n} is enabled by the additional supremum in the expression on the left of (16). The basic idea is that this allows us to define a sequence n_i such that $\mathbb{E}\left[\max_{g\in\mathcal{G}_i}\left|\hat{\mu}_{\mathbb{X}}(g) - \max_{m_{i,n_i}\le m\le n}\frac{1}{m}\sum_{t=1}^{m} g(X_t)\right|\right] \le \gamma_i$ for any $n \ge n_i$. We may then conclude by defining $m_i = m_{i,n_i}$, and i_n maximal such that $n \ge n_{i_n}$. The formal argument becomes somewhat more technical in order to verify such sequences are well-defined and to satisfy the monotonicity requirements of m_i and i_n from the lemma. The details follow.

Formally, for each $n \in \mathbb{N}$, define

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

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

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

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

$$\mathbb{E}\left[\sup_{n'\ge n}\max_{g\in\mathcal{G}_i}\left|\hat{\mu}_{\mathbb{X}}(g) - \max_{m_{i,n}\le m\le n'}\frac{1}{m}\sum_{t=1}^{m} g(X_t)\right|\right] \le \gamma_i. \tag{17}$$

Furthermore, since

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

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 $\gamma_i > 0$, (16) implies that $\exists n_i' \in \mathbb{N}$ such that, $\forall n \ge n_i'$,

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

Therefore $j_n \ge i$ for every $n \ge \max\left\{i, \max_{1\le i'\le i} n_{i'}'\right\}$. 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

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

Note that, given the value $n_{i-1} \in \mathbb{N}$, the value n_i is well-defined since $\lim_{n\to\infty} j_n = \infty$ and $\lim_{n\to\infty} m_{i,n} = \infty$. 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 $m_{i+1} = m_{i+1,n_{i+1}} > m_{i,n_i} = m_i$, so that m_i is strictly increasing, with $m_i \to \infty$. Finally, for each $n \in \mathbb{N}$, define $i_n = \max\{i \in \{1,\dots,n\} : n_i \le n\}$. Since $n_1 = 1$, 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, $\forall n \in \mathbb{N}$, we have $n \geq n_{i_n}$, which also implies $m_{i_n} \leq n_{i_n} \leq n$ (by the assumed property that $m_{i,n} \leq n$ for any 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 (17) 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. Indeed, in the special case of binary classification, this is actually the *only* use of Condition 1 needed for the proof. For the case of general $(\mathcal{Y}, \ell)$, one additional use of Condition 1 (in Lemma 24 below) will be needed, to extend this lemma from set approximation to function approximation.

Lemma 23 *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). Letting $C_{kj} = \bigcap_{i=1}^{j} B_{ki}$ 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 $(\mathcal{X} \setminus G_k) \triangle (\mathcal{X} \setminus A) = G_k \triangle A$, by subadditivity of $\mathbb{E}[\hat{\mu}_{\mathbb{X}}(\cdot)]$ (Lemma 11), we have that

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

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, let $A = \bigcup_{i=1}^{\infty} A_i$, and fix any $\varepsilon > 0$. Letting $B_k = \bigcup_{i=1}^{k} A_i$ 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 $A_i \in \Lambda$). Let $C_{k_\varepsilon} = \bigcup_{i=1}^{k_\varepsilon} G_i$. 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 subadditivity of $\mathbb{E}[\hat{\mu}_{\mathbb{X}}(\cdot)]$ (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 $\mathcal{X} = \mathbb{R}^d$ ($d \in \mathbb{N}$) 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 23 from set approximation to function approximation, again using Condition 1.

Lemma 24 *There exists a countable set $\tilde{\mathcal{F}}$ of measurable functions $\mathcal{X} \to \mathcal{Y}$ such that, for every $\mathbb{X} \in \mathcal{C}_1$, for every measurable $f : \mathcal{X} \to \mathcal{Y}$,*

$$\inf_{\tilde{f} \in \tilde{\mathcal{F}}} \mathbb{E}\left[\hat{\mu}_{\mathbb{X}}(\ell(\tilde{f}(\cdot), f(\cdot)))\right] = 0.$$

Proof The proof will establish this claim for the set $\tilde{\mathcal{F}}$ of finite-depth *decision list* functions, where the decision region of each node is specified by an element from the countable set $\mathcal{T}_1$ (from Lemma 23) and the values are taken from a countable dense set $\tilde{\mathcal{Y}} \subseteq \mathcal{Y}$.

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 $f : \mathcal{X} \to \mathcal{Y}$, $\exists \tilde{f}_\varepsilon \in \tilde{\mathcal{F}}$ s.t. $\mathbb{E}\left[\hat{\mu}_{\mathbb{X}}(\ell(\tilde{f}_\varepsilon(\cdot), f(\cdot)))\right] < 3c_\ell \varepsilon$. The lemma will follow immediately from this (for this same set $\tilde{\mathcal{F}}$) by taking $\varepsilon \to 0$. Let $\mathcal{T}_1$ be as in Lemma 23, and let $\tilde{\mathcal{Y}} \subseteq \mathcal{Y}$ be a countable set with $\sup_{y \in \mathcal{Y}} \inf_{\tilde{y} \in \tilde{\mathcal{Y}}} \ell(\tilde{y}, y) = 0$; 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 $B_{\varepsilon,1} = \{y \in \mathcal{Y} : \ell(\tilde{y}_1, y) \leq \varepsilon\}$, and for each integer $i \geq 2$, inductively define $B_{\varepsilon,i} = \{y \in \mathcal{Y} : \ell(\tilde{y}_i, y) \leq \varepsilon\} \setminus \bigcup_{j=1}^{i-1} B_{\varepsilon,j}$. 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 $\mathbb{E}\left[\hat{\mu}_{\mathbb{X}}\left(\bigcup_{i=k_\varepsilon+1}^{\infty} C_{\varepsilon,i}\right)\right] < c_\ell \varepsilon/\bar{\ell}$.

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 23. 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 $x \in \mathcal{X} = \bigcup_{i=1}^{\infty} C_{\varepsilon,i}$,

$$\begin{aligned}
\ell(f(x), \tilde{f}_\varepsilon(x)) &\le \bar{\ell} \mathbb{1}_{\bigcup_{i=k_\varepsilon+1}^{\infty} C_{\varepsilon,i}}(x) + \sum_{i=1}^{k_\varepsilon} \ell(f(x), \tilde{f}_\varepsilon(x)) \mathbb{1}_{C_{\varepsilon,i}}(x) \\
&\le \bar{\ell} \mathbb{1}_{\bigcup_{i=k_\varepsilon+1}^{\infty} C_{\varepsilon,i}}(x) + \sum_{i=1}^{k_\varepsilon} c_\ell \left(\ell(f(x), \tilde{y}_i) \mathbb{1}_{C_{\varepsilon,i}}(x) + \ell(\tilde{y}_i, \tilde{f}_\varepsilon(x)) \mathbb{1}_{C_{\varepsilon,i}}(x)\right) \\
&\le \bar{\ell} \mathbb{1}_{\bigcup_{i=k_\varepsilon+1}^{\infty} C_{\varepsilon,i}}(x) + c_\ell \varepsilon + c_\ell \sum_{i=1}^{k_\varepsilon} \ell(\tilde{y}_i, \tilde{f}_\varepsilon(x)) \mathbb{1}_{C_{\varepsilon,i}}(x). \qquad (18)
\end{aligned}$$

Focusing now on the rightmost summation, let $[k_\varepsilon] = \{1, \ldots, k_\varepsilon\}$. If $x \notin \bigcup_{i\in[k_\varepsilon]} C_{\varepsilon,i}$ then this term is trivially zero due to the $\mathbb{1}_{C_{\varepsilon,i}}(x)$ factors. Otherwise, let $j \in [k_\varepsilon]$ be such that $x \in C_{\varepsilon,j}$; this j is unique by disjointness of the $C_{\varepsilon,i}$ sets, and for this same reason we have

$$\sum_{i=1}^{k_\varepsilon} \ell(\tilde{y}_i, \tilde{f}_\varepsilon(x)) \mathbb{1}_{C_{\varepsilon,i}}(x) = \ell(\tilde{y}_j, \tilde{f}_\varepsilon(x)) \mathbb{1}_{C_{\varepsilon,j}}(x). \qquad (19)$$

Now note that if $x \in A_{\varepsilon,j} \setminus \bigcup_{i\in[k_\varepsilon]\setminus\{j\}} A_{\varepsilon,i}$, then $\ell(\tilde{y}_j, \tilde{f}_\varepsilon(x)) = 0$. Thus, if $\ell(\tilde{y}_j, \tilde{f}_\varepsilon(x)) \mathbb{1}_{C_{\varepsilon,j}}(x) \neq 0$, then either $x \in C_{\varepsilon,j} \setminus A_{\varepsilon,j}$, or else $\exists i \in [k_\varepsilon]\setminus\{j\}$ with $x \in C_{\varepsilon,j} \cap A_{\varepsilon,i} \subseteq A_{\varepsilon,i} \setminus C_{\varepsilon,i}$ (where this last inclusion follows from $C_{\varepsilon,j} \cap C_{\varepsilon,i} = \emptyset$). Either way, we see that if $\ell(\tilde{y}_j, \tilde{f}_\varepsilon(x)) \mathbb{1}_{C_{\varepsilon,j}}(x) \neq 0$ then $\exists i \in [k_\varepsilon]$ with $x \in C_{\varepsilon,i} \triangle A_{\varepsilon,i}$, so that

$$\ell(\tilde{y}_j, \tilde{f}_\varepsilon(x)) \mathbb{1}_{C_{\varepsilon,j}}(x) \le \ell(\tilde{y}_j, \tilde{f}_\varepsilon(x)) \sum_{i=1}^{k_\varepsilon} \mathbb{1}_{C_{\varepsilon,i} \triangle A_{\varepsilon,i}}(x) \le \bar{\ell} \sum_{i=1}^{k_\varepsilon} \mathbb{1}_{C_{\varepsilon,i} \triangle A_{\varepsilon,i}}(x). \qquad (20)$$

Combining (19) and (20) and plugging back into (18) yields

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

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

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

The lemma now follows directly from this (together with non-negativity and symmetry of ℓ), since each $\tilde{f}_\varepsilon \in \tilde{\mathcal{F}}$, so that $\inf_{\tilde{f}\in\tilde{\mathcal{F}}}\mathbb{E}\Big[\hat{\mu}_{\mathbb{X}}(\ell(\tilde{f}(\cdot),f(\cdot)))\Big] \le \lim_{\varepsilon\to 0}\mathbb{E}\Big[\hat{\mu}_{\mathbb{X}}(\ell(\tilde{f}_\varepsilon(\cdot),f(\cdot)))\Big] \le \lim_{\varepsilon\to 0} 3c_\ell\varepsilon = 0$. ■

Remark: Before proceeding, we remark that since $\mathcal{T}_1$ in the proof of Lemma 23 is defined as the set of finite unions of elements of $\mathcal{T}_0$ (where $\mathcal{T}_0$ is any countable base for the topology $\mathcal{T}$), we can in fact represent any $f \in \tilde{\mathcal{F}}$ as a function $\tilde{f}(\cdot;\{y_i\}_{i=1}^k,\{A_i\}_{i=1}^k)$, $k\in\mathbb{N}$, $\{y_i\}_{i=1}^k \in \tilde{\mathcal{Y}}^k$, with $\{A_i\}_{i=1}^k \in \mathcal{T}_0^k$ (for $\tilde{f}(\cdot;\cdot,\cdot)$ and $\tilde{\mathcal{Y}}$ as defined in the above proof: that is, in the definition of $\tilde{\mathcal{F}}$ in the proof of Lemma 24, we can replace $\mathcal{T}_1$ with $\mathcal{T}_0$ and the set $\tilde{\mathcal{F}}$ remains unchanged. For instance, in the special case of $\mathcal{X}=\mathbb{R}^d$ ($d\in\mathbb{N}$) and $\mathcal{Y}=[0,1]$ with ℓ the squared loss ($\ell(a,b)=(a-b)^2$), we can take $\tilde{\mathcal{F}}$ as the set of rational-valued finite-depth decision lists, with the region of each decision node being a rational-centered rational-radius open ball.

We will use Lemma 24 via the following immediate implication.

Lemma 25 *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 Enumerate the elements of the countable set $\tilde{\mathcal{F}}$ from Lemma 24 as $\tilde{f}_1,\tilde{f}_2,\ldots$, and define $\mathcal{F}_i = \left\{\tilde{f}_1,\ldots,\tilde{f}_i\right\}$. With this definition, by Lemma 24, any $\mathbb{X}\in\mathcal{C}_1$ and measurable $f:\mathcal{X}\to\mathcal{Y}$ satisfy $\lim_{i\to\infty}\min_{f_i\in\mathcal{F}_i}\mathbb{E}[\hat{\mu}_{\mathbb{X}}(\ell(f_i(\cdot),f(\cdot)))] = \inf_{\tilde{f}\in\tilde{\mathcal{F}}}\mathbb{E}\Big[\hat{\mu}_{\mathbb{X}}(\ell(\tilde{f}(\cdot),f(\cdot)))\Big] = 0$. ■

Additionally, we have the following property for the f-approximating sequences of sets $\mathcal{F}_i$ implied by Lemma 25.

Lemma 26 *Fix any process $\mathbb{X}$ on $\mathcal{X}$, any measurable function $f:\mathcal{X}\to\mathcal{Y}$, any nondecreasing sequence $\{u_i\}_{i=1}^\infty$ in $\mathbb{N}$ with $u_i\to\infty$, and any sequence $\{\mathcal{F}_i\}_{i=1}^\infty$ of sets of measurable functions $\mathcal{X}\to\mathcal{Y}$ with $\mathcal{F}_1\subseteq\mathcal{F}_2\subseteq\cdots$ such that $\lim_{i\to\infty}\inf_{g\in\mathcal{F}_i}\mathbb{E}[\hat{\mu}_{\mathbb{X}}(\ell(g(\cdot),f(\cdot)))] = 0$. 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 $\forall i\ge\iota_0$,*

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

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 $\forall j \in \mathbb{N}$,

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

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

$$\mathbb{E}\left[\sup_{u_{i_k} \le m < \infty} \frac{1}{m}\sum_{t=1}^{m} \ell(g_{j_k}(X_t), f(X_t))\right] \le \mathbb{E}[\hat{\mu}_{\mathbb{X}}(\ell(g_{j_k}(\cdot), f(\cdot)))] + 4^{-k}\bar{\ell} < 2 \cdot 4^{-k}\bar{\ell}. \tag{21}$$

Also note that, since the leftmost expression in (21) is nonincreasing in i_k, we may choose $i_k > i_{k-1}$ if $k \ge 2$ (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 (21) for every $k \in \mathbb{N}$. We may then note that, by Markov's inequality,

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

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

$$\sup_{u_{i_k} \le m < \infty} \frac{1}{m}\sum_{t=1}^{m} \ell(g_{j_k}(X_t), f(X_t)) \le 2^{(1/2)-k}\sqrt{\bar{\ell}}. \tag{22}$$

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

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

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 $\max\{i_0, j_0\} = 1 \le i$, so that the set on the right hand side is nonempty, and furthermore, since $\{i_k\}_{k=0}^{\infty}$ is strictly increasing, every $k \ge i$ has $\max\{i_k, j_k\} > i$, so that the set is finite, and hence has a maximum element. Also, since i_k and j_k are finite for every k, we have that $\lim_{i\to\infty} k_i = \infty$. In particular, this implies that, on the event K, $\exists \iota_0 \in \mathbb{N}$ such that $\forall i \ge \iota_0$, $k_i \ge \kappa_0$, so that (22) implies

$$\sup_{u_{i_{k_i}} \le m < \infty} \frac{1}{m}\sum_{t=1}^{m} \ell(g_{j_{k_i}}(X_t), f(X_t)) \le \alpha_i. \tag{23}$$

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 (23), 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 27 $\mathcal{C}_1 \subseteq$ SUIL.

Proof Suppose $\mathbb{X} \in \mathcal{C}_1$. Lemma 25 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 $f^\star : \mathcal{X} \to \mathcal{Y}$, $\lim_{i\to\infty} \min_{g_i \in \mathcal{G}_i} \mathbb{E}[\hat{\mu}_{\mathbb{X}}(\ell(g_i(\cdot), f^\star(\cdot)))] = 0$. Furthermore, applying Lemma 22 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 $\forall n \in \mathbb{N}$, $m_{i_n} \leq n$ and

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

Let $I = \{i_n : n \in \mathbb{N}\}$, and for each $i \in I$, define $n_i = \min\{n \in \mathbb{N} : i_n = i\}$. Markov's inequality and (24) imply

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

Therefore, the Borel-Cantelli Lemma implies that there exists an event K' of probability one, on which $\exists \iota_1 \in \mathbb{N}$ such that $\forall i \in I$ with $i \geq \iota_1$,

$$\max_{f,g \in \mathcal{G}_i} \left(\hat{\mu}_{\mathbb{X}}(\ell(f(\cdot), g(\cdot))) \; - \max_{m_i \leq m \leq n_i} \frac{1}{m} \sum_{t=1}^{m} \ell(f(X_t), g(X_t)) \right) \leq \sqrt{\gamma_i}. \tag{25}$$

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

$$\max_{m_{i_n} \leq m \leq n} \frac{1}{m} \sum_{t=1}^{m} \ell(f(X_t), g(X_t)) \geq \max_{m_{i_n} \leq m \leq n_{i_n}} \frac{1}{m} \sum_{t=1}^{m} \ell(f(X_t), g(X_t)). \tag{26}$$

Furthermore, since $i_n \to \infty$, on the event K', $\exists \nu_1 \in \mathbb{N}$ such that $\forall n \geq \nu_1$, we have $i_n \geq \iota_1$, so that (25) and (26) imply

$$\max_{f,g \in \mathcal{G}_{i_n}} \left(\hat{\mu}_{\mathbb{X}}(\ell(f(\cdot), g(\cdot))) \; - \max_{m_{i_n} \leq m \leq n} \frac{1}{m} \sum_{t=1}^{m} \ell(f(X_t), g(X_t)) \right) \leq \sqrt{\gamma_{i_n}}. \tag{27}$$

Now consider using the inductive learning rule $\hat{f}_n$ defined in (12), with $\mathcal{F}_n = \mathcal{G}_{i_n}$ and $\hat{m}_n = m_{i_n}$ for each $n \in \mathbb{N}$. Fix any measurable function $f^\star : \mathcal{X} \to \mathcal{Y}$. By the defining properties of the $\mathcal{G}_i$ sequence, and the fact that m_i is nondecreasing with $\lim_{i \to \infty} m_i = \infty$, Lemma 26 implies that there exists a (nonrandom) sequence $\{f_i^\star\}_{i=1}^\infty$ with $f_i^\star \in \mathcal{G}_i$ 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 $\forall i \geq \iota_0$,

$$\sup_{m_i \leq m < \infty} \frac{1}{m} \sum_{t=1}^{m} \ell(f_i^\star(X_t), f^\star(X_t)) \leq \alpha_i. \tag{28}$$

On this event, let $\nu_0 \in \mathbb{N}$ be a value such that $\forall n \in \mathbb{N}$ with $n \geq \nu_0$, we have $i_n \geq \iota_0$; such a ν_0 exists since $\lim_{n \to \infty} i_n = \infty$.

For brevity, define $\hat{g}_n(\cdot) = \hat{f}_n(X_{1:n}, f^\star(X_{1:n}), \cdot)$ for every $n \in \mathbb{N}$. Note that, by the definition of $\hat{f}_n$ from (12) and the fact that $f_n^\star \in \mathcal{F}_n = \mathcal{G}_{i_n}$ and $\hat{m}_n = m_{i_n}$, $\forall n \in \mathbb{N}$, we have

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

Thus, on the event K, $\forall n \in \mathbb{N}$ with $n \geq \nu_0$, (28) implies

$$\max_{m_{i_n} \leq m \leq n} \frac{1}{m} \sum_{t=1}^{m} \ell(\hat{g}_n(X_t), f^\star(X_t)) \leq \alpha_{i_n}. \tag{29}$$

Now suppose the event $K \cap K'$ occurs and fix any $n \in \mathbb{N}$ with $n \geq \max\{\nu_0, \nu_1\}$. The relaxed triangle inequality and subadditivity of $\hat{\mu}_{\mathbb{X}}$ (Lemma 8) imply

$$\hat{\mu}_{\mathbb{X}}(\ell(\hat{g}_n(\cdot), f^\star(\cdot))) \leq c_\ell \hat{\mu}_{\mathbb{X}}(\ell(\hat{g}_n(\cdot), f_{i_n}^\star(\cdot))) + c_\ell \hat{\mu}_{\mathbb{X}}(\ell(f_{i_n}^\star(\cdot), f^\star(\cdot))). \tag{30}$$

Furthermore, since $\hat{g}_n$ and $f_{i_n}^\star$ are both elements of $\mathcal{G}_{i_n}$, and since the event K' holds and $n \geq \nu_1$, the inequality (27) implies

$$\hat{\mu}_{\mathbb{X}}(\ell(\hat{g}_n(\cdot), f_{i_n}^\star(\cdot))) \leq \max_{m_{i_n} \leq m \leq n} \frac{1}{m} \sum_{t=1}^{m} \ell(\hat{g}_n(X_t), f_{i_n}^\star(X_t)) + \sqrt{\gamma_{i_n}}. \tag{31}$$

Then the relaxed triangle inequality and symmetry of ℓ, together with subadditivity of the max, imply

$$\begin{aligned} &\max_{m_{i_n} \leq m \leq n} \frac{1}{m} \sum_{t=1}^{m} \ell(\hat{g}_n(X_t), f_{i_n}^\star(X_t)) \\ &\leq c_\ell \max_{m_{i_n} \leq m \leq n} \frac{1}{m} \sum_{t=1}^{m} \ell(\hat{g}_n(X_t), f^\star(X_t)) + c_\ell \max_{m_{i_n} \leq m \leq n} \frac{1}{m} \sum_{t=1}^{m} \ell(f_{i_n}^\star(X_t), f^\star(X_t)). \end{aligned} \tag{32}$$

Also, since m_{i_n} is finite, we generally have

$$\hat{\mu}_{\mathbb{X}}(\ell(f^\star_{i_n}(\cdot), f^\star(\cdot))) \leq \sup_{m_{i_n} \leq m < \infty} \frac{1}{m} \sum_{t=1}^{m} \ell(f^\star_{i_n}(X_t), f^\star(X_t)).$$

Combining this with (31) and (32) and plugging into (30) yields

$$\begin{aligned}&\hat{\mu}_{\mathbb{X}}(\ell(\hat{g}_n(\cdot), f^\star(\cdot)))\\ &\leq c_\ell^2 \max_{m_{i_n} \leq m \leq n} \frac{1}{m} \sum_{t=1}^{m} \ell(\hat{g}_n(X_t), f^\star(X_t)) + c_\ell(c_\ell+1) \sup_{m_{i_n} \leq m < \infty} \frac{1}{m} \sum_{t=1}^{m} \ell(f^\star_{i_n}(X_t), f^\star(X_t)) + c_\ell \sqrt{\gamma_{i_n}}.\end{aligned}$$

Since the event K holds and $n \geq \nu_0$, the inequalities (29) and (28) provide upper bounds on the first two terms above, respectively, so that altogether we have

$$\hat{\mu}_{\mathbb{X}}(\ell(\hat{g}_n(\cdot), f^\star(\cdot))) \leq c_\ell^2 \alpha_{i_n} + c_\ell(c_\ell+1)\alpha_{i_n} + c_\ell\sqrt{\gamma_{i_n}} = c_\ell(2c_\ell+1)\alpha_{i_n} + c_\ell\sqrt{\gamma_{i_n}}.$$

In particular, recall that $i_n \to \infty$ and $\lim_{i\to\infty} \alpha_i = \lim_{i\to\infty} \gamma_i = 0$, so that the rightmost expression above converges to 0 as $n \to \infty$. Thus, on the event $K \cap K'$, since $\max\{\nu_0, \nu_1\} < \infty$, we have that

$$\limsup_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(\hat{f}_n, f^\star; n) = \limsup_{n\to\infty} \hat{\mu}_{\mathbb{X}}(\ell(\hat{g}_n(\cdot), f^\star(\cdot))) \leq \lim_{n\to\infty} c_\ell(2c_\ell+1)\alpha_{i_n} + c_\ell\sqrt{\gamma_{i_n}} = 0.$$

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 19, 20, and 27 completes the proof of Theorem 7.

Interestingly, we may note that the *only* reliance of the above proof of Lemma 27 on the assumption $\mathbb{X} \in \mathcal{C}_1$ is in the existence of the set $\tilde{\mathcal{F}}$ from Lemma 24 (used here via its implication in Lemma 25): that is, we have in fact established that any $\mathbb{X}$ for which there exists a countable set $\tilde{\mathcal{F}}$ with these properties admits strong universal inductive learning, so that the existence of such a set implies $\mathbb{X} \in$ SUIL. Together with Theorem 7 (implying $\mathcal{C}_1 =$ SUIL) and Lemma 24 (implying $\mathbb{X} \in \mathcal{C}_1$ suffices for such a set $\tilde{\mathcal{F}}$ to exist), this establishes that $\mathcal{C}_1$ is in fact *equivalent* to the set of processes for which such a set exists (and hence so are SUIL and SUAL, via Theorem 7). Thus, we have yet another useful equivalent way of expressing Condition 1, stated formally in the following corollary.

Corollary 28 *A process $\mathbb{X}$ satisfies Condition 1 if and only if there exists a countable set $\tilde{\mathcal{G}}$ of measurable functions $\mathcal{X} \to \mathcal{Y}$ such that, for every measurable $f : \mathcal{X} \to \mathcal{Y}$, $\inf_{\tilde{g}\in\tilde{\mathcal{G}}} \mathbb{E}[\hat{\mu}_{\mathbb{X}}(\ell(\tilde{g}(\cdot), f(\cdot)))] = 0$.*

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: namely, a process $\mathbb{X}$ satisfies Condition 1 if and only if there exists a countable set $\mathcal{T}_2 \subseteq \mathcal{B}$ with $\sup_{A\in\mathcal{B}} \inf_{G\in\mathcal{T}_2} \mathbb{E}[\hat{\mu}_{\mathbb{X}}(G \triangle A)] = 0$. Recall that this was the guarantee for the set $\mathcal{T}_1$ from Lemma 23. However, Lemma 23 also guarantees the stronger property that this *same* set $\mathcal{T}_1$ can serve as the above set $\mathcal{T}_2$ for *every* $\mathbb{X}$ satisfying Condition 1. Similarly, the set $\tilde{\mathcal{F}}$ supplied by Lemma 24 is also defined independent of $\mathbb{X}$, so that this same set $\tilde{\mathcal{F}}$ can serve as the set $\tilde{\mathcal{G}}$ in Corollary 28 for every $\mathbb{X}$ satisfying Condition 1. This universality of $\mathcal{T}_1$ and $\tilde{\mathcal{F}}$ will be crucial in the next section when discussing *optimistically* universal learning.

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 25. 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

$$\hat{i}_{n,m}(x_{1:m}) = \max\Bigg\{ i \in \mathbb{N} : u_i \leq n \text{ and } \max_{f,g\in\mathcal{F}_i} \left(\max_{u_i\leq s\leq m} \frac{1}{s}\sum_{t=1}^{s} \ell(f(x_t), g(x_t)) - \max_{u_i\leq s\leq n} \frac{1}{s}\sum_{t=1}^{s} \ell(f(x_t), g(x_t)) \right) \leq \gamma_i \Bigg\}. \tag{33}$$

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). 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

$$\operatorname*{argmin}_{f\in\mathcal{F}_{\hat{i}_{n,m}(x_{1:m})}} \max_{u_{\hat{i}_{n,m}(x_{1:m})}\leq s\leq n} \frac{1}{s}\sum_{t=1}^{s} \ell(f(x_t), y_t). \tag{34}$$

We break ties in the argmin based on a fixed preference ordering of $\mathcal{F}_i$. Since the sets $\mathcal{F}_i$ are finite, one can easily verify that this makes $\hat{f}_{n,m}$ a measurable function, and hence (34) defines a valid self-adaptive learning rule. 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.

The essential difference between the self-adaptive learning rule (34) and the inductive learning rule (12) is that the self-adaptive rule uses the sequence of test samples $X_{1:m}$ for the *model selection* component, selecting which class $\mathcal{F}_i$ to use in the optimization in (34), whereas (12) uses a *distribution-dependent* selection. Specifically, the self-adaptive rule replaces the distribution-dependent value i_n from Lemma 22, used in the proof of Lemma 27, with a data-dependent value $\hat{i}_{n,m}(X_{1:m})$, thus removing all dependence on the distribution of $\mathbb{X}$. In the proof of Lemma 27, the value i_n is chosen to guarantee (via Lemma 22) that the estimator $\max_{m_{i_n} \leq m \leq n} \frac{1}{m} \sum_{t=1}^{m} \ell(f(X_t), g(X_t))$ is close to $\hat{\mu}_{\mathbb{X}}(\ell(f(\cdot), g(\cdot)))$ uniformly over all f, g in the class $\mathcal{G}_{i_n}$ defined in the proof. The value $\hat{i}_{n,m}(X_{1:m})$ in (33) is designed to provide this guarantee *directly*. Specifically, in the analysis of $\hat{f}_{n,m}$ below, the value $\hat{i}_{n,m}(X_{1:m})$ ensures (essentially) that for large m, for all $f, g \in \mathcal{F}_{\hat{i}_{n,m}(X_{1:m})}$, $\max_{u_{\hat{i}_{n,m}(X_{1:m})} \leq s \leq n} \frac{1}{s} \sum_{t=1}^{s} \ell(f(X_t), g(X_t))$ is close to $\hat{\mu}_{\mathbb{X}}(\ell(f(\cdot), g(\cdot)))$. These can then be related to the losses relative to $f^\star$ via relaxed triangle inequalities and the approximation guarantees from Lemma 26, to conclude that the function $f \in \mathcal{F}_{\hat{i}_{n,m}(X_{1:m})}$ minimizing $\max_{u_{\hat{i}_{n,m}(X_{1:m})} \leq s \leq n} \frac{1}{s} \sum_{t=1}^{s} \ell(f(X_t), f^\star(X_t))$ achieves a relatively small value of $\hat{\mu}_{\mathbb{X}}(\ell(f(\cdot), f^\star(\cdot)))$. In both the inductive and self-adaptive cases, this approach is analogous to the traditional principles of model selection, whereby we constrain the function class so that empirical estimates of the risk are close enough to the corresponding population risks to guarantee that optimizing the estimate yields a function with relatively small population risk, but while also allowing the constraint to become less restrictive as n grows, to admit increasingly good approximations of $f^\star$.

As discussed in the proofs of Lemmas 23, 24, and 25, and the remark following the proof of Lemma 24, the sets $\mathcal{F}_i$ can be constructed based on an enumeration of finite-depth decision lists, with the region of each decision node being an element of a countable base for the topology $\mathcal{T}$, and with values from a countable dense subset of $\mathcal{Y}$. For instance, in the special case of $\mathcal{X} = \mathbb{R}^d$ ($d \in \mathbb{N}$) with the Euclidean topology, and $\mathcal{Y} = [0, 1]$ with the squared loss ($\ell(a, b) = (a-b)^2$), we can let $\tilde{f}_1, \tilde{f}_2, \ldots$ be an enumeration of the rational-valued finite-depth decision lists, with the region of each decision node being a rational-centered rational-radius open ball. Then we can let $\mathcal{F}_i = \{\tilde{f}_1, \ldots, \tilde{f}_i\}$ for each $i \in \mathbb{N}$. In this case, in principle the learning rule $\hat{f}_{n,m}$ can be approximated by a digital computer (up to some finite precision for the points and predictions).

Continuing with the general case, we have the following theorem for this $\hat{f}_{n,m}$ rule.

Theorem 29 *The self-adaptive learning rule $\hat{f}_{n,m}$ is optimistically universal.*

Proof The proof proceeds along similar lines to that of Lemma 27, 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 27. 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 $f, g \in \mathcal{F}_i$, $\max_{u_i \le s \le m} \frac{1}{s} \sum_{t=1}^{s} \ell(f(X_t), g(X_t))$ 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 $m_n^* \ge n$ such that $\forall m \ge m_n^*$, $\hat{i}_{n,m}(X_{1:m}) = \hat{i}_{n,m_n^*}(X_{1:m_n^*})$. 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 $m \ge m_n^*$ satisfies

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

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

$$\max_{f,g \in \mathcal{F}_{\hat{i}_n}} \left(\sup_{u_{\hat{i}_n} \le s < \infty} \frac{1}{s} \sum_{t=1}^{s} \ell(f(X_t), g(X_t)) - \max_{u_{\hat{i}_n} \le s \le n} \frac{1}{s} \sum_{t=1}^{s} \ell(f(X_t), g(X_t)) \right) \le \gamma_{\hat{i}_n}. \tag{35}$$

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

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

Together with finiteness of every u_i, this implies

$$\lim_{n \to \infty} \hat{i}_n = \infty. \tag{36}$$

Next note that, by our choices of the sequences $\{\mathcal{F}_i\}_{i=1}^{\infty}$ and $\{u_i\}_{i=1}^{\infty}$, Lemma 26 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 $\forall i \ge \iota_0$,

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

In particular, since $\lim_{n \to \infty} \hat{i}_n = \infty$ by (36), this implies that, on the event K, $\exists \nu_0 \in \mathbb{N}$ such that $\forall n \ge \nu_0$, we have $\hat{i}_n \ge \iota_0$, so that the above implies

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

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 that ties are broken in the argmin based on a fixed ordering). Define $\hat{g}_n = \hat{g}_{n,m_n^*}$. Combining the definition of $\hat{f}_{n,m_n^*}$ with (37) we have that, on the event K, $\forall n \in \mathbb{N}$ with $n \geq \nu_0$,

$$\begin{aligned}\max_{u_{\hat{i}_n} \leq s \leq n} \frac{1}{s} \sum_{t=1}^{s} \ell(\hat{g}_n(X_t), f^\star(X_t)) &\leq \max_{u_{\hat{i}_n} \leq s \leq n} \frac{1}{s} \sum_{t=1}^{s} \ell(f^\star_{\hat{i}_n}(X_t), f^\star(X_t)) \\ &\leq \sup_{u_{\hat{i}_n} \leq s < \infty} \frac{1}{s} \sum_{t=1}^{s} \ell(f^\star_{\hat{i}_n}(X_t), f^\star(X_t)) \leq \alpha_{\hat{i}_n}. \end{aligned} \tag{38}$$

Now suppose the event K occurs and fix any $n \in \mathbb{N}$ with $n \geq \nu_0$. The relaxed triangle inequality and subadditivity of the supremum imply

$$\begin{aligned}&\sup_{u_{\hat{i}_n} \leq s < \infty} \frac{1}{s} \sum_{t=1}^{s} \ell(\hat{g}_n(X_t), f^\star(X_t)) \\ &\leq c_\ell \sup_{u_{\hat{i}_n} \leq s < \infty} \frac{1}{s} \sum_{t=1}^{s} \ell(\hat{g}_n(X_t), f^\star_{\hat{i}_n}(X_t)) + c_\ell \sup_{u_{\hat{i}_n} \leq s < \infty} \frac{1}{s} \sum_{t=1}^{s} \ell(f^\star_{\hat{i}_n}(X_t), f^\star(X_t)). \end{aligned} \tag{39}$$

Since $\hat{g}_n$ and $f^\star_{\hat{i}_n}$ are both elements of $\mathcal{F}_{\hat{i}_n}$, (35) implies

$$\sup_{u_{\hat{i}_n} \leq s < \infty} \frac{1}{s} \sum_{t=1}^{s} \ell(\hat{g}_n(X_t), f^\star_{\hat{i}_n}(X_t)) \leq \max_{u_{\hat{i}_n} \leq s \leq n} \frac{1}{s} \sum_{t=1}^{s} \ell(\hat{g}_n(X_t), f^\star_{\hat{i}_n}(X_t)) + \gamma_{\hat{i}_n}. \tag{40}$$

The relaxed triangle inequality and symmetry of ℓ, together with subadditivity of the max, then imply

$$\begin{aligned}&\max_{u_{\hat{i}_n} \leq s \leq n} \frac{1}{s} \sum_{t=1}^{s} \ell(\hat{g}_n(X_t), f^\star_{\hat{i}_n}(X_t)) \\ &\leq c_\ell \max_{u_{\hat{i}_n} \leq s \leq n} \frac{1}{s} \sum_{t=1}^{s} \ell(\hat{g}_n(X_t), f^\star(X_t)) + c_\ell \max_{u_{\hat{i}_n} \leq s \leq n} \frac{1}{s} \sum_{t=1}^{s} \ell(f^\star_{\hat{i}_n}(X_t), f^\star(X_t)). \end{aligned}$$

Combining this with (40) and plugging into (39) yields

$$\begin{aligned}&\sup_{u_{\hat{i}_n} \leq s < \infty} \frac{1}{s} \sum_{t=1}^{s} \ell(\hat{g}_n(X_t), f^\star(X_t)) \\ &\leq c_\ell^2 \max_{u_{\hat{i}_n} \leq s \leq n} \frac{1}{s} \sum_{t=1}^{s} \ell(\hat{g}_n(X_t), f^\star(X_t)) + c_\ell(c_\ell + 1) \sup_{u_{\hat{i}_n} \leq s < \infty} \frac{1}{s} \sum_{t=1}^{s} \ell(f^\star_{\hat{i}_n}(X_t), f^\star(X_t)) + c_\ell \gamma_{\hat{i}_n}. \end{aligned}$$

Since the event K holds and $n \geq \nu_0$, the inequalities (38) and (37) provide upper bounds on the first two terms above, respectively, so that altogether we have

$$\sup_{u_{\hat{i}_n} \leq s < \infty} \frac{1}{s} \sum_{t=1}^{s} \ell(\hat{g}_n(X_t), f^\star(X_t)) \leq c_\ell(2c_\ell + 1)\alpha_{\hat{i}_n} + c_\ell \gamma_{\hat{i}_n}. \tag{41}$$

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

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

Combined with (41), this implies that, on the event K, every $n \in \mathbb{N}$ with $n \geq \nu_0$ satisfies

$$\hat{\mathcal{L}}_{\mathbb{X}}\Big(\hat{f}_{n,\cdot}, f^\star; n\Big) \leq c_\ell(2c_\ell + 1)\alpha_{\hat{i}_n} + c_\ell \gamma_{\hat{i}_n}.$$

Recalling that (by their definitions) $\lim_{i\to\infty} \alpha_i = 0$ and $\lim_{i\to\infty} \gamma_i = 0$, and that $\lim_{n\to\infty} \hat{i}_n = \infty$ by (36), we have that on the event K,

$$\limsup_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}\Big(\hat{f}_{n,\cdot}, f^\star; n\Big) \leq \lim_{n\to\infty} c_\ell(2c_\ell + 1)\alpha_{\hat{i}_n} + c_\ell \gamma_{\hat{i}_n} = 0.$$

Since the event K has probability one, and $\hat{\mathcal{L}}_{\mathbb{X}}$ is nonnegative, this establishes that $\hat{\mathcal{L}}_{\mathbb{X}}\Big(\hat{f}_{n,\cdot}, f^\star; n\Big) \to 0$ (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: that is, $\hat{f}_{n,m}$ is optimistically universal. ■

An immediate consequence of Theorem 29 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 30 *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 $\lambda\left(U_\varepsilon \setminus \bigcup_{i=1}^{j_\delta} (a_i, b_i)\right) < \delta/2$. Let $k_\delta = \left\lceil \log_2\left(\frac{4j_\delta}{\delta}\right)\right\rceil$. Since $\lambda(U_\varepsilon) < \infty$, we know that every i has $a_i > -\infty$ and $b_i < \infty$. Also, 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

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

Thus,

$$\begin{aligned}
&\lambda\Big(U_\varepsilon \setminus \bigcup\{C_{k_\delta,t} : C_{k_\delta,t} \subseteq U_\varepsilon, t \in \mathbb{Z}\}\Big) \\
&\leq \lambda\left(U_\varepsilon \setminus \bigcup_{i=1}^{j_\delta}(a_i,b_i)\right) + \lambda\left(\bigcup_{i=1}^{j_\delta}(a_i,b_i) \setminus \bigcup\{C_{k_\delta,t} : C_{k_\delta,t} \subseteq U_\varepsilon, t \in \mathbb{Z}\}\right) \\
&< \delta/2 + \sum_{i=1}^{j_\delta} \lambda\Big((a_i,b_i) \setminus \bigcup\{C_{k_\delta,t} : C_{k_\delta,t} \subseteq U_\varepsilon, t \in \mathbb{Z}\}\Big) \\
&\leq \delta/2 + \sum_{i=1}^{j_\delta} \lambda\Big((a_i,b_i) \setminus \bigcup\{C_{k_\delta,t} : C_{k_\delta,t} \subseteq (a_i,b_i), t \in \mathbb{Z}\}\Big) \leq \delta/2 + \sum_{i=1}^{j_\delta} \frac{\delta}{2j_\delta} = \delta. \qquad (42)
\end{aligned}$$

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)

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

Therefore $\forall k > k_\delta$,

$$\begin{aligned}
&\lambda(U_\varepsilon \cap B_k) \\
&= \lambda\Big(B_k \cap \bigcup\{C_{k_\delta,i} : C_{k_\delta,i} \subseteq U_\varepsilon, i \in \mathbb{Z}\}\Big) + \lambda\Big(B_k \cap U_\varepsilon \setminus \bigcup\{C_{k_\delta,i} : C_{k_\delta,i} \subseteq U_\varepsilon, i \in \mathbb{Z}\}\Big) \\
&= (1/2)\lambda\Big(\bigcup\{C_{k_\delta,i} : C_{k_\delta,i} \subseteq U_\varepsilon, i \in \mathbb{Z}\}\Big) + \lambda\Big(B_k \cap U_\varepsilon \setminus \bigcup\{C_{k_\delta,i} : C_{k_\delta,i} \subseteq U_\varepsilon, i \in \mathbb{Z}\}\Big). \qquad (43)
\end{aligned}$$

The first term in (43) 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 (42) is greater than $(1/2)\lambda(U_\varepsilon) - \delta/2$. Furthermore, the second term in (43) 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,

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

where this last inequality is by (42).

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

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

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

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

and

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

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 learning rule converging to the stationary regression function 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* have convergent relative frequencies (CRF) and are ergodic (indeed, they are *product* processes). Thus, this establishes the stronger fact that (when $\mathcal{X}$ is uncountable) there is no inductive learning rule that is strongly universally consistent for every ergodic $\mathbb{X} \in$ CRF (as stated in Corollary 32 below); this suffices to establish Theorem 6 since Theorems 18 and 7 imply CRF $\subseteq$ SUIL.

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 30 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. Let $\mathbf{U} = \{U_{k,j}\}_{k,j\in\mathbb{N}}$, $\mathbf{V} = \{V_{k,j}\}_{k,j\in\mathbb{N}}$, $\mathbf{W} = \{W_j\}_{j\in\mathbb{N}}$.

Fix any $y_0, y_1 \in \mathcal{Y}$ with $\ell(y_0, y_1) > 0$, and define $\Delta_{01} = \ell(y_0, y_1)/(2c_\ell)$. Importantly, the near-metric properties of ℓ imply that any $y \in \mathcal{Y}$ with $\ell(y, y_1) < \Delta_{01}$ necessarily has $\ell(y, y_0) > \ell(y, y_0) + \ell(y, y_1) - \Delta_{01} \geq \ell(y_0, y_1)/c_\ell - \Delta_{01} = \Delta_{01}$, where the second inequality is due to the relaxed triangle inequality and symmetry. Thus, it is not possible to simultaneously achieve $\ell(y, y_1) < \Delta_{01}$ and $\ell(y, y_0) \leq \Delta_{01}$.

For any array $\mathbf{v} = \{v_{k,j}\}_{k,j\in\mathbb{N}}$, and any $K \in \mathbb{N}$, define $\mathbf{v}_{<K} = \{v_{k,j}\}_{k,j\in\mathbb{N},k<K}$, and define $\mathbf{v}_K = \{v_{K,j}\}_{j\in\mathbb{N}}$. 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. Note that the above functions are measurable, since each R_K is measurable, and $\mathbf{v}$, $\mathbf{v}_{<K}$, $\mathbf{w}$, $\mathbf{u}_{<K}$ are all countable and hence measurable (recalling that singleton sets $\{x\}$ are closed, hence measurable).

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

$$\mathbb{P}\Big(\pi_0\Big(\Big\{x : \ell\Big(f_n\Big(\hat{X}_{1:n}^{(k)}, f_k^\star\Big(\hat{X}_{1:n}^{(k)}; \mathbf{U}_{<k}, \mathbf{V}_{<k}, \mathbf{W}\Big), x\Big), y_0\Big) \geq \Delta_{01}\Big\}\Big) \geq 3/4\Big) < 2^{-k}, \quad (44)$$

then fix the minimum such 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 (44), 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 (44), 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 (44) 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 (e.g., based on Fatou's lemma), defining $\mathbf{Q}_{k^*} = (\mathbf{U}_{<k^*}, \mathbf{V}_{<k^*}, \mathbf{W})$,

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{P}\left(\limsup_{n\to\infty}\left\{\pi_0\left(\left\{x : \ell\left(f_n\left(\hat{X}_{1:n}^{(k^*)}, f_{k^*}^\star\left(\hat{X}_{1:n}^{(k^*)}; \mathbf{Q}_{k^*}\right), x\right), y_0\right) \geq \Delta_{01}\right\}\right) \geq 3/4\right\} \,\middle|\, \mathbf{Q}_{k^*}\right)\right] \\
&= \mathbb{P}\left(\limsup_{n\to\infty}\left\{\pi_0\left(\left\{x : \ell\left(f_n\left(\hat{X}_{1:n}^{(k^*)}, f_{k^*}^\star\left(\hat{X}_{1:n}^{(k^*)}; \mathbf{Q}_{k^*}\right), x\right), y_0\right) \geq \Delta_{01}\right\}\right) \geq 3/4\right\}\right) \\
&\geq \limsup_{n\to\infty} \mathbb{P}\left(\pi_0\left(\left\{x : \ell\left(f_n\left(\hat{X}_{1:n}^{(k^*)}, f_{k^*}^\star\left(\hat{X}_{1:n}^{(k^*)}; \mathbf{Q}_{k^*}\right), x\right), y_0\right) \geq \Delta_{01}\right\}\right) \geq 3/4\right).
\end{aligned}$$

The negation of (44) implies this last expression is at least 2^{-k^*} (noting that the negation of (44) 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 $f^\star(\cdot) = f_{k^*}^\star(\cdot; \mathbf{u}_{<k^*}, \mathbf{v}_{<k^*}, \mathbf{w})$, we have

$$\mathbb{P}\left(\limsup_{n\to\infty}\left\{\pi_0(\{x : \ell(f_n(X_{1:n}, f^\star(X_{1:n}), x), y_0) \geq \Delta_{01}\}) \geq 3/4\right\}\right) \geq 2^{-k^*}.$$

Define the event

$$E' = \left\{\limsup_{n\to\infty} \pi_0(\{x \in R_{k^*} : \ell(f_n(X_{1:n}, f^\star(X_{1:n}), x), y_0) \geq \Delta_{01}\}) \geq 1/4\right\}.$$

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

$$\begin{aligned}
&\limsup_{n\to\infty}\left\{\pi_0(\{x : \ell(f_n(X_{1:n}, f^\star(X_{1:n}), x), y_0) \geq \Delta_{01}\}) \geq 3/4\right\} \\
&\subseteq \limsup_{n\to\infty}\left\{\pi_0(\{x \in R_{k^*} : \ell(f_n(X_{1:n}, f^\star(X_{1:n}), x), y_0) \geq \Delta_{01}\}) \geq 1/4\right\} \subseteq E',
\end{aligned}$$

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.

Define $t_i = n_{k^*-1} + k^*(i-1) + 1$ for each $i \in \mathbb{N}$, and let $I_{k^*} = \{t_i : i \in \mathbb{N}\}$. Note that, since every $V_{k^*,j}$ is 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

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

Since $k^* \sum_{t=n+1}^{n+m} \mathbb{1}_{I_{k^*}}(t) > m - 2k^*$, letting $i_n = \max\{i \in \mathbb{N} : t_i \leq n\}$, the right hand side above is at least as large as

$$\begin{aligned}
&\limsup_{s\to\infty} \frac{1}{k^* s + 2k^*} \sum_{j=1}^{s} \ell\big(f_n\big(X_{1:n}, f^\star(X_{1:n}), X_{t_{i_n+j}}\big), y_0\big) \\
&= \limsup_{s\to\infty} \frac{1}{k^* s} \sum_{j=1}^{s} \ell\big(f_n\big(X_{1:n}, f^\star(X_{1:n}), X_{t_{i_n+j}}\big), y_0\big) \\
&\geq \limsup_{s\to\infty} \frac{1}{k^* s} \sum_{j=1}^{s} \mathbb{1}\big[\ell\big(f_n\big(X_{1:n}, f^\star(X_{1:n}), X_{t_{i_n+j}}\big), y_0\big) \geq \Delta_{01}\big]\, \Delta_{01}.
\end{aligned}$$

Furthermore, the subsequence $\{X_{t_{i_n+j}}\}_{j=1}^{\infty}$ 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 $\{X_t : t \notin \{t_{i_n+j} : j \in \mathbb{N}\}\}$ and f_n. This implies that

$$\big\{\mathbb{1}\big[\ell\big(f_n\big(X_{1:n}, f^\star(X_{1:n}), X_{t_{i_n+j}}\big), y_0\big) \geq \Delta_{01}\big]\big\}_{j=1}^{\infty}$$

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 $E \cap E_n''$,

$$\begin{aligned}
&\limsup_{s\to\infty} \frac{1}{k^* s} \sum_{j=1}^{s} \mathbb{1}\big[\ell\big(f_n\big(X_{1:n}, f^\star(X_{1:n}), X_{t_{i_n+j}}\big), y_0\big) \geq \Delta_{01}\big]\, \Delta_{01} \\
&= \frac{\Delta_{01}}{k^*} \pi_0\Big(\big\{x : \ell(f_n(X_{1:n}, f^\star(X_{1:n}), x), y_0) \geq \Delta_{01}\big\} \,\Big|\, R_{k^*}\Big) \\
&= \frac{2\Delta_{01}}{k^*} \pi_0(\{x \in R_{k^*} : \ell(f_n(X_{1:n}, f^\star(X_{1:n}), x), y_0) \geq \Delta_{01}\}).
\end{aligned}$$

Altogether, we have that on the event $E \cap E' \cap \bigcap_{n > n_{k^*-1}} E_n''$,

$$\begin{aligned}
&\limsup_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(f_n, f^\star; n) \\
&\geq \frac{2\Delta_{01}}{k^*} \limsup_{n\to\infty} \pi_0(\{x \in R_{k^*} : \ell(f_n(X_{1:n}, f^\star(X_{1:n}), x), y_0) \geq \Delta_{01}\}) \geq \frac{\Delta_{01}}{2k^*}.
\end{aligned}$$

Since $\frac{\Delta_{01}}{2k^*} > 0$, and since $E \cap E' \cap \bigcap_{n > n_{k^*-1}} E_n''$ has probability at least $2^{-k^*} > 0$ (by the union bound), 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 claim that $\mathbb{X} \in$ CRF. Note that for every $t > n_{k^*-1}$, if $t - n_{k^*-1} - 1$ is an integer multiple of k^*, then $t \in I_{k^*}$, in which case $X_t = V_{k^*,t}$, and otherwise $X_t = U_{k^*,t}$. Thus, since all X_t are independent, all $V_{k^*,t}$ are identically distributed, and all $U_{k^*,t}$ are identically distributed, we have that for any $n > n_{k^*-1}$, $\{X_t\}_{t=n}^{\infty}$ and $\{X_t\}_{t=n+k^*}^{\infty}$ have identical distributions. Thus,

the process $\{X_t\}_{t=n_{k^*-1}+1}^{\infty}$ is k^*-stationary (see Gray, 2009, Section 5.10), and hence also asymptotically mean stationary (recall the definition from Section 3). Since $\{X_t\}_{t=n_{k^*-1}+1}^{\infty}$ differs from $\mathbb{X}$ only by removing an initial finite segment, this immediately implies $\mathbb{X}$ is also asymptotically mean stationary. Thus, since (as discussed in Section 3 above) Theorem 8.1 of Gray (2009) implies that every asymptotically mean stationary process has convergent relative frequencies, and Theorem 18 of Section 3 establishes that CRF $\subseteq \mathcal{C}_1$, we have that $\mathbb{X} \in \mathcal{C}_1$, and since Theorem 7 establishes that SUIL $= \mathcal{C}_1$, this implies $\mathbb{X} \in$ 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 $k \in \mathbb{N}$, (44) and the definition of n_k imply that, defining $\mathbf{Q}_k = (\mathbf{U}_{<k}, \mathbf{V}_{<k}, \mathbf{W})$,

$$\mathbb{P}\Big(\pi_0\Big(\Big\{x : \ell\Big(f_{\sqrt{n_k}}\Big(\hat{X}^{(k)}_{1:\sqrt{n_k}}, f_k^\star\Big(\hat{X}^{(k)}_{1:\sqrt{n_k}}; \mathbf{Q}_k\Big), x\Big), y_0\Big) \geq \Delta_{01}\Big\}\Big) \geq 3/4\Big) < 2^{-k}.$$

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

$$\begin{aligned}
&\mathbb{E}\left[\sum_{k=1}^{\infty} \mathbb{P}\Big(\pi_0\Big(\Big\{x : \ell\Big(f_{\sqrt{n_k}}\Big(\hat{X}^{(k)}_{1:\sqrt{n_k}}, f_k^\star\Big(\hat{X}^{(k)}_{1:\sqrt{n_k}}; \mathbf{Q}_k\Big), x\Big), y_0\Big) \geq \Delta_{01}\Big\}\Big) \geq 3/4 \Big| \mathbf{V}\Big)\right] \\
&= \sum_{k=1}^{\infty} \mathbb{P}\Big(\pi_0\Big(\Big\{x : \ell\Big(f_{\sqrt{n_k}}\Big(\hat{X}^{(k)}_{1:\sqrt{n_k}}, f_k^\star\Big(\hat{X}^{(k)}_{1:\sqrt{n_k}}; \mathbf{Q}_k\Big), x\Big), y_0\Big) \geq \Delta_{01}\Big\}\Big) \geq 3/4\Big) < 1.
\end{aligned}$$

In particular, this implies that with probability one,

$$\sum_{k=1}^{\infty} \mathbb{P}\Big(\pi_0\Big(\Big\{x : \ell\Big(f_{\sqrt{n_k}}\Big(\hat{X}^{(k)}_{1:\sqrt{n_k}}, f_k^\star\Big(\hat{X}^{(k)}_{1:\sqrt{n_k}}; \mathbf{Q}_k\Big), x\Big), y_0\Big) \geq \Delta_{01}\Big\}\Big) \geq 3/4 \Big| \mathbf{V}\Big) < \infty.$$

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 $X_i = X_i^{(k)}(\mathbf{U}_{<k}, \mathbf{U}_k, \mathbf{v}_{<k}, \mathbf{v}_k, \mathbf{W})$ for every $k \in \mathbb{N}$ and $i \in \{n_{k-1}+1, \ldots, n_k\}$,

$$\sum_{k=1}^{\infty} \mathbb{P}\Big(\pi_0\Big(\Big\{x : \ell\Big(f_{\sqrt{n_k}}\Big(X_{1:\sqrt{n_k}}, f_k^\star\Big(X_{1:\sqrt{n_k}}; \mathbf{U}_{<k}, \mathbf{v}_{<k}, \mathbf{W}\Big), x\Big), y_0\Big) \geq \Delta_{01}\Big\}\Big) \geq 3/4\Big) < \infty.$$

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 $k > k_0$,

$$\pi_0\Big(\Big\{x : \ell\Big(f_{\sqrt{n_k}}\Big(X_{1:\sqrt{n_k}}, f_k^\star\Big(X_{1:\sqrt{n_k}}; \mathbf{U}_{<k}, \mathbf{v}_{<k}, \mathbf{W}\Big), x\Big), y_0\Big) \geq \Delta_{01}\Big\}\Big) < 3/4.$$

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 $f_k^\star$, $\forall j \in \mathbb{N}$, $f_k^\star(W_j; \mathbf{U}_{<k}, \mathbf{v}_{<k}, \mathbf{W}) = y_1$, and $\forall k', j \in \mathbb{N}$ with $k' < k$, $f_k^\star(U_{k',j}; \mathbf{U}_{<k}, \mathbf{v}_{<k}, \mathbf{W}) = y_1$. 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 $f_k^\star$, this implies $f_k^\star(U_{k,j}; \mathbf{U}_{<k}, \mathbf{v}_{<k}, \mathbf{W}) = y_1$ on the event H. The definition of $f_k^\star$ further implies that, on H, for every $k', k, j \in \mathbb{N}$ with $k' < k$, $f_k^\star(v_{k',j}; \mathbf{U}_{<k}, \mathbf{v}_{<k}, \mathbf{W}) = y_0$. 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 $f_k^\star(v_{k,j}; \mathbf{U}_{<k}, \mathbf{v}_{<k}, \mathbf{W}) = y_0$. Furthermore, by definition of $f_0^\star$, every $k, j \in \mathbb{N}$ has $f_0^\star(v_{k,j}; \mathbf{v}) = y_0$, and on the event H, every $j \in \mathbb{N}$ has $f_0^\star(W_j; \mathbf{v}) = y_1$, and $\forall k, j \in \mathbb{N}$, $f_0^\star(U_{k,j}; \mathbf{v}) = y_1$. Altogether we have that, on the event H, every $k', k, j \in \mathbb{N}$ with $k' \le k$ has $f_k^\star(v_{k',j}; \mathbf{U}_{<k}, \mathbf{v}_{<k}, \mathbf{W}) = f_0^\star(v_{k',j}; \mathbf{v})$, $f_k^\star(U_{k',j}; \mathbf{U}_{<k}, \mathbf{v}_{<k}, \mathbf{W}) = f_0^\star(U_{k',j}; \mathbf{v})$, and $f_k^\star(W_j; \mathbf{U}_{<k}, \mathbf{v}_{<k}, \mathbf{W}) = f_0^\star(W_j; \mathbf{v})$. In particular, note that for any $k \in \mathbb{N}$, every $i \in \{1, \ldots, \sqrt{n_k}\}$ has $X_i \in \{v_{k',i} : k' \le k\} \cup \{U_{k',i} : k' \le k\} \cup \{W_i\}$, so that, on the event H, $f_k^\star(X_i; \mathbf{U}_{<k}, \mathbf{v}_{<k}, \mathbf{W}) = f_0^\star(X_i; \mathbf{v})$. Thus, taking $f^\star(\cdot) = f_0^\star(\cdot; \mathbf{v})$, on the event $H \cap H'$, $\forall k \in \mathbb{N}$ with $k > k_0$,

$$\pi_0\Big(\Big\{x : \ell\Big(f_{\sqrt{n_k}}\Big(X_{1:\sqrt{n_k}}, f^\star(X_{1:\sqrt{n_k}}), x\Big), y_0\Big) \ge \Delta_{01}\Big\}\Big) < 3/4.$$

As mentioned above, the near-metric properties of ℓ imply that any $y \in \mathcal{Y}$ with $\ell(y, y_1) < \Delta_{01}$ necessarily has $\ell(y, y_0) > \Delta_{01}$. Therefore, on $H \cap H'$, $\forall k \in \mathbb{N}$ with $k > k_0$,

$$\begin{aligned}
&\pi_0\Big(\Big\{x : \ell\Big(f_{\sqrt{n_k}}\Big(X_{1:\sqrt{n_k}}, f^\star(X_{1:\sqrt{n_k}}), x\Big), y_1\Big) \ge \Delta_{01}\Big\}\Big) \\
&= 1 - \pi_0\Big(\Big\{x : \ell\Big(f_{\sqrt{n_k}}\Big(X_{1:\sqrt{n_k}}, f^\star(X_{1:\sqrt{n_k}}), x\Big), y_1\Big) < \Delta_{01}\Big\}\Big) \\
&\ge 1 - \pi_0\Big(\Big\{x : \ell\Big(f_{\sqrt{n_k}}\Big(X_{1:\sqrt{n_k}}, f^\star(X_{1:\sqrt{n_k}}), x\Big), y_0\Big) > \Delta_{01}\Big\}\Big) > 1/4. \qquad (45)
\end{aligned}$$

Now fix any $k, k' \in \mathbb{N}$ with $k' \ge k$ 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,

$$\begin{aligned}
&\frac{1}{n_{k'} - \sqrt{n_{k'}}} \sum_{t=\sqrt{n_{k'}}+1}^{n_{k'}} \ell\Big(f_{\sqrt{n_k}}\Big(X_{1:\sqrt{n_k}}, f^\star\Big(X_{1:\sqrt{n_k}}\Big), X_t\Big), f^\star(X_t)\Big) \\
&= \frac{1}{n_{k'} - \sqrt{n_{k'}}} \sum_{t=\sqrt{n_{k'}}+1}^{n_{k'}} \ell\Big(f_{\sqrt{n_k}}\Big(X_{1:\sqrt{n_k}}, f^\star\Big(X_{1:\sqrt{n_k}}\Big), X_t\Big), y_1\Big) \\
&\ge \frac{1}{n_{k'} - \sqrt{n_{k'}}} \sum_{t=\sqrt{n_{k'}}+1}^{n_{k'}} \mathbb{1}\Big[\ell\Big(f_{\sqrt{n_k}}\Big(X_{1:\sqrt{n_k}}, f^\star\Big(X_{1:\sqrt{n_k}}\Big), X_t\Big), y_1\Big) \ge \Delta_{01}\Big] \Delta_{01}.
\end{aligned}$$

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 $k \le k'$) 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 $1 - \frac{1}{(k')^3}$,

$$\frac{1}{n_{k'} - \sqrt{n_{k'}}} \sum_{t=\sqrt{n_{k'}}+1}^{n_{k'}} \mathbb{1}\Big[\ell\Big(f_{\sqrt{n_k}}\Big(X_{1:\sqrt{n_k}}, f^\star\Big(X_{1:\sqrt{n_k}}\Big), X_t\Big), y_1\Big) \geq \Delta_{01}\Big]$$
$$\geq \pi_0\Big(\Big\{x : \ell\Big(f_{\sqrt{n_k}}\Big(X_{1:\sqrt{n_k}}, f^\star\Big(X_{1:\sqrt{n_k}}\Big), x\Big), y_1\Big) \geq \Delta_{01}\Big\}\Big) - \sqrt{\frac{(3/2)\ln(k')}{n_{k'} - \sqrt{n_{k'}}}}.$$

Combining with (45) 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

$$\frac{1}{n_{k'} - \sqrt{n_{k'}}} \sum_{t=\sqrt{n_{k'}}+1}^{n_{k'}} \ell\Big(f_{\sqrt{n_k}}\Big(X_{1:\sqrt{n_k}}, f^\star\Big(X_{1:\sqrt{n_k}}\Big), X_t\Big), f^\star(X_t)\Big)$$
$$> \Delta_{01}\left(\frac{1}{4} - \sqrt{\frac{(3/2)\ln(k')}{n_{k'} - \sqrt{n_{k'}}}}\right).$$

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

$$\limsup_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(f_n, f^\star; n) \geq \limsup_{k\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}\Big(f_{\sqrt{n_k}}, f^\star; \sqrt{n_k}\Big)$$
$$= \limsup_{k\to\infty} \limsup_{m\to\infty} \frac{1}{m} \sum_{t=1}^{m} \ell\Big(f_{\sqrt{n_k}}\Big(X_{1:\sqrt{n_k}}, f^\star\Big(X_{1:\sqrt{n_k}}\Big), X_t\Big), f^\star(X_t)\Big)$$
$$\geq \limsup_{k\to\infty} \limsup_{k'\to\infty} \frac{1}{n_{k'}} \sum_{t=\sqrt{n_{k'}}+1}^{n_{k'}} \ell\Big(f_{\sqrt{n_k}}\Big(X_{1:\sqrt{n_k}}, f^\star\Big(X_{1:\sqrt{n_k}}\Big), X_t\Big), f^\star(X_t)\Big)$$
$$\geq \limsup_{k'\to\infty} \frac{n_{k'} - \sqrt{n_{k'}}}{n_{k'}} \Delta_{01}\left(\frac{1}{4} - \sqrt{\frac{(3/2)\ln(k')}{n_{k'} - \sqrt{n_{k'}}}}\right). \tag{46}$$

Since $n_{k'}$ is strictly increasing in k', we have that for any $k' \geq 4$, $0 \leq \frac{(3/2)\ln(k')}{n_{k'} - \sqrt{n_{k'}}} \leq \frac{3\ln(n_{k'})}{n_{k'}}$, 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 (46) equals $\Delta_{01}/4$. By the union bound, the event $H \cap H' \cap \bigcap_{k' \in \mathbb{N}\setminus\{1\}} \bigcap_{k \leq k'} H''_{k,k'}$ has probability at least

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

so that there is a nonzero probability that $\limsup_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(f_n, f^\star; n) \geq \Delta_{01}/4 > 0$. Thus, 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 $\mathbb{X} \in \text{CRF}$. For this, for each

$k \in \mathbb{N}$, let $J_k = \{n_{k-1} + (i-1)k + 1 : i \in \mathbb{N}, n_{k-1} + (i-1)k + 1 \leq \sqrt{n_k}\}$. For any $n \in \mathbb{N}$, define $k_n = \max\{k \in \mathbb{N} : n_{k-1} < n\}$; 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, $\forall n \in \mathbb{N}$,

$$\frac{1}{n}\sum_{t=1}^{n} \mathbb{1}_A(X_t) = \frac{1}{n}\sum_{k=1}^{k_n}\left(\left(\sum_{t=n_{k-1}+1}^{\min\{\sqrt{n_k},n\}} \left(\mathbb{1}_{J_k}(t)\mathbb{1}_A(v_{k,t}) + \mathbb{1}_{\mathbb{N}\setminus J_k}(t)\mathbb{1}_A(U_{k,t})\right)\right) + \sum_{t=\sqrt{n_k}+1}^{\min\{n_k,n\}} \mathbb{1}_A(W_t)\right). \tag{47}$$

By Kolmogorov's strong law of large numbers (Ash and Doléans-Dade, 2000, Theorem 6.2.2), with probability one we have

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{k_n}\left(\left(\sum_{t=n_{k-1}+1}^{\min\{\sqrt{n_k},n\}} \left(\mathbb{1}_{J_k}(t)\left(\mathbb{1}_A(v_{k,t}) - \mathbb{1}_A(v_{k,t})\right) + \mathbb{1}_{\mathbb{N}\setminus J_k}(t)\left(\mathbb{1}_A(U_{k,t}) - \pi_0(A|\mathcal{X}\setminus R_k)\right)\right)\right) + \sum_{t=\sqrt{n_k}+1}^{\min\{n_k,n\}} \left(\mathbb{1}_A(W_t) - \pi_0(A)\right)\right) = 0. \tag{48}$$

We therefore focus on establishing convergence of

$$\frac{1}{n}\sum_{k=1}^{k_n}\left(\left(\sum_{t=n_{k-1}+1}^{\min\{\sqrt{n_k},n\}} \left(\mathbb{1}_{J_k}(t)\mathbb{1}_A(v_{k,t}) + \mathbb{1}_{\mathbb{N}\setminus J_k}(t)\pi_0(A|\mathcal{X}\setminus R_k)\right)\right) + \sum_{t=\sqrt{n_k}+1}^{\min\{n_k,n\}} \pi_0(A)\right). \tag{49}$$

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,

$$\begin{aligned} 0 \leq \frac{1}{n}\sum_{k=1}^{k_n}\sum_{t=n_{k-1}+1}^{\min\{\sqrt{n_k},n\}} \mathbb{1}_{J_k}(t)\mathbb{1}_A(v_{k,t}) &\leq \frac{1}{n}\sum_{k=1}^{k_n}\sum_{t=n_{k-1}+1}^{\min\{\sqrt{n_k},n\}} \mathbb{1}_{J_k}(t) \\ &\leq \frac{\sqrt{n_{k_n-1}}}{n} + \frac{1}{n}\left(\frac{n}{k_n}+1\right) = \frac{\sqrt{n_{k_n-1}}}{n} + \frac{1}{k_n} + \frac{1}{n}. \end{aligned} \tag{50}$$

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,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{k_n}\sum_{t=n_{k-1}+1}^{\min\{\sqrt{n_k},n\}} \mathbb{1}_{J_k}(t)\mathbb{1}_A(v_{k,t}) = 0. \tag{51}$$

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 (50) has limit 0, which in particular also implies that $\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{k_n}\sum_{t=n_{k-1}+1}^{\min\{\sqrt{n_k},n\}} \mathbb{1}_{J_k}(t) = 0$. Furthermore, for any fixed $\varepsilon \in (0,1)$, $\lim_{n\to\infty}\frac{n_{k_\varepsilon-1}}{n} = 0$. Thus, since $k_n \to \infty$ implies $k_n \geq k_\varepsilon$ for all sufficiently large n, we have

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

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 (51), (48), and (47), we have

$$\frac{1}{n}\sum_{t=1}^{n} \mathbb{1}_A(X_t) \to \pi_0(A) \text{ (a.s.).} \tag{52}$$

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 $\mathbb{X} \in \text{CRF}$. Since (as argued above) it holds that $\text{CRF} \subseteq \mathcal{C}_1 = \text{SUIL}$, this further implies $\mathbb{X} \in \text{SUIL}$. 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 31 *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 $\hat{f}_n(x_{1:n}, y_{1:n}, x) = y_{i(x;x_{1:n})}$; define $\hat{f}_n(x_{1:n}, y_{1:n}, x) = y_0$ for some arbitrary fixed $y_0 \in \mathcal{Y}$ if $x \notin \{x_1, \ldots, x_n\}$. 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 $k \leq |\mathcal{X}|$, 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 Corollary 15, we have

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

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 $f^\star : \mathcal{X} \to \mathcal{Y}$,

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

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) $\text{SUIL} = \mathcal{C}_1$, 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 satisfies 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 that are ergodic (since they are product processes) with convergent relative frequencies (CRF), as argued in the proof. Formally, we have established the following corollary.

Corollary 32 *If $\mathcal{X}$ is uncountable, then there does not exist an inductive learning rule that is strongly universally consistent under every ergodic $\mathbb{X} \in$ 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, and at the same time positively resolve Open Problem 1 for countable $\mathcal{X}$. However, for the *general* case with uncountable $\mathcal{X}$, we leave open both the question in Open Problem 1 and the question of whether Condition 2 is sufficient for $\mathbb{X}$ to admit strong universal online learning (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. We also construct an online learning rule that is universally consistent under every $\mathbb{X} \in$ SUAL. 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 separable but *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 33 *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 Define $W_n = \sum_{i=1}^{\infty} w_{n,i}$ 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 $n \in \mathbb{N}$,

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

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

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

Lemma 34 *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 measurable 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. $\lim_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(\hat{h}_\cdot^{(i_n)}, f^\star; n) = 0$, then $\lim_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(\hat{f}_\cdot, f^\star; n) = 0$ (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 $\sum_{i=1}^{\infty} p_i = 1$. Fix any $b \in (0,1)$, and for $n, i \in \mathbb{N}$ define $v_{n,i}$ as in Lemma 33, 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 $\bar{z}_n(x_{1:n}, y_{1:n}) = \sum_{i=1}^{\infty} v_{n,i} \hat{z}_{n,i}(x_{1:n}, y_{1:n})$. 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 for each n) 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 33 (see e.g., Littlestone and Warmuth, 1994), together with concentration inequalities to argue that the bound from Lemma 33 almost surely becomes valid in the limit of $n \to \infty$. However, instead of this approach, we will analyze a method that avoids randomization.[6] 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 $\hat{f}_n(x_{1:n}, y_{1:n}, x_{n+1}) = \hat{y}_{n+1}$ for some value $\hat{y}_{n+1} \in \mathcal{Y}$ satisfying[7]

$$\sum_{i=1}^{\infty} v_{(n+1),i} \ell\left(\hat{y}_{n+1}, \hat{h}_n^{(i)}(x_{1:n}, y_{1:n}, x_{n+1})\right) \leq \varepsilon_n + \inf_{y\in\mathcal{Y}} \sum_{i=1}^{\infty} v_{(n+1),i} \ell\left(y, \hat{h}_n^{(i)}(x_{1:n}, y_{1:n}, x_{n+1})\right).$$

6. In general, randomization is known to be necessary for achieving optimal regret guarantees in online learning (see Cesa-Bianchi and Lugosi, 2006, Chapter 4). However, since the reference sequence $\hat{h}_\cdot^{(i_n)}$ itself has $\hat{\mathcal{L}}_{\mathbb{X}}(\hat{h}_\cdot^{(i_n)}, f^\star; n) \to 0$ (a.s.), we are not concerned with multiplicative constant factors for the purpose of achieving $\hat{\mathcal{L}}_{\mathbb{X}}(\hat{f}_\cdot, f^\star; n) \to 0$ (a.s.), and thus we can avoid randomization.

7. Here we suppose the choice of $\hat{y}_{n+1}$ is such that the function $\hat{f}_n(\cdot,\cdot,\cdot)$ is measurable: for instance, it would suffice to consider an enumeration of a countable dense subset of $\mathcal{Y}$ (which exists by the separability assumption on $\mathcal{Y}$) and then choose the first y in this enumeration satisfying the ε_n-excess criterion in the definition of $\hat{y}_{n+1}$.

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 relaxed triangle inequality and the fact that $\sum_{i=1}^{\infty} v_{(t+1),i} = 1$, we have that

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

Then the definition of $\hat{f}_t$ guarantees the right hand side is at most

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

so that

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

Together with Lemma 33, we have that

$$\begin{aligned}
&\frac{1}{n} \sum_{t=0}^{n-1} \ell\Big(\hat{f}_t(x_{1:t}, y_{1:t}, x_{t+1}), y_{t+1}\Big) \qquad (53) \\
&\leq \left(\frac{1}{n} \sum_{t=0}^{n-1} \varepsilon_t\right) + 2c_\ell \inf_{i \in \mathbb{N}} \left(\frac{\ln(1/b)}{1-b} \left(\frac{1}{n} \sum_{t=0}^{n-1} \ell\Big(\hat{h}_t^{(i)}(x_{1:t}, y_{1:t}, x_{t+1}), y_{t+1}\Big)\right) + \frac{\bar{\ell}}{(1-b)n} \ln\left(\frac{1}{p_i}\right)\right).
\end{aligned}$$

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 (53) implies

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

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 35 *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* $i_n \le j_n$ *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 $\sup_{n\ge n_i}\beta_{i,n} \le \frac{1}{i} + \limsup_{n\to\infty}\beta_{i,n}$; such an n_i is guaranteed to exist by the definition of the limsup. For each $n\in\mathbb{N}$ with $n < n_1$, define $i_n = 1$, and for each $n\in\mathbb{N}$ with $n\ge n_1$, define $i_n = \max\{i\in\{1,\ldots,j_n\} : n\ge n_i\}$. By definition, we have $i_n\le j_n$ for every $n\in\mathbb{N}$. Furthermore, by definition, we have $n\ge n_{i_n}$ for every $n\ge n_1$, so that $\beta_{i_n,n}\le\frac{1}{i_n}+\limsup_{n'\to\infty}\beta_{i_n,n'}$. 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

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

Since $\liminf_{n\to\infty}\beta_{i_n,n}\ge 0$ 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* for $\mathbb{X}$ to admit strong universal online learning, 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 SUOL $= \mathcal{C}_2$. Formally, we make partial progress toward resolving the following question, which remains open at this writing.

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

In this subsection, we show that in general, 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. In particular, we may note an interesting parallel to Corollary 15.

Lemma 36 *A process $\mathbb{X}$ satisfies Condition 2 if and only if every disjoint sequence $\{A_i\}_{i=1}^{\infty}$ in $\mathcal{B}$ with $\bigcup_{i=1}^{\infty} A_i = \mathcal{X}$ (i.e., every countable measurable partition) satisfies*

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\Big[X_t \in \bigcup \big\{A_i : X_{1:(t-1)} \cap A_i = \emptyset\big\}\Big] = 0 \text{ (a.s.).}$$

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 $k \geq 2$, we have that $\{B_k\}_{k=1}^{\infty}$ is a disjoint sequence in $\mathcal{B}$ with $\bigcup_{k=1}^{\infty} B_k = \mathcal{X}$ and $|\{k : X_{1:T} \cap A_k \neq \emptyset\}| \leq |\{k : X_{1:T} \cap B_k \neq \emptyset\}| \leq |\{k : X_{1:T} \cap A_k \neq \emptyset\}| + 1$, so that $|\{k : X_{1:T} \cap B_k \neq \emptyset\}| = o(T)$ (a.s.) if and only if $|\{k : X_{1:T} \cap A_k \neq \emptyset\}| = o(T)$ (a.s.). 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 disjoint sequence $\{A_i\}_{i=1}^{\infty}$ in $\mathcal{B}$ with $\bigcup_{i=1}^{\infty} A_i = \mathcal{X}$. Then note that, for any $T \in \mathbb{N}$, $|\{k \in \mathbb{N} : X_{1:T} \cap A_k \neq \emptyset\}| = \mathbb{1}\big[X_T \in \bigcup \big\{A_i : X_{1:(T-1)} \cap A_i = \emptyset\big\}\big] + \big|\big\{k \in \mathbb{N} : X_{1:(T-1)} \cap A_k \neq \emptyset\big\}\big|$. By induction (taking $T = 1$ as a trivially-satisfied base case), this implies that $\forall T \in \mathbb{N}$,

$$|\{k \in \mathbb{N} : X_{1:T} \cap A_k \neq \emptyset\}| = \sum_{t=1}^{T} \mathbb{1}\Big[X_t \in \bigcup \big\{A_i : X_{1:(t-1)} \cap A_i = \emptyset\big\}\Big].$$

In particular, this implies that $\sum_{t=1}^{T} \mathbb{1}\big[X_t \in \bigcup \big\{A_i : X_{1:(t-1)} \cap A_i = \emptyset\big\}\big] = o(T)$ (a.s.) if and only if $|\{k \in \mathbb{N} : X_{1:T} \cap A_k \neq \emptyset\}| = o(T)$ (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 37 $\mathrm{SUOL} \subseteq \mathcal{C}_2$.

Proof This proof follows essentially the same outline as that of Lemma 20. We prove the result in the contrapositive. Suppose $\mathbb{X} \notin \mathcal{C}_2$. By Lemma 36, there exists a disjoint sequence $\{A_i\}_{i=1}^{\infty}$ in $\mathcal{B}$ with $\bigcup_{i=1}^{\infty} A_i = \mathcal{X}$ such that, with probability strictly greater than 0,

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\Big[X_t \in \bigcup \big\{A_i : X_{1:(t-1)} \cap A_i = \emptyset\big\}\Big] > 0.$$

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

$$\mathbb{E}\left[\limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\Big[X_t \in \bigcup \big\{A_i : X_{1:(t-1)} \cap A_i = \emptyset\big\}\Big]\right] > 0. \tag{54}$$

Now take any two distinct values $y_0, y_1 \in \mathcal{Y}$, and (as we did in the proof of Lemma 20) for each $\kappa \in [0,1)$, $i \in \mathbb{N}$, and $x \in A_i$, letting $\kappa_i = \lfloor 2^i \kappa \rfloor - 2\lfloor 2^{i-1}\kappa \rfloor \in \{0,1\}$, define $f^\star_\kappa(x) = y_{\kappa_i}$. Recall that we established in the proof of Lemma 20 that $(x,\kappa) \mapsto f^\star_\kappa(x)$ is measurable in the appropriate product σ-algebra. Also for every $t \in \mathbb{N}$ define i_t as the unique $i \in \mathbb{N}$ with $X_t \in A_i$, 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 $f^\kappa_n(\cdot) = g_n(X_{1:n}, f^\star_\kappa(X_{1:n}), \cdot)$ for each $n \in \mathbb{N}$. Then

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

By Fubini's theorem, this is equal

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

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

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

and linearity of integration implies this equals

$$\mathbb{E}\left[\limsup_{n\to\infty} \frac{1}{n}\sum_{t=0}^{n-1} \mathbb{1}_{\bar{\mathcal{A}}(X_{1:t})}(X_{t+1}) \int_0^1 \ell(f^\kappa_t(X_{t+1}), f^\star_\kappa(X_{t+1}))\, \mathrm{d}\kappa\right]. \tag{55}$$

For any $t \in \mathbb{N}\cup\{0\}$, the value of $f^\kappa_t(X_{t+1})$ 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 $f^\kappa_t(X_{t+1})$ 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

$$\int_0^1 \ell(f^\kappa_t(X_{t+1}), f^\star_\kappa(X_{t+1}))\, \mathrm{d}\kappa = \mathbb{E}\Big[\ell\big(f^K_t(X_{t+1}), f^\star_K(X_{t+1})\big)\,\Big|\,\mathbb{X}, g_t\Big]$$
$$= \mathbb{E}\Big[\mathbb{E}\Big[\ell\Big(g_t(X_{1:t}, \{y_{K_{i_j}}\}_{j=1}^t, X_{t+1}), y_{K_{t+1}}\Big)\,\Big|\,\mathbb{X}, g_t, K_{i_1}, \ldots, K_{i_t}\Big]\,\Big|\,\mathbb{X}, g_t\Big]$$
$$= \mathbb{E}\left[\sum_{b\in\{0,1\}} \frac{1}{2}\ell\Big(g_t(X_{1:t}, \{y_{K_{i_j}}\}_{j=1}^t, X_{t+1}), y_b\Big)\,\Big|\,\mathbb{X}, g_t\right].$$

By the relaxed triangle inequality, this is no smaller than $\mathbb{E}\Big[\frac{1}{2c_\ell}\ell(y_0,y_1)\Big|\mathbb{X}, g_t\Big] = \frac{1}{2c_\ell}\ell(y_0,y_1)$, so that (55) is at least as large as

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

where this last inequality is immediate from (54) and the fact that (since ℓ is a near-metric) $\ell(y_0, y_1) > 0$. Altogether, we have that

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

In particular, this implies $\exists \kappa \in [0,1)$ such that $\mathbb{E}\left[\limsup_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(g_\cdot, f^\star_\kappa; n)\right] > 0$. Since any random variable equal 0 (a.s.) necessarily has expected value 0, this further implies that with probability strictly greater than 0, $\limsup_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(g_\cdot, f^\star_\kappa; n) > 0$. 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, $\mathbb{X} \notin \text{SUOL}$. 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 $\mathcal{X}$. 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.

Lemma 38 *If $\mathcal{X}$ is countable, then a process $\mathbb{X}$ satisfies Condition 2 if and only if*

$$|\{x \in \mathcal{X} : X_{1:T} \cap \{x\} \neq \emptyset\}| = o(T) \text{ (a.s.).}$$

Proof Enumerate the elements of $\mathcal{X}$ as $z_1, z_2, \ldots$ (or $z_1, \ldots, z_{|\mathcal{X}|}$ in the case of finite $|\mathcal{X}|$). If Condition 2 is satisfied, then choose $A_k = \{z_k\}$ for every $k \in \mathbb{N}$ with $k \leq |\mathcal{X}|$, and if $|\mathcal{X}| < \infty$ then let $A_k = \emptyset$ for every $k > |\mathcal{X}|$. It immediately follows from Condition 2 that $|\{x \in \mathcal{X} : X_{1:T} \cap \{x\} \neq \emptyset\}| = o(T)$ (a.s.). For the other direction, if Condition 2 fails, there exists a disjoint sequence $\{A_k\}_{k=1}^\infty$ in $\mathcal{B}$ that has $|\{k \in \mathbb{N} : X_{1:T} \cap A_k \neq \emptyset\}| \neq o(T)$ with nonzero probability. Noting that $|\{k \in \mathbb{N} : X_{1:T} \cap A_k \neq \emptyset\}| \leq |\{x \in \mathcal{X} : X_{1:T} \cap \{x\}\}|$, we have that, on this same event of nonzero probability, $|\{x \in \mathcal{X} : X_{1:T} \cap \{x\}\}| \neq o(T)$ as well. ■

We now state our result for strong universal online learning when $\mathcal{X}$ is countable.

Theorem 39 *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,* $\text{SUOL} = \mathcal{C}_2$. *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 $\text{SUOL} \subseteq \mathcal{C}_2$ from Theorem 37, it suffices to show $\mathcal{C}_2 \subseteq \text{SUOL}$, 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 $i \leq t$ 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

$$\limsup_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(f_\cdot, f^\star; n) = \limsup_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \ell(f_t(X_{1:t}, f^\star(X_{1:t}), X_{t+1}), f^\star(X_{t+1}))$$

$$\leq \limsup_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \bar{\ell} \mathbb{1}[\nexists i \in \{1, \ldots, t\} : X_{t+1} = X_i] = \bar{\ell} \limsup_{n\to\infty} \frac{1}{n} \left| \{x \in \mathcal{X} : X_{1:n} \cap \{x\} \neq \emptyset\} \right|.$$

Lemma 38 implies that the rightmost expression above is equal 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. ■

For uncountable $\mathcal{X}$, we can at least state a corollary holding for all *deterministic* processes, via a reduction to the case of countable $\mathcal{X}$.

Corollary 40 *For any deterministic process $\mathbb{X}$, Condition 2 is necessary and sufficient for $\mathbb{X}$ to admit strong universal online learning: that is, $\mathbb{X} \in$ SUOL if and only if $\mathbb{X} \in \mathcal{C}_2$.*

Proof Sketch This result follows from essentially the same proof used for Theorem 39; the only significant change is in the proof of Lemma 38, which can be established for deterministic processes on general $\mathcal{X}$ by replacing the z_k sequence defined in the proof by the distinct entries of the sequence $\mathbb{X}$ (noting that the intersection of $\mathbb{X}$ with the complement of this z_k sequence is empty). 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 variables. Furthermore, it is straightforward to show that any deterministic $\mathbb{X}$ 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 39 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 41 SUAL $\subseteq$ SUOL*, and the inclusion is* strict *if and only if* $|\mathcal{X}| = \infty$.

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 (34), as established by Theorem 29. 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 $\hat{h}_n^{(i)}(x_{1:n}, y_{1:n}, x_{n+1}) = \hat{g}_{i,n}(x_{1:n}, y_{1:i}, x_{n+1})$. Measurability of $\hat{h}_n^{(i)}$ 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 34 (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 $i, n \in \mathbb{N}$, define $\hat{\beta}_{i,n} = \hat{\mathcal{L}}_{\mathbb{X}}(\hat{h}_\cdot^{(i)}, f^\star; n)$. 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 an event E of probability one. In particular, on E, Lemma 35 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 34 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 40, 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 $A_k = \{z_i : i \geq k\}$ for each $k \in \mathbb{N}$, and note that each A_k is countable, hence in $\mathcal{B}$, and that $A_k \downarrow \emptyset$. 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 $k \in \mathbb{N}$, so that $\lim_{k\to\infty} \hat{\mu}_{\mathbb{X}}(A_k) \geq \frac{1}{2} > 0$. Since $\mathbb{X}$ is deterministic, this violates the requirement of Condition 1, and therefore $\mathbb{X} \notin \mathcal{C}_1$. ■

The proof of Theorem 41 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 42 *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 41) 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 37 establishes that SUOL $\subseteq \mathcal{C}_2$, and Theorem 7 establishes that SUAL $= \mathcal{C}_1$, Theorem 41 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 40 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 $\mathcal{C}_1 = \mathcal{C}_2$. Altogether, we conclude that the above proof also establishes the following result.

Corollary 43 $\mathcal{C}_1 \subseteq \mathcal{C}_2$*, and the inclusion is* strict *if and only if* $|\mathcal{X}| = \infty$.

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 any L_p loss $(y, y') \mapsto |y - y'|^p$ $(p \in (0, \infty))$ 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 44 *The set* SUOL *is invariant to the specification of* $(\mathcal{Y}, \ell)$*, subject to* $(\mathcal{Y}, \ell)$ *being totally bounded with* $\bar{\ell} > 0$*.*

Proof To disambiguate notation in this proof, for any near-metric space $(\mathcal{Y}', \ell')$, we denote by $\text{SUOL}_{(\mathcal{Y}', \ell')}$ the family SUOL as it would be defined if $(\mathcal{Y}, \ell)$ were specified as $(\mathcal{Y}', \ell')$. As above, 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 $\text{SUOL}_{(\mathcal{Y}', \ell')} = \text{SUOL}_{(\{0,1\}, \ell_{01})}$ for all totally bounded near-metric spaces $(\mathcal{Y}', \ell')$ with $\sup_{y, y' \in \mathcal{Y}'} \ell'(y, y') > 0$. Fix any such $(\mathcal{Y}', \ell')$.

The inclusion $\text{SUOL}_{(\mathcal{Y}', \ell')} \subseteq \text{SUOL}_{(\{0,1\}, \ell_{01})}$ is quite straightforward, as follows. For any $\mathbb{X} \in \text{SUOL}_{(\mathcal{Y}', \ell')}$, 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 $\hat{f}_n^{01}(x_{1:n}, y_{1:n}, x_{n+1}) = \operatorname{argmin}_{y \in \{0,1\}} \ell'\left(\hat{f}_n(x_{1:n}, y'_{1:n}, x_{n+1}), z_y\right)$ (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 $\forall t \in \mathbb{N} \cup \{0\}$,

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

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

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

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 $\hat{f}_n^{01}$ is strongly universally consistent under $\mathbb{X}$ (for the specification $(\mathcal{Y}, \ell) = (\{0,1\}, \ell_{01})$), so that $\mathbb{X} \in \text{SUOL}_{(\{0,1\},\ell_{01})}$. Since this argument holds for any $\mathbb{X} \in \text{SUOL}_{(\mathcal{Y}',\ell')}$, we conclude that $\text{SUOL}_{(\mathcal{Y}',\ell')} \subseteq \text{SUOL}_{(\{0,1\},\ell_{01})}$.

The proof of the converse inclusion is somewhat more involved. Specifically, fix any $\mathbb{X} \in \text{SUOL}_{(\{0,1\},\ell_{01})}$, and let $\hat{f}_n^{01}$ 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 $(\mathcal{Y}, \ell) = (\mathcal{Y}', \ell')$ totally bounded, as follows. For each $\varepsilon > 0$, let $\mathcal{Y}_\varepsilon' \subseteq \mathcal{Y}'$ be such that $|\mathcal{Y}_\varepsilon'| < \infty$ and $\sup_{y\in\mathcal{Y}'} \inf_{y_\varepsilon\in\mathcal{Y}_\varepsilon'} \ell'(y_\varepsilon, y) \leq \varepsilon$, as guaranteed to exist by total boundedness. For each $y \in \mathcal{Y}'$, let $g_\varepsilon(y) = \operatorname{argmin}_{y_\varepsilon\in\mathcal{Y}_\varepsilon'} \ell'(y_\varepsilon, y)$, 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 $h_\varepsilon^{(y_\varepsilon)}(y) = \mathbb{1}[g_\varepsilon(y) = y_\varepsilon]$. One can easily verify that g_ε and $h_\varepsilon^{(y_\varepsilon)}$ are measurable functions, and furthermore that for every $y \in \mathcal{Y}'$, exactly one $y_\varepsilon \in \mathcal{Y}_\varepsilon'$ has $h_\varepsilon^{(y_\varepsilon)}(y) = 1$ while every $y_\varepsilon' \in \mathcal{Y}_\varepsilon' \setminus \{y_\varepsilon\}$ has $h_\varepsilon^{(y_\varepsilon')}(y) = 0$.

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

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

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 $\hat{f}_n^{01}$, the $h_\varepsilon^{(y_\varepsilon)}$ functions, and the argmax. Thus, $\hat{f}_n^{(\varepsilon)}$ 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 $x \mapsto h_\varepsilon^{(y_\varepsilon)}(f(x))$ is a measurable function $\mathcal{X} \to \{0,1\}$, and therefore (by strong universal consistency of $\hat{f}_n^{01}$) with probability one,

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

By the union bound, this holds simultaneously for all $y_\varepsilon \in \mathcal{Y}_\varepsilon'$ with probability one. Furthermore, note that if $\hat{f}_t^{01}(X_{1:t}, h_\varepsilon^{(y_\varepsilon)}(f(X_{1:t})), X_{t+1}) = h_\varepsilon^{(y_\varepsilon)}(f(X_{t+1}))$ for every $y_\varepsilon \in \mathcal{Y}_\varepsilon'$, then

$\hat{f}_t^{(\varepsilon)}(X_{1:t}, f(X_{1:t}), X_{t+1}) = g_\varepsilon(f(X_{t+1}))$. We therefore have that, under the specification $(\mathcal{Y}, \ell) = (\mathcal{Y}', \ell')$,

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

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 $i, n \in \mathbb{N}$, define $\beta_{i,n}^{f^\star} = \hat{\mathcal{L}}_{\mathbb{X}}\Big(\hat{f}_\cdot^{(1/i)}, f^\star; n\Big)$ (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

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

Thus, since these $\beta_{i,n}^{f^\star}$ are also nonnegative, Lemma 35 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 $\lim_{n\to\infty} \beta_{i_n,n}^{f^\star} = 0$. Therefore, applying Lemma 34 to the sequence $\left\{\hat{f}_n^{(1/i)}\right\}_{i=1}^\infty$ 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 45 *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 44, 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

near-metric spaces $(\mathcal{Y}', \ell')$ with $\sup_{y,y' \in \mathcal{Y}'} \ell'(y, y') < \infty$ 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 $\sup_{y \in \mathcal{Y}} \inf_{y_\varepsilon \in \mathcal{Y}'_\varepsilon} \ell'(y_\varepsilon, y) \leq \varepsilon$ 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 $\min_{i \leq k} \ell'(z_i, z_{k+1}) > \varepsilon$. 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 $\inf_{i,j \in \mathbb{N}: i \neq j} \ell'(z_i, z_j) \geq \varepsilon > 0$. 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 $\ell'(\hat{g}_n(x_{1:n}, y'_{1:n}, x_{n+1}), z_y) < \varepsilon/(2c_\ell)$, 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 $(\mathcal{Y}, \ell) = (\mathbb{N}, \ell_{01})$)

$$\begin{aligned}
&\limsup_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}\Big(\hat{g}_\cdot^{\mathbb{N}}, f; n\Big) = \limsup_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \mathbb{1}\Big[\hat{g}_t^{\mathbb{N}}(X_{1:t}, f(X_{1:t}), X_{t+1}) \neq f(X_{t+1})\Big] \\
&\leq \limsup_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \mathbb{1}\big[\ell'\big(\hat{g}_t(X_{1:t}, f'(X_{1:t}), X_{t+1}), f'(X_{t+1})\big) \geq \varepsilon/(2c_\ell)\big] \\
&\leq \frac{2c_\ell}{\varepsilon} \limsup_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \ell'\big(\hat{g}_t(X_{1:t}, f'(X_{1:t}), X_{t+1}), f'(X_{t+1})\big) = 0 \text{ (a.s.)},
\end{aligned}$$

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 $\sup_{y \in \mathcal{Y}'} \inf_{\tilde{y} \in \tilde{\mathcal{Y}}'} \ell'(\tilde{y}, y) = 0$; 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 $\hat{f}^{\mathbb{N}}_n$)

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

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

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

The rest of this proof follows identically to the analogous part of the proof of Theorem 44. Briefly, for any measurable $f^\star : \mathcal{X} \to \mathcal{Y}'$, for each $i, n \in \mathbb{N}$, defining $\beta^{f^\star}_{i,n} = \hat{\mathcal{L}}_{\mathbb{X}}\Big(\hat{f}^{(1/i)}_{\cdot}, f^\star; n\Big)$ (under the specification $(\mathcal{Y}, \ell) = (\mathcal{Y}', \ell')$), by the union bound, on an event of probability one, we have

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

Therefore Lemma 35 (with $j_n = n$) and Lemma 34 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 $(\mathcal{Y}, \ell) = (\mathcal{Y}', \ell')$, $\lim_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}\Big(\hat{f}_{\cdot}, f^\star; n\Big) = 0$ 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 near-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 online learning rules. Specifically, the proofs of Theorems 44 and 45 can also be used to establish the following corollary.

Corollary 46 *For any separable near-metric space $(\mathcal{Y}', \ell')$ with $0 < \sup_{y,y'\in\mathcal{Y}'} \ell'(y, y') < \infty$, the following hold.*

- *If $(\mathcal{Y}', \ell')$ is totally bound, there exists an optimistically universal online learning rule when $(\mathcal{Y}, \ell) = (\mathcal{Y}', \ell')$ if and only if there exists an optimistically universal online learning rule when $(\mathcal{Y}, \ell) = (\{0, 1\}, \ell_{01})$.*
- *If $(\mathcal{Y}', \ell')$ is not totally bound, there exists an optimistically universal online learning rule when $(\mathcal{Y}, \ell) = (\mathcal{Y}', \ell')$ if and only if there exists an optimistically universal online learning rule when $(\mathcal{Y}, \ell) = (\mathbb{N}, \ell_{01})$.*

The question of whether the two SUOL sets from the above Theorems 44 and 45 are equivalent remains an interesting open problem.

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

In particular, in the notation of the above proofs, Theorems 44 and 45 imply this problem is equivalent to the question of whether $\text{SUOL}_{(\{0,1\},\ell_{01})} = \text{SUOL}_{(\mathbb{N},\ell_{01})}$: 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)[8] 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.[9]

Theorem 47 *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 $t \leq n_{k-1}$. Let $X_t^{(k)} = X_t$ for every $t \in \mathbb{N}$ with $t \leq n_{k-1}$. 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

$$\mathbb{P}\Big(\hat{t}_n(X_{1:n}^{(k)}) = \mathbb{1}[(k+1)/2 \in \mathbb{N}]\Big) > 1/2, \tag{56}$$

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).

8. In the case of random $\hat{t}_n$, we will suppose $\hat{t}_n$ is independent from $\mathbb{X}$.
9. 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 versus 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.

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 (56) with $k = k^*$. Furthermore, by the definition of $X_t^{(k^*)}$ for every $t \leq n_{k^*-1}$, 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 $n > n_{k^*-1}$,

$$\mathbb{P}\big(\hat{t}_n(X_{1:n}) = \mathbb{1}[(k^*+1)/2 \in \mathbb{N}]\big) \leq 1/2. \tag{57}$$

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 41 implies SUAL $\subset$ SUOL, we have $\mathbb{X} \in$ SUOL as well. However, (57) 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 37 implies SUOL $\subseteq \mathcal{C}_2$, and Theorems 7 and 41 imply SUIL = SUAL $\subset$ SUOL, we also have that $\mathbb{X} \notin$ SUOL, $\mathbb{X} \notin$ SUAL, and $\mathbb{X} \notin$ SUIL. However, (57) 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 (56) 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 $\mathbb{X} \in \text{SUIL} \cap \text{SUAL} \cap \text{SUOL}$ by Theorems 7 and 41. 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 48 *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 $\text{SUIL} = \mathcal{C}_1$, 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 near-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 $\sup_{y_1, y_2 \in \mathcal{Y}} \ell(y_1, y_2) = \infty$. All of the other restrictions from Section 1.1 (e.g., that $(\mathcal{Y}, \ell)$ is a separable near-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, 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. We can also state an equivalent form of Condition 3 in terms of countable measurable partitions of $\mathcal{X}$, as follows.

Lemma 49 *A process $\mathbb{X}$ satisfies Condition 3 if and only if every disjoint sequence $\{A_i\}_{i=1}^{\infty}$ in $\mathcal{B}$ with $\bigcup_{i=1}^{\infty} A_i = \mathcal{X}$ (i.e., every countable measurable partition) satisfies*

$$|\{k \in \mathbb{N} : \mathbb{X} \cap A_k \neq \emptyset\}| < \infty \text{ (a.s.).} \tag{58}$$

Proof First suppose $\mathbb{X}$ satisfies Condition 3. Given any disjoint sequence $\{A_k\}_{k=1}^{\infty}$ in $\mathcal{B}$ with $\bigcup_{k=1}^{\infty} A_k = \mathcal{X}$, we can define a sequence $B_k = \bigcup_{i=k}^{\infty} A_i$ in $\mathcal{B}$ with $B_k \downarrow \emptyset$. Then note that $|\{k \in \mathbb{N} : \mathbb{X} \cap A_k \neq \emptyset\}| \leq \sup\{k \in \mathbb{N} : \mathbb{X} \cap A_k \neq \emptyset\} = |\{k \in \mathbb{N} : \mathbb{X} \cap B_k \neq \emptyset\}|$, and Condition 3 implies the rightmost expression is finite almost surely. Thus, (58) holds for all such sequences $\{A_k\}_{k=1}^{\infty}$.

For the converse direction, suppose $\mathbb{X}$ satisfies (58) for every disjoint sequence $\{A_i\}_{i=1}^{\infty}$ in $\mathcal{B}$ with $\bigcup_{i=1}^{\infty} A_i = \mathcal{X}$. Let $\{B_k\}_{k=1}^{\infty}$ be any monotone sequence in $\mathcal{B}$ with $B_k \downarrow \emptyset$, and for simplicity also define $B_0 = \mathcal{X}$. We can define a disjoint sequence $\{A_k\}_{k=1}^{\infty}$ in $\mathcal{B}$ with $\bigcup_{k=1}^{\infty} A_k = \mathcal{X}$ by letting $A_k = B_{k-1} \setminus B_k$ for each $k \in \mathbb{N}$. Then note that $|\{k \in \mathbb{N} : \mathbb{X} \cap B_k \neq \emptyset\}| = \sup\{k \in \mathbb{N} \cup \{0\} : \mathbb{X} \cap B_k \neq \emptyset\} = \sup\{k \in \mathbb{N} \cup \{0\} : \mathbb{X} \cap A_{k+1} \neq \emptyset\}$, and this rightmost quantity is finite if and only if $|\{k \in \mathbb{N} : \mathbb{X} \cap A_k \neq \emptyset\}| < \infty$. Together with (58), this implies $|\{k \in \mathbb{N} : \mathbb{X} \cap B_k \neq \emptyset\}| < \infty$ (a.s.). Since this holds for every such sequence $\{B_k\}_{k=1}^{\infty}$, it follows that $\mathbb{X}$ satisfies Condition 3. ■

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 $\lim_{k\to\infty} \hat{\mu}_{\mathbb{X}}(A_k) = 0$; 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 50 *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 $\bar{\ell} = \infty$, $\mathrm{SUOL} = \mathrm{SUAL} = \mathrm{SUIL} = \mathcal{C}_3$.

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 41) 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 50 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. This

again contrasts with the results for bounded losses, in the inductive setting (cf. Theorem 6). We have the following theorem, the proof of which is given in Section 8.3 below.

Theorem 51 *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 (68) 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 50 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 somewhat 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;[10] 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 52 *A process $\mathbb{X}$ satisfies Condition 3 if and only if every measurable function $f : \mathcal{X} \to \mathbb{R}$ satisfies*

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

10. 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.

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 $\sup_{t \in \mathbb{N}} f(X_t) \leq k_0 < \infty$.

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). Therefore $\sup_{t \in \mathbb{N}} f(X_t) < \infty$ (a.s.). Also note that monotonicity of the sequence $\{A_k\}_{k=1}^{\infty}$ implies $\forall x \in \mathcal{X}$, $f(x) = \max(\{k \in \mathbb{N} : x \in A_k\} \cup \{0\})$. Thus, defining $\hat{k} = \sup_{t \in \mathbb{N}} f(X_t)$, 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 $|\{k \in \mathbb{N} : \mathbb{X} \cap A_k \neq \emptyset\}| \leq \hat{k} < \infty$ (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 50 and 51. As with Theorem 7, we prove Theorem 50 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 19, showing that processes admitting strong universal inductive learning also admit strong universal self-adaptive learning. The proof is identical to that of Lemma 19, and as such is omitted.

Lemma 53 *When $\bar{\ell} = \infty$,* SUIL $\subseteq$ SUAL.

Next, we have a result analogous to Lemma 20, showing that any process admitting strong universal self-adaptive or online learning necessarily satisfies Condition 3.

Lemma 54 *When $\bar{\ell} = \infty$,* SUAL $\cup$ 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,

$$|\{k \in \mathbb{N} : \mathbb{X} \cap B_k \neq \emptyset\}| = \infty. \tag{59}$$

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, (59) implies that

$$\forall k \in \mathbb{N}, \mathbb{X} \cap B_k \neq \emptyset. \tag{60}$$

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 in $\mathcal{B}$ with

$\bigcup_{i=1}^{\infty} A_i = \mathcal{X}$. Thus, for every $t \in \mathbb{N}$, there exists a unique (X_t-dependent) variable $i_t \in \mathbb{N}$ 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, (60) 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 is almost surely finite, 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 $i \in \mathbb{N}$, define $\kappa_i = \lfloor 2^i \kappa \rfloor - 2\lfloor 2^{i-1}\kappa \rfloor$: 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 $(x,\kappa) \mapsto f_\kappa^\star(x)$ is measurable in the product σ-algebra (under $\mathcal{B}$ for the x argument, and the usual Borel σ-algebra on $[0,1)$ for the κ argument), since the inverse image of any measurable set $C \subseteq \mathcal{Y}$ is a countable union of measurable rectangle sets: namely, $\bigcup_{\substack{i \in \mathbb{N}, b \in \{0,1\}: \\ y_{i,b} \in C}} (A_i \times \{\kappa : \kappa_i = b\})$.

For the purpose of treating both self-adaptive and online learning simultaneously, for any $n, m \in \mathbb{N} \cup \{0\}$, let $f_{n,m}$ denote any (possibly random) 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 define $f_{n,m}^\kappa(x) = f_{n,m}(X_{1:m}, f_\kappa^\star(X_{1:m}), x)$ (a composition of measurable functions, and therefore measurable); equivalently, $f_{n,m}^\kappa(x) = f_{n,m}(X_{1:m}, \{y_{i_t,\kappa_{i_t}}\}_{t=1}^m, x)$. We generally have

$$\begin{aligned} &\sup_{\kappa \in [0,1)} \mathbb{E}\left[\operatorname*{limsup}_{n\to\infty} \operatorname*{limsup}_{t\to\infty} \frac{1}{t} \sum_{m=0}^{t-1} \ell\Big(f_{n,m}^\kappa(X_{m+1}), y_{i_{m+1},\kappa_{i_{m+1}}}\Big)\right] \\ &\geq \int_0^1 \mathbb{E}\left[\operatorname*{limsup}_{n\to\infty} \operatorname*{limsup}_{t\to\infty} \frac{1}{t} \sum_{m=0}^{t-1} \ell\Big(f_{n,m}^\kappa(X_{m+1}), y_{i_{m+1},\kappa_{i_{m+1}}}\Big)\right] \mathrm{d}\kappa. \end{aligned} \tag{61}$$

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 (61) equals

$$\mathbb{E}\left[\int_0^1 \limsup_{n\to\infty}\limsup_{t\to\infty}\frac{1}{t}\sum_{m=0}^{t-1}\ell\Big(f^{\kappa}_{n,m}(X_{m+1}), y_{i_{m+1},\kappa_{i_{m+1}}}\Big)\,\mathrm{d}\kappa\right]$$
$$\geq \mathbb{E}\left[\mathbb{1}_{E\cap E'}\int_0^1 \limsup_{n\to\infty}\limsup_{t\to\infty}\frac{1}{t}\sum_{m=0}^{t-1}\ell\Big(f^{\kappa}_{n,m}(X_{m+1}), y_{i_{m+1},\kappa_{i_{m+1}}}\Big)\,\mathrm{d}\kappa\right]. \tag{62}$$

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 (60) 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 on the right hand side of (62) is at least as large as

$$\mathbb{E}\left[\mathbb{1}_{E\cap E'}\int_0^1 \limsup_{n\to\infty}\limsup_{j\to\infty}\frac{1}{\tau_j}\sum_{m=0}^{\tau_j-1}\ell\Big(f^{\kappa}_{n,m}(X_{m+1}), y_{i_{m+1},\kappa_{i_{m+1}}}\Big)\,\mathrm{d}\kappa\right]$$
$$\geq \mathbb{E}\left[\mathbb{1}_{E\cap E'}\int_0^1 \limsup_{n\to\infty}\limsup_{j\to\infty}\frac{1}{\tau_j}\left(\ell\Big(f^{\kappa}_{n,\tau_j-1}(X_{\tau_j}), y_{i_{\tau_j},\kappa_{i_{\tau_j}}}\Big)\wedge\tau_j\right)\mathrm{d}\kappa\right]. \tag{63}$$

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 (63) is at least as large as

$$\mathbb{E}\left[\mathbb{1}_{E\cap E'}\limsup_{n\to\infty}\limsup_{j\to\infty}\frac{1}{\tau_j}\int_0^1\left(\ell\Big(f^{\kappa}_{n,\tau_j-1}(X_{\tau_j}), y_{i_{\tau_j},\kappa_{i_{\tau_j}}}\Big)\wedge\tau_j\right)\mathrm{d}\kappa\right]. \tag{64}$$

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

$$\int_0^1\left(\ell\Big(f^{\kappa}_{n,\tau_j-1}(X_{\tau_j}), y_{i_{\tau_j},\kappa_{i_{\tau_j}}}\Big)\wedge\tau_j\right)\mathrm{d}\kappa = \mathbb{E}\Big[\Big(\ell\Big(f^{K}_{n,\tau_j-1}(X_{\tau_j}), y_{i_{\tau_j},K_{i_{\tau_j}}}\Big)\wedge\tau_j\Big)\Big|\mathbb{X}, f_{n,\tau_j-1}\Big]$$
$$= \mathbb{E}\Big[\mathbb{E}\Big[\Big(\ell\Big(f_{n,\tau_j-1}\Big(X_{1:(\tau_j-1)}, \{y_{i_t,K_{i_t}}\}_{t=1}^{\tau_j-1}, X_{\tau_j}\Big), y_{i_{\tau_j},K_{i_{\tau_j}}}\Big)\wedge\tau_j\Big)\Big|\mathbb{X}, \{K_{i_t}\}_{t=1}^{\tau_j-1}, f_{n,\tau_j-1}\Big]\Big|\mathbb{X}, f_{n,\tau_j-1}\Big]$$
$$= \mathbb{E}\left[\sum_{b\in\{0,1\}}\frac{1}{2}\left(\ell\Big(f_{n,\tau_j-1}\Big(X_{1:(\tau_j-1)}, \{y_{i_t,K_{i_t}}\}_{t=1}^{\tau_j-1}, X_{\tau_j}\Big), y_{i_{\tau_j},b}\Big)\wedge\tau_j\right)\Big|\mathbb{X}, f_{n,\tau_j-1}\right]. \tag{65}$$

Since $\tau_j \geq 0$, one can easily verify that $\ell(\cdot,\cdot) \wedge \tau_j$ is a pseudo-near-metric (i.e., a near-metric except that $\ell(y,y')$ might sometimes be 0 even for $y \neq y'$) with c_ℓ as the constant in the relaxed triangle inequality. Thus, by the relaxed triangle inequality,

$$\sum_{b\in\{0,1\}} \left(\ell\Big(f_{n,\tau_j-1}\left(X_{1:(\tau_j-1)}, \{y_{i_s,K_{i_s}}\}_{s=1}^{\tau_j-1}, X_{\tau_j}\right), y_{i_{\tau_j},b}\Big) \wedge \tau_j\right)$$
$$\geq \frac{1}{c_\ell}\left(\ell\Big(y_{i_{\tau_j},0}, y_{i_{\tau_j},1}\Big) \wedge \tau_j\right) \geq \frac{1}{c_\ell}\left(T_{i_{\tau_j}} \wedge \tau_j\right). \quad (66)$$

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 (65) and (66) yields that, on the event $E \cap E'$, $\forall n, j \in \mathbb{N}$ with $j \geq \iota_0$,

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

where the rightmost equality follows from $\sigma(\mathbb{X})$-measurability of τ_j. Therefore, the expression in (64) is at least as large as

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

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 (61) is strictly greater than 0. By the inequality in (61) this implies $\exists \kappa \in [0,1)$ such that

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

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

$$\operatorname*{limsup}_{n\to\infty} \operatorname*{limsup}_{t\to\infty} \frac{1}{t}\sum_{m=0}^{t-1} \ell\Big(f^\kappa_{n,m}(X_{m+1}), y_{i_{m+1},\kappa_{i_{m+1}}}\Big) > 0.$$

This argument applies to any measurable functions $f_{n,m} : \mathcal{X}^m \times \mathcal{Y}^m \times \mathcal{X} \to \mathcal{Y}$ (possibly random). 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

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

Therefore, the above argument implies that $\exists \kappa \in [0,1)$ for which, with probability strictly greater than 0, $\operatorname*{limsup}_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(h_\cdot, f^\star_\kappa; n) > 0$, 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 $\mathbb{X} \notin \text{SUOL}$, and since the argument applies to any process $\mathbb{X}$ failing to satisfy Condition 3, we conclude that $\text{SUOL} \subseteq \mathcal{C}_3$.

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 $\mathbb{X} \notin \text{SUAL}$, and since the argument applies to any process $\mathbb{X}$ failing to satisfy Condition 3, we conclude that $\text{SUAL} \subseteq \mathcal{C}_3$, 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 $\varepsilon_k = (2c_\ell)^{-k}$. Given a sequence $\{\tilde{f}_i\}_{i=1}^\infty$ of measurable functions $\mathcal{X} \to \mathcal{Y}$ (described below), 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

$$\begin{aligned}\hat{i}_{n,k}(x_{1:n}, y_{1:n}) = \min\Bigg\{i \in \mathbb{N} : \max_{1\leq t\leq n} \ell\Big(\tilde{f}_i(x_t), y_t\Big) \leq \varepsilon_k, \text{ and} \\ \sup_{x\in\mathcal{X}} \ell\Big(\tilde{f}_i(x), \tilde{f}_{\hat{i}_{n,k-1}(x_{1:n},y_{1:n})}(x)\Big) \leq c_\ell \varepsilon_{k-1} + \varepsilon_k\Bigg\}, \quad (67)\end{aligned}$$

if it exists. For completeness, if the set on the right hand side of (67) is empty for a given $k \in \mathbb{N}$, let us define $\hat{i}_{n,k}(x_{1:n}, y_{1:n}) = \hat{i}_{n,k-1}(x_{1:n}, y_{1:n})$. Fix any sequence $\{k_n\}_{n=1}^\infty$ in $\mathbb{N}$ with $k_n \to \infty$. Then, 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

$$\hat{f}_n(x_{1:n}, y_{1:n}, x) = \tilde{f}_{\hat{i}_{n,k_n}(x_{1:n},y_{1:n})}(x). \quad (68)$$

We will argue in the proof of Lemma 57 below that $\hat{f}_n$ is measurable, and hence (68) defines a valid inductive learning rule. 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 23 and 25, under the restriction of $\mathbb{X}$ to $\mathcal{C}_3$.

Lemma 55 *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 23, 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 23. As in the proof of Lemma 23, 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 23, we will instead establish the claim with a *smaller* set $\mathcal{T}_1$, which thereby simplifies the problem of implementing the resulting learning rule. Specifically, as in the proof of Lemma 23, 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

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

For any $A \in \mathcal{T}$, as mentioned in the proof of Lemma 23, $\exists \{B_i\}_{i=1}^{\infty}$ 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). Defining $C_k = \bigcap_{i=1}^{k} B_i$ 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), $\mathbb{X} \cap \hat{A}_{k_0} = \mathbb{X} \cap (\mathcal{X} \setminus A)$. Thus, every $A \in \mathcal{T}_1$ has $(\mathcal{X} \setminus A) \in \Lambda$. Now define $E^{(\mathcal{T}_1)} = \bigcap_{A \in \mathcal{T}_1} \bigcap_{k=0}^{\infty} E_k^{(A)}$, which has probability one by the union bound (since $\mathcal{T}_1$ is countable).

Next, consider any $A \in \Lambda$, and suppose the event (of probability one), denoted E', that $\exists \hat{A} \in \mathcal{T}_1$ s.t. $\mathbb{X} \cap \hat{A} = \mathbb{X} \cap A$ holds, 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 let $A = \bigcup_{i=1}^{\infty} A_i$. 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 follows. ■

Lemma 56 *There exists a countable set $\tilde{\mathcal{F}}$ of measurable functions $\mathcal{X} \to \mathcal{Y}$ such that, for every $\mathbb{X} \in \mathcal{C}_3$, for every measurable $f : \mathcal{X} \to \mathcal{Y}$, with probability one, $\forall \varepsilon > 0$, $\forall \tilde{f}_1 \in \tilde{\mathcal{F}}$, $\exists \tilde{f}_2 \in \tilde{\mathcal{F}}$ with*

$$\sup_{x \in \mathcal{X}} \ell\Big(\tilde{f}_2(x), \tilde{f}_1(x)\Big) \leq \varepsilon + c_\ell \sup_{t \in \mathbb{N}} \ell\Big(\tilde{f}_1(X_t), f(X_t)\Big)$$

and $$\sup_{t \in \mathbb{N}} \ell\Big(\tilde{f}_2(X_t), f(X_t)\Big) \leq \varepsilon.$$

Proof The construction, and first half of this proof, will proceed analogously to the proof of Lemma 24, but with a few important changes. Specifically, let $\mathcal{T}_1$ be as in Lemma 55, let $\tilde{\mathcal{Y}}$ be a countable subset of $\mathcal{Y}$ with $\sup_{y \in \mathcal{Y}} \inf_{\tilde{y} \in \tilde{\mathcal{Y}}} \ell(\tilde{y}, y) = 0$ (which exists by separability of $(\mathcal{Y}, \ell)$), let $A_0 = \mathcal{X}$, let y_0 be an arbitrary value in $\mathcal{Y}$, and for any $k \in \mathbb{N}$, any sequence $\{A_i\}_{i=1}^{k}$ in $\mathcal{B}$, and any sequence $\{y_i\}_{i=1}^{k}$ in $\mathcal{Y}$, define $\tilde{f}(x; \{y_i\}_{i=1}^{k}, \{A_i\}_{i=1}^{k}) = y_{\max\{i \in \{0,\ldots,k\} : x \in A_i\}}$. Define $\mathcal{T}_2 = \left\{ \bigcap_{i=1}^{k} A_i : k \in \mathbb{N}, A_1, \ldots, A_k \in \mathcal{T}_1 \right\}$: the finite intersections of sets in $\mathcal{T}_1$. This is a countable set since $\mathcal{T}_1$ is countable. Then define

$$\tilde{\mathcal{F}} = \left\{ \tilde{f}(\cdot; \{y_i\}_{i=1}^{k}, \{A_i\}_{i=1}^{k}) : k \in \mathbb{N}, \forall i \leq k, y_i \in \tilde{\mathcal{Y}}, A_i \in \mathcal{T}_2 \right\},$$

which is a countable set (since $\tilde{\mathcal{Y}}$ and $\mathcal{T}_2$ are countable). Enumerate the elements of $\tilde{\mathcal{Y}}$ as $\tilde{y}_1, \tilde{y}_2, \ldots$; for simplicity, we will suppose this sequence is infinite, which is always the case

if $\bar{\ell} = \infty$, and otherwise can be achieved by repeating elements if necessary in the general case. As in the proof of Lemma 24, for each $\varepsilon > 0$, let $B_{\varepsilon,1} = \{y \in \mathcal{Y} : \ell(\tilde{y}_1, y) \leq \varepsilon\}$ and for each integer $i \geq 2$ inductively define $B_{\varepsilon,i} = \{y \in \mathcal{Y} : \ell(\tilde{y}_i, y) \leq \varepsilon\} \setminus \bigcup_{j=1}^{i-1} B_{\varepsilon,j}$. For each $\varepsilon > 0$, this defines a disjoint sequence $\{B_{\varepsilon,i}\}_{i=1}^{\infty}$ in $\mathcal{B}_y$ with $\bigcup_{i=1}^{\infty} B_{\varepsilon,i} = \mathcal{Y}$.

Fix any $\mathbb{X} \in \mathcal{C}_3$, any measurable $f : \mathcal{X} \to \mathcal{Y}$, and any $\varepsilon > 0$. For each $i \in \mathbb{N}$, define $C_{\varepsilon,i} = f^{-1}(B_{\varepsilon,i})$, 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 since the $B_{\varepsilon,i}$ sets are disjoint over the values of i, the sets $C_{\varepsilon,i}$ are also disjoint over i. 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 3 implies that, on an event $E_{\varepsilon,1}$ of probability one, $\exists k_0 \in \mathbb{N}$ s.t. $\mathbb{X} \cap \bigcup_{i=k_0+1}^{\infty} C_{\varepsilon,i} = \emptyset$. Since $\bigcup_{i=1}^{\infty} C_{\varepsilon,i} = \mathcal{X}$, this also means $\mathbb{X} \cap \bigcup_{i=1}^{k_0} C_{\varepsilon,i} = \mathbb{X}$. Furthermore, by the union bound and the defining property of $\mathcal{T}_1$ from Lemma 55, on an event $E_{\varepsilon,2}$ of probability one, $\forall i \in \mathbb{N}$, $\exists \tilde{A}_{\varepsilon,i} \in \mathcal{T}_1$ with $\mathbb{X} \cap \tilde{A}_{\varepsilon,i} = \mathbb{X} \cap C_{\varepsilon,i}$. This also means that, when the events $E_{\varepsilon,1}$ and $E_{\varepsilon,2}$ occur simultaneously, we have $\mathbb{X} \cap \bigcup_{i=1}^{k_0} \tilde{A}_{\varepsilon,i} = \mathbb{X}$.

At this point, we may note that the function $\tilde{f}(\cdot; \{\tilde{y}_i\}_{i=1}^{k_0}, \{\tilde{A}_{\varepsilon,i}\}_{i=1}^{k_0})$ would suffice as a specification of $\tilde{f}_2$ for the purpose of satisfying the *second* requirement in the lemma: that is, $\sup_{t\in\mathbb{N}} \ell\Big(\tilde{f}_2(X_t), f(X_t)\Big) \leq \varepsilon$. Indeed if this were the only requirement, we could have used $\mathcal{T}_1$ in place of $\mathcal{T}_2$ in the specification of $\tilde{\mathcal{F}}$ above. However, we must modify this function in order to also satisfy the first requirement, constraining the supremum distance between $\tilde{f}_2$ and a given $\tilde{f}_1$ in $\tilde{\mathcal{F}}$.

Toward this end, supposing the event $E_{\varepsilon,1} \cap E_{\varepsilon,2}$ occurs and that k_0 and $\{\tilde{A}_{\varepsilon,i}\}_{i=1}^{k_0}$ are as above, fix any function $\tilde{f}_1 \in \tilde{\mathcal{F}}$, and let $k_1 \in \mathbb{N}$, $\{y_i\}_{i=1}^{k_1} \in \tilde{\mathcal{Y}}^{k_1}$, and $\{A_i\}_{i=1}^{k_1} \in \mathcal{T}_2^{k_1}$ be such that $\tilde{f}_1(\cdot) = \tilde{f}(\cdot; \{y_i\}_{i=1}^{k_1}, \{A_i\}_{i=1}^{k_1})$. Let A_0 and y_0 be as specified above, and define $\tilde{A}_{\varepsilon,0} = \mathcal{X}$ and $\tilde{y}_0 = y_0$. Let $k_2 = (k_0+1)(k_1+1) - 1$. For each $i \in \{0, \ldots, k_0\}$ and $j \in \{0, 1, \ldots, k_1\}$, define $\hat{A}_{j(k_0+1)+i} = \tilde{A}_{\varepsilon,i} \cap A_j$. Also, for each $i \in \{0, \ldots, k_0\}$ and $j \in \{0, 1, \ldots, k_1\}$, define $\hat{D}_{i,j} = \hat{A}_{j(k_0+1)+i} \setminus \bigcup_{j'=j(k_0+1)+i+1}^{k_2} \hat{A}_{j'}$; if $\mathbb{X} \cap \hat{D}_{i,j} \neq \emptyset$, define $\hat{y}_{j(k_0+1)+i} = \tilde{y}_i$, and otherwise define $\hat{y}_{j(k_0+1)+i} = y_j$. Note that $\hat{A}_1, \ldots, \hat{A}_{k_2}$ are elements of $\mathcal{T}_2$ and $\hat{y}_1, \ldots, \hat{y}_{k_2}$ are elements of $\tilde{\mathcal{Y}}$, and therefore, defining $\tilde{f}_2(\cdot) = \tilde{f}(\cdot; \{\hat{y}_i\}_{i=1}^{k_2}, \{\hat{A}_i\}_{i=1}^{k_2})$, we have that $\tilde{f}_2 \in \tilde{\mathcal{F}}$.

Now, for every $x \in \mathcal{X}$, denote by $(\mathrm{i}(x), \mathrm{j}(x))$ the unique value in $\{0, \ldots, k_0\} \times \{0, \ldots, k_1\}$ such that $\mathrm{j}(x)(k_0+1) + \mathrm{i}(x) = \max\Big\{j' \in \{0, \ldots, k_2\} : x \in \hat{A}_{j'}\Big\}$ (noting that $\hat{A}_0 = \mathcal{X}$, so that this is always well-defined). By definition, we have $\tilde{f}_2(x) = \hat{y}_{\mathrm{j}(x)(k_0+1)+\mathrm{i}(x)}$ (noting that $\hat{y}_0 = y_0$, so that this equality holds even when $(\mathrm{i}(x), \mathrm{j}(x)) = (0,0)$). Since $\tilde{A}_{\varepsilon,0} = \mathcal{X}$, it holds that every $j \in \{0, \ldots, k_1\}$ has $\hat{A}_{j(k_0+1)} = A_j$. In particular, this implies that, for every $x \in \mathcal{X}$, it holds that $\mathrm{j}(x) = \max\{j' \in \{0, \ldots, k_1\} : x \in A_{j'}\}$: that is, by definition of $(\mathrm{i}(x), \mathrm{j}(x))$, we have $x \in \hat{A}_{\mathrm{j}(x)(k_0+1)+\mathrm{i}(x)} \subseteq A_{\mathrm{j}(x)}$, and maximality of $\mathrm{j}(x)(k_0+1)+\mathrm{i}(x)$ implies

that every $j' \in \{\mathrm{j}(x)+1, \dots, k_1\}$ has $x \notin \hat{A}_{j'(k_0+1)} = A_{j'}$. Moreover, for any $i \in \{0, \dots, k_0\}$, if $x \in \tilde{A}_{\varepsilon,i}$, then since $x \in A_{\mathrm{j}(x)}$ as well, we have $x \in \hat{A}_{\mathrm{j}(x)(k_0+1)+i} = \tilde{A}_{\varepsilon,i} \cap A_{\mathrm{j}(x)}$; in particular, this is true of the *largest* $i \in \{0, \dots, k_0\}$ with $x \in \tilde{A}_{\varepsilon,i}$. It immediately follows that $\mathrm{i}(x) = \max\{i' \in \{0, \dots, k_0\} : x \in \tilde{A}_{\varepsilon,i'}\}$.

Now note that, for any $t \in \mathbb{N}$, since $X_t \in \hat{D}_{\mathrm{i}(X_t),\mathrm{j}(X_t)}$, we have $\mathbb{X} \cap \hat{D}_{\mathrm{i}(X_t),\mathrm{j}(X_t)} \neq \emptyset$. Therefore, by definition, $\hat{y}_{\mathrm{j}(X_t)(k_0+1)+\mathrm{i}(X_t)} = \tilde{y}_{\mathrm{i}(X_t)}$. Furthermore, since $\hat{A}_{\mathrm{j}(X_t)(k_0+1)+\mathrm{i}(X_t)} \subseteq \tilde{A}_{\varepsilon,\mathrm{i}(X_t)}$, we have $X_t \in \tilde{A}_{\varepsilon,\mathrm{i}(X_t)}$, and since $\mathbb{X} \cap \bigcup_{i=1}^{k_0} \tilde{A}_{\varepsilon,i} = \mathbb{X}$, there exists $i \in \{1, \dots, k_0\}$ with $X_t \in \tilde{A}_{\varepsilon,i}$. Therefore, the fact (established above) that $\mathrm{i}(X_t) = \max\{i' \in \{0, \dots, k_0\} : X_t \in \tilde{A}_{\varepsilon,i'}\}$ implies $\mathrm{i}(X_t) \neq 0$. Since every $i \in \mathbb{N}$ has $\mathbb{X} \cap \tilde{A}_{\varepsilon,i} = \mathbb{X} \cap C_{\varepsilon,i}$, this further implies that $X_t \in C_{\varepsilon,\mathrm{i}(X_t)}$, and therefore $\ell\big(\tilde{y}_{\mathrm{i}(X_t)}, f(X_t)\big) \leq \varepsilon$, so that altogether we have $\ell\Big(\tilde{f}_2(X_t), f(X_t)\Big) \leq \varepsilon$. Since this is true of every $t \in \mathbb{N}$, we conclude that $\sup_{t\in\mathbb{N}} \ell\Big(\tilde{f}_2(X_t), f(X_t)\Big) \leq \varepsilon$.

Next, note that for any $x \in \mathcal{X}$, since (as established above) $\mathrm{j}(x) = \max\{j' \in \{0, \dots, k_1\} : x \in A_{j'}\}$, we have $\tilde{f}_1(x) = y_{\mathrm{j}(x)}$. In particular, if $\mathbb{X} \cap \hat{D}_{\mathrm{i}(x),\mathrm{j}(x)} = \emptyset$, then, by definition, we have $\tilde{f}_2(x) = y_{\mathrm{j}(x)}$, so that in this case $\ell\Big(\tilde{f}_2(x), \tilde{f}_1(x)\Big) = \ell(y_{\mathrm{j}(x)}, y_{\mathrm{j}(x)}) = 0$. On the other hand, if $\mathbb{X} \cap \hat{D}_{\mathrm{i}(x),\mathrm{j}(x)} \neq \emptyset$, then, by definition, we have $\tilde{f}_2(x) = \tilde{y}_{\mathrm{i}(x)}$. In this case, letting $t \in \mathbb{N}$ be such that $X_t \in \hat{D}_{\mathrm{i}(x),\mathrm{j}(x)}$, it immediately follows that $(\mathrm{i}(X_t), \mathrm{j}(X_t)) = (\mathrm{i}(x), \mathrm{j}(x))$, so that we also have $\tilde{f}_1(X_t) = y_{\mathrm{j}(x)}$ and $\tilde{f}_2(X_t) = \tilde{y}_{\mathrm{i}(x)}$. Thus, in this case we have $\ell\Big(\tilde{f}_2(x), \tilde{f}_1(x)\Big) = \ell\big(\tilde{y}_{\mathrm{i}(x)}, y_{\mathrm{j}(x)}\big) = \ell\Big(\tilde{f}_2(X_t), \tilde{f}_1(X_t)\Big)$. Since every $x \in \mathcal{X}$ satisfies one of these two cases, and since each X_t itself takes a value in $\mathcal{X}$, we conclude that

$$\sup_{x\in\mathcal{X}} \ell\Big(\tilde{f}_2(x), \tilde{f}_1(x)\Big) = \sup_{t\in\mathbb{N}} \ell\Big(\tilde{f}_2(X_t), \tilde{f}_1(X_t)\Big).$$

Combining this with the relaxed triangle inequality and the fact (established above) that $\sup_{t\in\mathbb{N}} \ell\Big(\tilde{f}_2(X_t), f(X_t)\Big) \leq \varepsilon$, we conclude that

$$\begin{aligned}\sup_{x\in\mathcal{X}} \ell\Big(\tilde{f}_2(x), \tilde{f}_1(x)\Big) &\leq c_\ell \sup_{t\in\mathbb{N}} \ell\Big(\tilde{f}_2(X_t), f(X_t)\Big) + c_\ell \sup_{t\in\mathbb{N}} \ell\Big(\tilde{f}_1(X_t), f(X_t)\Big) \\ &\leq c_\ell\varepsilon + c_\ell \sup_{t\in\mathbb{N}} \ell\Big(\tilde{f}_1(X_t), f(X_t)\Big).\end{aligned}$$

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 $\bigcap_{k=1}^{\infty} \Big(E_{\varepsilon'_k,1} \cap E_{\varepsilon'_k,2}\Big)$, for any $\tilde{f}_1 \in \tilde{\mathcal{F}}$ and any $\varepsilon > 0$, letting $k = \lceil \log_2((c_\ell/\varepsilon) \vee 2) \rceil$, we have that $\exists \tilde{f}_2 \in \tilde{\mathcal{F}}$ with

$$\sup_{t\in\mathbb{N}} \ell(\tilde{f}_2(X_t), f(X_t)) \leq \varepsilon'_k \leq \varepsilon,$$

and

$$\sup_{x\in\mathcal{X}} \ell(\tilde{f}_2(x), \tilde{f}_1(x)) \leq c_\ell\varepsilon'_k + c_\ell \sup_{t\in\mathbb{N}} \ell(\tilde{f}_1(X_t), f(X_t)) \leq \varepsilon + c_\ell \sup_{t\in\mathbb{N}} \ell(\tilde{f}_1(X_t), f(X_t)).$$

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

We are now ready to present a result establishing that any process satisfying Condition 3 necessarily admits strong universal inductive (and online) learning. This is analogous to Lemma 27 from the bounded case. 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 57 *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 (68), where the sequence $\{\tilde{f}_i\}_{i=1}^{\infty}$ is chosen as an enumeration of the elements of the countable set $\tilde{\mathcal{F}}$ from Lemma 56. 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.

To verify that $\hat{f}_n$ is a measurable function, we note that any measurable $B \subseteq \mathcal{Y}$ has $\hat{f}_n^{-1}(B) = \bigcup_{i\in\mathbb{N}}\left(\hat{i}_{n,k_n}^{-1}(\{i\}) \times \mathcal{X}\right) \cap \left(\mathcal{X}^n \times \mathcal{Y}^n \times \tilde{f}_i^{-1}(B)\right)$. Since each $\tilde{f}_i$ is a measurable function, it suffices to verify measurability of $\hat{i}_{n,k}$ for all n, k. Note that $\hat{i}_{n,0}$ is constant, hence trivially measurable. For the purpose of induction, let us suppose some $k \in \mathbb{N}$ has $\hat{i}_{n,k-1}$ measurable. For any $i \in \mathbb{N}$, let $J_{i,k} = \left\{j \in \mathbb{N} : \sup_{x\in\mathcal{X}} \ell\left(\tilde{f}_i(x), \tilde{f}_j(x)\right) \le c_\ell \varepsilon_{k-1} + \varepsilon_k\right\}$, and $A_{i,k} = \hat{i}_{n,k-1}^{-1}(J_{i,k}) \cap \left\{(x_{1:n}, y_{1:n}) : \max_{1\le t\le n} \ell\left(\tilde{f}_i(x_t), y_t\right) \le \varepsilon_k\right\}$. Since $\hat{i}_{n,k-1}$ is measurable (by assumption) and ℓ and $\tilde{f}_i$ are measurable functions, we observe that $A_{i,k}$ is a measurable set. Then note that any $i \in \mathbb{N}$ has $\hat{i}_{n,k}^{-1}(\{i\}) = \left(A_{i,k} \setminus \bigcup_{i'<i} A_{i',k}\right) \cup \left(\hat{i}_{n,k-1}^{-1}(\{i\}) \setminus \bigcup_{i'\in\mathbb{N}} A_{i',k}\right)$, where the second term is due to the case when the set on the right hand side of (67) is empty. Thus, $\hat{i}_{n,k}^{-1}(\{i\})$ is a measurable set. Since any $C \subseteq \mathbb{N}$ has $\hat{i}_{n,k}^{-1}(C) = \bigcup_{i\in C} \hat{i}_{n,k}^{-1}(\{i\})$, we conclude that $\hat{i}_{n,k}$ is measurable, and this holds for all k by the principle of induction. Therefore, $\hat{f}_n$ is a valid inductive learning rule.

Now 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 56, 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

$$\sup_{x\in\mathcal{X}} \ell\left(\tilde{f}_j(x), \tilde{f}_i(x)\right) \le \varepsilon + c_\ell \sup_{t\in\mathbb{N}} \ell\left(\tilde{f}_i(X_t), f^\star(X_t)\right) \tag{69}$$

$$\text{and} \quad \sup_{t\in\mathbb{N}} \ell\left(\tilde{f}_j(X_t), f^\star(X_t)\right) \le \varepsilon. \tag{70}$$

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, $\forall n \ge n_k^*$, $\hat{i}_{n,k} = i_k^*$ and $\sup_{t\in\mathbb{N}} \ell\left(\tilde{f}_{i_k^*}(X_t), f^\star(X_t)\right) \le \varepsilon_k$, for ε_k as defined above (67). In particular, as

a base case, let us define $i_0^* = 1$ and $n_0^* = 1$, for which the claims trivially hold since we have defined $\hat{i}_{n,0} = 1$ for every $n \in \mathbb{N}$, and 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, $\forall n \geq n_{k-1}^*$, it holds that $\hat{i}_{n,k-1} = i_{k-1}^*$ 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

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

Note that, taking $\varepsilon = \varepsilon_k$ and $i = i_{k-1}^*$ in (69) and (70), 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 $\sup_{x\in\mathcal{X}} \ell\Big(\tilde{f}_j(x), \tilde{f}_{i_{k-1}^*}(x)\Big) \leq c_\ell\varepsilon_{k-1} + \varepsilon_k$ (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 $\ell\Big(\tilde{f}_j(X_{t_{j,k}}), f^\star(X_{t_{j,k}})\Big) > \varepsilon_k$. Now define

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

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 $\sup_{x\in\mathcal{X}} \ell\Big(\tilde{f}_j(x), \tilde{f}_{i_{k-1}^*}(x)\Big) \leq c_\ell\varepsilon_{k-1} + \varepsilon_k$ 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 (67) (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 $\sup_{x\in\mathcal{X}} \ell\Big(\tilde{f}_j(x), \tilde{f}_{i_{k-1}^*}(x)\Big) > c_\ell\varepsilon_{k-1} + \varepsilon_k$ is clearly also not included in the set on the right hand side of (67) in this case (again, since $\hat{i}_{n,k-1} = i_{k-1}^*$). On the other hand, by definition we have $\sup_{x\in\mathcal{X}} \ell\Big(\tilde{f}_{i_k^*}(x), \tilde{f}_{i_{k-1}^*}(x)\Big) \leq c_\ell\varepsilon_{k-1} + \varepsilon_k$, 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 $\hat{i}_{n,k-1} = i_{k-1}^*$, we have that i_k^* *is* included in the set on the right hand side of (67) (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 $n \geq n_k^*$, it holds that $\hat{i}_{n,k} = i_k^*$.

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,

$$\sup_{t\in\mathbb{N}} \ell\Big(\tilde{f}_{\hat{i}_{n,k_n^*}}(X_t), f^\star(X_t)\Big) \le \varepsilon_{k_n^*}. \tag{71}$$

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 $\sup_{x\in\mathcal{X}} \ell\Big(\tilde{f}_{\hat{i}_{n,k}}(x), \tilde{f}_{\hat{i}_{n,k-1}}(x)\Big) \le c_\ell\varepsilon_{k-1} + \varepsilon_k$ (noting that this is true even when the set on the right hand side of (67) 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 relaxed triangle inequality and subadditivity of the supremum, and noting that $k_n^* \le k_n$ (by definition), this implies

$$\begin{aligned}
&\sup_{x\in\mathcal{X}} \ell\Big(\tilde{f}_{\hat{i}_{n,k_n}}(x), \tilde{f}_{\hat{i}_{n,k_n^*}}(x)\Big) \le \sup_{x\in\mathcal{X}} \sum_{k=k_n^*+1}^{k_n} c_\ell^{k-k_n^*} \ell\Big(\tilde{f}_{\hat{i}_{n,k}}(x), \tilde{f}_{\hat{i}_{n,k-1}}(x)\Big) \\
&\le \sum_{k=k_n^*+1}^{k_n} \sup_{x\in\mathcal{X}} c_\ell^{k-k_n^*} \ell\Big(\tilde{f}_{\hat{i}_{n,k}}(x), \tilde{f}_{\hat{i}_{n,k-1}}(x)\Big) \\
&\le \sum_{k=k_n^*+1}^{k_n} c_\ell^{k-k_n^*} \left(c_\ell\varepsilon_{k-1} + \varepsilon_k\right) \le \sum_{k=k_n^*+1}^{\infty} c_\ell^{k-k_n^*} \left(c_\ell\varepsilon_{k-1} + \varepsilon_k\right).
\end{aligned}$$

If $k_n^* \ge 1$, then by our choice of $\varepsilon_k = (2c_\ell)^{-k}$ for every $k \in \mathbb{N}$, the rightmost expression above equals $c_\ell^{-k_n^*}(2c_\ell^2+1)\cdot 2^{-k_n^*} = (2c_\ell^2+1)\varepsilon_{k_n^*}$; on the other hand, if $k_n^* = 0$, then our choice of $\varepsilon_0 = \infty$ implies the expression is $\infty = (2c_\ell^2+1)\varepsilon_0$. Thus, either way, we have

$$\sup_{x\in\mathcal{X}} \ell\Big(\tilde{f}_{\hat{i}_{n,k_n}}(x), \tilde{f}_{\hat{i}_{n,k_n^*}}(x)\Big) \le (2c_\ell^2+1)\varepsilon_{k_n^*}. \tag{72}$$

Therefore, by the relaxed triangle inequality, $\forall n \in \mathbb{N}$,

$$\begin{aligned}
&\sup_{t\in\mathbb{N}} \ell\Big(\tilde{f}_{\hat{i}_{n,k_n}}(X_t), f^\star(X_t)\Big) \le \sup_{t\in\mathbb{N}} c_\ell \left(\ell\Big(\tilde{f}_{\hat{i}_{n,k_n^*}}(X_t), f^\star(X_t)\Big) + \ell\Big(\tilde{f}_{\hat{i}_{n,k_n}}(X_t), \tilde{f}_{\hat{i}_{n,k_n^*}}(X_t)\Big)\right) \\
&\le c_\ell \sup_{t\in\mathbb{N}} \ell\Big(\tilde{f}_{\hat{i}_{n,k_n^*}}(X_t), f^\star(X_t)\Big) + c_\ell \sup_{x\in\mathcal{X}} \ell\Big(\tilde{f}_{\hat{i}_{n,k_n}}(x), \tilde{f}_{\hat{i}_{n,k_n^*}}(x)\Big) \le 2(c_\ell^3 + c_\ell)\varepsilon_{k_n^*},
\end{aligned}$$

where the last inequality is due to (71) and (72). 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 (68), and ℓ is non-negative, 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 $f^\star : \mathcal{X} \to \mathcal{Y}$,

$$\sup_{t\in\mathbb{N}} \ell\Big(\hat{f}_n(X_{1:n}, f^\star(X_{1:n}), X_t), f^\star(X_t)\Big) \to 0 \text{ (a.s.)}. \tag{73}$$

This further implies that, for any measurable $f^\star : \mathcal{X} \to \mathcal{Y}$,

$$\lim_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}\Big(\hat{f}_n, f^\star; n\Big) = \lim_{n\to\infty} \operatorname*{limsup}_{m\to\infty} \frac{1}{m} \sum_{t=n+1}^{n+m} \ell\Big(\hat{f}_n(X_{1:n}, f^\star(X_{1:n}), X_t), f^\star(X_t)\Big)$$
$$\leq \lim_{n\to\infty} \sup_{t\in\mathbb{N}} \ell\Big(\hat{f}_n(X_{1:n}, f^\star(X_{1:n}), X_t), f^\star(X_t)\Big) = 0 \text{ (a.s.)}.$$

Thus, since $\hat{\mathcal{L}}_{\mathbb{X}}$ is non-negative, we conclude that 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, $\mathbb{X} \in \text{SUIL}$.

The above argument can also be used to show that $\mathbb{X} \in \text{SUOL}$. Specifically, consider this same $\hat{f}_n$ function defined above, but now interpreted as an *online* learning rule. We then have, for any measurable $f^\star : \mathcal{X} \to \mathcal{Y}$,

$$\lim_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}\Big(\hat{f}_\cdot, f^\star; n\Big) = \lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \ell\Big(\hat{f}_t(X_{1:t}, f^\star(X_{1:t}), X_{t+1}), f^\star(X_{t+1})\Big)$$
$$\leq \lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \sup_{m\in\mathbb{N}} \ell\Big(\hat{f}_t(X_{1:t}, f^\star(X_{1:t}), X_m), f^\star(X_m)\Big). \tag{74}$$

The convergence in (73) 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 final expression in (74) equals 0 almost surely. Since this holds for any measurable $f^\star : \mathcal{X} \to \mathcal{Y}$, and ℓ is non-negative, 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, $\mathbb{X} \in \text{SUOL}$.

Finally, since the above arguments hold for *any* choice of $\mathbb{X} \in \mathcal{C}_3$, we may conclude that $\mathcal{C}_3 \subseteq \text{SUIL} \cap \text{SUOL}$, which completes the proof. ■

Combining the above lemmas immediately provides the following proof of Theorem 50.

Proof of Theorem 50 Taking Lemmas 53, 54, and 57 together, we have that $\text{SUIL} \cup \text{SUOL} \subseteq \text{SUAL} \cup \text{SUOL} \subseteq \mathcal{C}_3 \subseteq \text{SUIL} \cap \text{SUOL} \subseteq \text{SUAL} \cap \text{SUOL}$. This further implies that $\text{SUAL} \triangle \text{SUOL} = (\text{SUAL} \cup \text{SUOL}) \setminus (\text{SUAL} \cap \text{SUOL}) = \emptyset$, and similarly $\text{SUIL} \triangle \text{SUOL} = (\text{SUIL} \cup \text{SUOL}) \setminus (\text{SUIL} \cap \text{SUOL}) = \emptyset$, so that $\text{SUIL} = \text{SUOL} = \text{SUAL}$. Combining this with Lemmas 54 and 57, we obtain $\text{SUOL} = \text{SUAL} \cup \text{SUOL} \subseteq \mathcal{C}_3 \subseteq \text{SUIL} \cap \text{SUOL} = \text{SUOL}$, so that $\text{SUOL} = \mathcal{C}_3$. Hence $\text{SUIL} = \text{SUAL} = \text{SUOL} = \mathcal{C}_3$, which completes the proof. ■

We may also note that the proof of Lemma 57 specifically establishes that the inductive learning rule $\hat{f}_n$ specified in (68) (with $\{\tilde{f}_i\}_{i=1}^\infty$ an enumeration of the countable set $\tilde{\mathcal{F}}$ from Lemma 56) is strongly universally consistent for every $\mathbb{X} \in \mathcal{C}_3$, and therefore by Theorem 50 (just established), for every $\mathbb{X} \in \text{SUIL}$ 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 $x_{1:m} \in \mathcal{X}^m$, $y_{1:n} \in \mathcal{Y}^n$, 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 57, 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 $\mathrm{SUAL} = \mathcal{C}_3$ (from Theorem 50, just established). The proof of Lemma 57 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 $\mathcal{C}_3 = \mathrm{SUOL}$ when $\bar{\ell} = \infty$, again by Theorem 50) $\hat{f}_n$ is also an optimistically universal *online* learning rule when $\bar{\ell} = \infty$. We summarize these findings in the following theorem.

Theorem 58 *When $\bar{\ell} = \infty$, with $\{\tilde{f}_i\}_{i=1}^{\infty}$ an enumeration of the countable set $\tilde{\mathcal{F}}$ from Lemma 56, the learning rule $\hat{f}_n$ from* (68) *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 51 immediately follows.

Remark: Interestingly, the proof of Lemma 57 in fact establishes a much stronger kind of convergence for $\hat{f}_n$ under any $\mathbb{X} \in \mathcal{C}_3$: for any measurable $f^\star : \mathcal{X} \to \mathcal{Y}$,

$$\sup_{t \in \mathbb{N}} \ell\Big(\hat{f}_n(X_{1:n}, f^\star(X_{1:n}), X_t), f^\star(X_t)\Big) \to 0 \text{ (a.s.)}. \tag{76}$$

Denoting by $\mathrm{SUIL}^{\mathrm{sup}}$ the set of processes $\mathbb{X}$ that admit the existence of an inductive learning rule $\hat{f}_n$ satisfying (76) for every measurable $f^\star : \mathcal{X} \to \mathcal{Y}$, we have thus established that $\mathcal{C}_3 \subseteq \mathrm{SUIL}^{\mathrm{sup}}$ when $\bar{\ell} = \infty$. Furthermore, as shown in the proof of Lemma 57, this type of convergence itself implies strong universal consistency of $\hat{f}_n$ in the original sense of Definition 1, so that $\mathrm{SUIL}^{\mathrm{sup}} \subseteq \mathrm{SUIL}$. Thus, since $\mathrm{SUIL} = \mathcal{C}_3$ when $\bar{\ell} = \infty$ (from Theorem 50, just established), we have established that, when $\bar{\ell} = \infty$, $\mathrm{SUIL}^{\mathrm{sup}} = \mathrm{SUIL}$: 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 57 with a straightforward variation on the proof of Lemma 54, 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 (76) for all measurable functions $f^\star : \mathcal{X} \to \mathcal{Y}$: that is, $\mathrm{SUIL}^{\mathrm{sup}} = \mathcal{C}_3$. 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 $f^\star : \mathcal{X} \to \mathcal{Y}$,

$$\ell\Big(\hat{h}_n(X_{1:n}, f^\star(X_{1:n}), X_{n+1}), f^\star(X_{n+1})\Big) \to 0 \text{ (a.s.)}.$$

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 59 *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 50 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 47, relative to this hypothesis test $\hat{t}_n$. The proof of Theorem 47 (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 disjoint sequence $\{A_i\}_{i=1}^\infty$ in $\mathcal{B}$,

$$|\{i \in \mathbb{N} : \mathbb{X} \cap A_i \neq \emptyset\}| = \big|\big\{i \in \mathbb{N} : X_{1:(n_{k-1}+1)} \cap A_i \neq \emptyset\big\}\big| \leq n_{k-1} + 1 < \infty.$$

Therefore, Lemma 49 implies 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 47, $\hat{t}_n(X_{1:n})$ 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 as well. ■

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 50), we have the following immediate corollary.

Corollary 60 *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. Noisy Responses

In much of the statistical learning theory literature, it is common to suppose that the response Y_t is *noisy*, so that rather than $Y_t = f^\star(X_t)$ always, we merely have that among all measurable functions $f : \mathcal{X} \to \mathcal{Y}$, the conditional expectation $\mathbb{E}[\ell(f(X_t), Y_t)|X_t]$ is minimized for $f = f^\star$. For instance, in classification with ℓ the 0-1 loss and $\mathcal{Y} = \{0, 1\}$, this corresponds to having $Y_t = f^\star(X_t)$ with a probability at least $1/2$ given X_t. In regression with $\mathcal{Y}$ an interval in $\mathbb{R}$ and ℓ the squared loss ($\ell(a, b) = (a - b)^2$), it is well known that the point-wise minimizer of $\mathbb{E}[\ell(f(X), Y)|X]$ is the *conditional mean*: $f^\star(X) = \mathbb{E}[Y|X]$ (a.s.).

It is interesting to consider how the theory developed in the sections above can be modified to accommodate noisy responses. Here we are still interested in obtaining low long-run average loss. However, in the presence of noise we generally cannot hope to achieve *zero* average loss in the limit. We must therefore adjust our goal. Instead, we will aim to achieve zero *excess* loss, relative to a fixed *optimal* function: that is, we will still suppose there is a function $f^\star$ representing an optimal predictor, and we will evaluate our performance relative to this function.

In this context, we achieve two different strengths of results. First, we show that for certain restricted types of losses ℓ, there exists a self-adaptive learning rule that is strongly universally consistent for any $(\mathbb{X}, \mathbb{Y}) = \{(X_t, Y_t)\}_{t=1}^\infty$ with $\mathbb{X} \in \mathcal{C}_1$ and $\mathbb{Y}$ satisfying a conditional independence property: that is, the Y_t variables are conditionally independent given their respective X_t variables. In particular, this result applies to the *squared loss* for $\mathcal{Y}$ any bounded interval in $\mathbb{R}$. However, it turns out classification with the 0-1 loss does not satisfy the requirements on ℓ for this result. To address classification, we propose a second, stronger condition, where we suppose Y_t is a *noisy function* of X_t: that is, the Y_t values are conditionally independent and the conditional distribution of Y_t given X_t is a t-invariant function of X_t. We show that there exists a self-adaptive learning rule that is strongly universally consistent for classification with finite $\mathcal{Y}$ and the 0-1 loss, for all processes of this type with $\mathbb{X} \in \mathcal{C}_1$. The question of learning, either with general conditionally independent $\mathbb{Y}$ or with noisy functions, for general bounded separable losses ℓ is left for future work.

9.1 Definitions

In this general setting, for any measurable $\bar{f} : \mathcal{X} \to \mathcal{Y}$, for any process $(\mathbb{X}, \mathbb{Y}) = \{(X_t, Y_t)\}_{t=1}^\infty$ on $\mathcal{X} \times \mathcal{Y}$, for any inductive learning rule f_n and self-adaptive learning rule $g_{n,m}$, and any $n \in \mathbb{N}$, we define the long-run average *excess* loss

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

$$\hat{\mathcal{L}}_{\mathbb{X}}(g_{n,\cdot}, \mathbb{Y}; n, \bar{f}) = \limsup_{t\to\infty} \frac{1}{t+1} \sum_{m=n}^{n+t} \left(\ell(g_{n,m}(X_{1:m}, Y_{1:n}, X_{m+1}), Y_{m+1}) - \ell\big(\bar{f}(X_{m+1}), Y_{m+1}\big)\right).$$

We are then interested in learning rules f_n, $g_{n,m}$ that guarantee that, for all measurable $\bar{f}$, $\limsup_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(f_n, \mathbb{Y}; n, \bar{f}) \le 0$ almost surely, or $\limsup_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(g_{n,\cdot}, \mathbb{Y}; n, \bar{f}) \le 0$ almost surely, respectively, for some specific family of processes $(\mathbb{X}, \mathbb{Y})$. For brevity, we will not discuss

the online setting in detail in this section, though an analogous generalization is possible there: that is, $\hat{\mathcal{L}}_{\mathbb{X}}(f_{\cdot}, \mathbb{Y}; \bar{f}) = \limsup_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} \left(\ell(f_t(X_{1:t}, Y_{1:t}, X_{t+1}), Y_{t+1}) - \ell\big(\bar{f}(X_{t+1}), Y_{t+1}\big)\right)$.

It is clear that some kind of restriction to the dependences among the (X_t, Y_t) variables would be required for any positive result to be possible. While the argument we follow here can also be applied in more general scenarios (within certain limits), as a simple scenario to consider we restrict to the following two key requirements.

Y1. We restrict to processes $\mathbb{Y} = \{Y_t\}_{t=1}^{\infty}$ (dependent on $\mathbb{X}$) with the property that there exists a measurable function $f^\star : \mathcal{X} \to \mathcal{Y}$ with

$$\mathbb{E}[\ell(f^\star(X_t), Y_t)|X_t] = \inf_{y\in\mathcal{Y}} \mathbb{E}[\ell(y, Y_t)|X_t] \text{ (a.s.)}$$

for every $t \in \mathbb{N}$. In other words, we assume there is a time-invariant optimal function.[11]

Y2. We suppose the Y_t variables are *conditionally independent* given their respective X_t variables. Formally, we suppose $\mathbb{Y} = \{Y_t\}_{t=1}^{\infty}$ has the property that $\forall t \in \mathbb{N}$, Y_t is conditionally independent of $\{(X_{t'}, Y_{t'})\}_{t'\neq t}$ given X_t.

In particular, note that both of these conditions would be satisfied for any i.i.d. process $(\mathbb{X}, \mathbb{Y})$ if $\mathcal{Y}$ is sequentially compact (so that the infimum exists in Y1). Henceforth in this section, mentions of $f^\star$ will always refer to the function $f^\star$ guaranteed to exist by Y1. As we argue below in Lemma 64, for any process $(\mathbb{X}, \mathbb{Y})$ satisfying Y1 and Y2, for any learning rule $f_{n,m}$, we have $\hat{\mathcal{L}}_{\mathbb{X}}(f_{n,\cdot}, \mathbb{Y}; n, f^\star) \geq 0$ (a.s.). Moreover, by similar arguments to the proof of Lemma 64, it is not hard to show that for any measurable function $\bar{f} : \mathcal{X} \to \mathcal{Y}$, we have $\hat{\mathcal{L}}_{\mathbb{X}}(f_{n,\cdot}, \mathbb{Y}; n, f^\star) \geq \hat{\mathcal{L}}_{\mathbb{X}}(f_{n,\cdot}, \mathbb{Y}; n, \bar{f})$ (a.s.). For these reasons, for $(\mathbb{X}, \mathbb{Y})$ satisfying Y1 and Y2, we say a learning rule $f_{n,m}$ is *consistent* if $\hat{\mathcal{L}}_{\mathbb{X}}(f_{n,\cdot}, \mathbb{Y}; n, f^\star) \to 0$ (a.s.).

Below we obtain two different results, corresponding to two different families of processes $(\mathbb{X}, \mathbb{Y})$. These families are stated formally in the following definitions.

Definition 61 *We say a process* $(\mathbb{X}, \mathbb{Y})$ *has* independent noise *if it satisfies properties* Y1 *and* Y2 *above. We say* $\mathbb{Y}$ *is a* noisy function *of* $\mathbb{X}$ *if* $(\mathbb{X}, \mathbb{Y})$ *has independent noise, and the conditional distribution of* Y_t *given* X_t *is a* t*-invariant function of* X_t*.*

The main difference between $(\mathbb{X}, \mathbb{Y})$ merely having independent noise and $\mathbb{Y}$ being a noisy function of $\mathbb{X}$ is that the former case admits processes where the conditional distribution of Y_t given X_t varies over time. For instance, as a simple example of a process $(\mathbb{X}, \mathbb{Y})$ having independent noise, consider $\mathcal{Y} = [-1, 1]$ and $\ell = \ell_{\text{sq}}$ the squared loss ($\ell_{\text{sq}}(a, b) = (a - b)^2$), and take $\mathbb{X}$ as any process, while $Y_t = f^\star(X_t) + \left(1 - \frac{1}{t}\right) V_t$, where $f^\star$ is any measurable function with $f^\star(x) \in [-1/2, 1/2]$ for all $x \in \mathcal{X}$, and where $\{V_t\}_{t=1}^{\infty}$ are independent (and independent of $\mathbb{X}$) with $V_t \sim \text{Uniform}([-1/2, 1/2])$. This process $\mathbb{Y}$ gets *noisier* over time, but the optimal function is always $f^\star$ (the conditional mean of Y_t given X_t being $f^\star(X_t)$), and each Y_t is conditionally independent of $\{(X_{t'}, Y_{t'})\}_{t'\neq t}$ given X_t. Note that $\mathbb{Y}$ is *not* a noisy function of $\mathbb{X}$. However, if instead we had defined $Y_t = f^\star(X_t) + V_t$, then it would be.

We now formally state the criteria for universal consistency with noise.

11. In principle, the theory below can be extended to cases where the infimum does not exist, but there exist $f^\star_\varepsilon$ functions with $\mathbb{E}[\ell(f^\star(X_t), Y_t)|X_t] \leq \inf_{y\in\mathcal{Y}} \mathbb{E}[\ell(y, Y_t)|X_t] + \varepsilon$ (a.s.). We restrict to cases where an optimal function $f^\star$ exists to simplify the exposition.

Definition 62 *For a self-adaptive learning rule $g_{n,m}$ and a process $\mathbb{X}$ on $\mathcal{X}$, we say $g_{n,m}$ is strongly universally consistent* with independent noise *under $\mathbb{X}$ if, for every process $\mathbb{Y}$ such that $(\mathbb{X},\mathbb{Y})$ has independent noise, it holds that $\lim_{n\to\infty}\hat{\mathcal{L}}_{\mathbb{X}}(g_{n,\cdot},\mathbb{Y};n,f^\star)=0$ (a.s.). Similarly, for an inductive learning rule f_n, we say f_n is strongly universally consistent with independent noise under $\mathbb{X}$ if, for every $\mathbb{Y}$ such that $(\mathbb{X},\mathbb{Y})$ has independent noise, $\lim_{n\to\infty}\hat{\mathcal{L}}_{\mathbb{X}}(f_n,\mathbb{Y};n,f^\star)=0$ (a.s.). We say a process $\mathbb{X}$ admits strong universal (inductive/self-adaptive) learning with independent noise if there exists an (inductive/self-adaptive) learning rule that is strongly universally consistent with independent noise under $\mathbb{X}$. We say a self-adaptive learning rule $g_{n,m}$ is optimistically universal with independent noise if it is strongly universally consistent with independent noise under every $\mathbb{X}$ that admits strong universal self-adaptive learning with independent noise.*

Definition 63 *For a self-adaptive learning rule $g_{n,m}$ and a process $\mathbb{X}$ on $\mathcal{X}$, we say $g_{n,m}$ is strongly universally consistent* for noisy functions *under $\mathbb{X}$ if, for every process $\mathbb{Y}$ that is a noisy function of $\mathbb{X}$, it holds that $\lim_{n\to\infty}\hat{\mathcal{L}}_{\mathbb{X}}(g_{n,\cdot},\mathbb{Y};n,f^\star)=0$ (a.s.). Similarly, for an inductive learning rule f_n, we say f_n is strongly universally consistent for noisy functions under $\mathbb{X}$ if, for every $\mathbb{Y}$ that is a noisy function of $\mathbb{X}$, $\lim_{n\to\infty}\hat{\mathcal{L}}_{\mathbb{X}}(f_n,\mathbb{Y};n,f^\star)=0$ (a.s.). We say a process $\mathbb{X}$ admits strong universal (inductive/self-adaptive) learning for noisy functions if there exists an (inductive/self-adaptive) learning rule that is strongly universally consistent for noisy functions under $\mathbb{X}$. We say a self-adaptive learning rule $g_{n,m}$ is optimistically universal for noisy functions if it is strongly universally consistent for noisy functions under every $\mathbb{X}$ that admits strong universal self-adaptive learning for noisy functions.*

In particular, since any i.i.d. process $(\mathbb{X},\mathbb{Y})$ would have $\mathbb{Y}$ a noisy function of $\mathbb{X}$ (if $\mathcal{Y}$ is sequentially compact, so that the infimum exists in Y1), we note that the theory we develop here represents a proper generalization of the standard theory of universal consistency under i.i.d. processes with bounded losses (e.g., Devroye, Györfi, and Lugosi, 1996; Györfi, Kohler, Krzyżak, and Walk, 2002).

In Section 9.2 below, we show that for certain restricted types of losses (essentially, those guaranteeing uniqueness of $f^\star$), any process $\mathbb{X}$ satisfying Condition 1 admits strong universal inductive and self-adaptive learning with independent noise. In other words, a process $\mathbb{X}$ admits strong universal learning with independent noise if (and only if) it admits strong universal learning *without* noise. In particular, this includes the squared loss ($\ell(a,b)=(a-b)^2$) for $\mathcal{Y}$ any bounded interval in $\mathbb{R}$. We also argue that there is a self-adaptive learning rule that is optimistically universal with independent noise. Thus, the theory developed above for the noiseless setting completely generalizes to allow independent noise, for these loss functions. However, it happens that the 0-1 loss does not satisfy the requirements for this result. As this is an important loss for the classification setting, in Section 9.3 we extend the theory to hold for learning with the 0-1 loss ($\ell(a,b)=\mathbb{1}[a\neq b]$) for any finite $\mathcal{Y}$, but only for the stronger *noisy function* setting. Specifically, we show that for this loss, any process $\mathbb{X}$ satisfying Condition 1 admits strong universal inductive and self-adaptive learning for noisy functions. We also show that there is a self-adaptive learning rule that is optimistically universal for noisy functions. Again, this result generalizes the results for the noiseless setting to the setting of noisy functions. The approach to obtaining this result is via the traditional *plug-in* technique (see e.g., Devroye, Györfi, and Lugosi, 1996),

making use of consistent regression estimators (guaranteed to exist by the aforementioned result for learning with independent noise) to identify which element in $\mathcal{Y}$ has the highest likelihood given X_t.

Before proceeding with our results for learning with independent noise, we first discuss the motivation for the special role of $f^\star$ in the definition of universal consistency above. This is motivated by the fact that, in any process $(\mathbb{X}, \mathbb{Y})$ that has independent noise, $f^\star$ is guaranteed to be an *optimal* function (almost surely), so that no prediction rule can be *better* than $f^\star$ in the limit. Formally, we have the following lemma.

Lemma 64 *If $\bar{\ell} < \infty$, for any deterministic self-adaptive learning rule $\hat{f}_{n,m}$, for any process $(\mathbb{X}, \mathbb{Y})$ that has independent noise, with probability one, $\forall n \in \mathbb{N}$, $\hat{\mathcal{L}}_{\mathbb{X}}(\hat{f}_{n,\cdot}, \mathbb{Y}; n, f^\star) \geq 0$.*

Proof For any $n, m \in \mathbb{N}$ with $m \geq n$, let $g_{n,m}(X_{m+1}) = \hat{f}_{n,m}(X_{1:m}, Y_{1:n}, X_{m+1})$, and define $\Delta^n_{m+1} = \ell(g_{n,m}(X_{m+1}), Y_{m+1}) - \ell(f^\star(X_{m+1}), Y_{m+1})$. For any $n \in \mathbb{N}$ note that, by the conditional independence property Y2, the sequence

$$\left\{\left(\Delta^n_{m+1} - \mathbb{E}\left[\Delta^n_{m+1} \middle| X_{m+1}, g_{n,m}(X_{m+1})\right]\right)\right\}_{m=n}^{\infty}$$

is a martingale difference sequence with respect to $\left\{\left(X_{1:(m+2)}, Y_{1:(m+1)}\right)\right\}_{m=n}^{\infty}$. Therefore, Azuma's inequality (Devroye, Györfi, and Lugosi, 1996, Theorem 9.1) implies that, for any $t \in \mathbb{N} \cup \{0\}$, with probability at least $1 - \frac{1}{(t+1)^2}$,

$$\frac{1}{t+1}\left|\sum_{m=n}^{n+t} \Delta^n_{m+1} - \mathbb{E}\left[\Delta^n_{m+1} \middle| X_{m+1}, g_{n,m}(X_{m+1})\right]\right| \leq \bar{\ell}\sqrt{\frac{2\ln(2(t+1)^2)}{t+1}}.$$

Since $\sum_{t=0}^{\infty} \frac{1}{(t+1)^2} < \infty$, the Borel-Cantelli lemma implies that, with probability one,

$$\limsup_{t\to\infty} \frac{1}{t+1}\left|\sum_{m=n}^{n+t} \Delta^n_{m+1} - \mathbb{E}\left[\Delta^n_{m+1} \middle| X_{m+1}, g_{n,m}(X_{m+1})\right]\right| = 0.$$

By the definition of $f^\star$ from Y1, and the conditional independence property Y2, for any $m \geq n$, $\mathbb{E}\left[\Delta^n_{m+1} \middle| X_{m+1}, g_{n,m}(X_{m+1})\right] \geq 0$ almost surely. Altogether, by the union bound, on an event of probability one, we have

$$\hat{\mathcal{L}}_{\mathbb{X}}(\hat{f}_{n,\cdot}, \mathbb{Y}; n, f^\star) = \limsup_{t\to\infty} \frac{1}{t+1} \sum_{m=n}^{n+t} \mathbb{E}\left[\Delta^n_{m+1} \middle| X_{m+1}, g_{n,m}(X_{m+1})\right] \geq 0.$$

Since this holds for any fixed $n \in \mathbb{N}$, the lemma follows by the union bound over all $n \in \mathbb{N}$. ■

9.2 Learning with Independent Noise

We begin with the general setting of learning with independent noise. In this subsection we will restrict to the case of bounded losses ($\bar{\ell} < \infty$), with ℓ satisfying some special properties. Specifically, we suppose that there exist functions $\underline{\phi}$ and $\overline{\phi}$ mapping $[0, \bar{\ell}] \to [0, \infty)$, strictly

increasing and continuous, with $\underline{\phi}$ convex and $\overline{\phi}$ concave, such that $\underline{\phi}(0) = \overline{\phi}(0) = 0$ and for any $\mathcal{Y}$-valued random variable Y, $\exists y^* \in \mathcal{Y}$ with $\mathbb{E}[\ell(y^*, Y)] = \inf_{y \in \mathcal{Y}} \mathbb{E}[\ell(y, Y)]$ (i.e., the infimum is realized in $\mathcal{Y}$), and $\forall y \in \mathcal{Y}$,

$$\underline{\phi}(\ell(y, y^*)) \leq \mathbb{E}[\ell(y, Y) - \ell(y^*, Y)] \leq \overline{\phi}(\ell(y, y^*)). \tag{77}$$

As a simple example of this, consider the case of bounded regression with the squared loss: $\mathcal{Y} = [0, 1]$ and $\ell = \ell_{\text{sq}}$ (where $\ell_{\text{sq}}(y, y') = (y - y')^2$ is the *squared loss*). In this case, as mentioned above, it is well known that $y^* = \mathbb{E}[Y]$ uniquely, and that for any $y \in [0, 1]$, $\mathbb{E}[\ell(y, Y) - \ell(y^*, Y)] = \ell(y, y^*)$, since for any $[0, 1]$-valued random variable Y and for $y^* = \mathbb{E}[Y]$, it holds that $\mathbb{E}[(Y - y)^2] - \mathbb{E}[(Y - y^*)^2] = y^2 - 2y\mathbb{E}[Y] + 2y^*\mathbb{E}[Y] - (y^*)^2 = (y - y^*)^2$. Thus, for $(\mathcal{Y}, \ell) = ([0, 1], \ell_{\text{sq}})$, the above condition holds with $\underline{\phi}(x) = \overline{\phi}(x) = x$. As we discuss below, the results also have implications for the classification setting via the well-known *plug-in* technique (Devroye, Györfi, and Lugosi, 1996).

For losses satisfying (77), we will argue that the following specifications yield learning rules that are strongly universally consistent with independent noise. First we specify an inductive learning rule. Let $\{\mathcal{G}_i\}_{i=1}^{\infty}$, $\{m_i\}_{i=1}^{\infty}$, and $\{i_n\}_{i=1}^{\infty}$ be as in the proof of Lemma 27, and note that without loss of generality we can suppose m_i is strictly increasing (since, for the γ_i values from the proof of Lemma 27, the sequence $i'_n = \min\{i : m_i = m_{i_n}\}$ can be used in place of i_n in Lemma 22 while still retaining the guarantee in the lemma, and while still satisfying $i'_n \to \infty$), and that $|\mathcal{G}_i| = i$ for all $i \in \mathbb{N}$ (and indeed, this is the case in the construction given in Lemma 25). Then for any $n \in \mathbb{N}$, $x_{1:n} \in \mathcal{X}^n$, and $y_{1:n} \in \mathcal{Y}^n$, for each $s \in \{m_{i_n}, \ldots, n\}$ define the function $\tilde{f}_{n,s}(x_{1:n}, y_{1:n}, \cdot)$ as

$$\operatorname*{argmin}_{f \in \mathcal{G}_{i_n}} \frac{1}{s} \sum_{t=1}^{s} \ell(f(x_t), y_t).$$

Then define the function $\hat{f}_n(x_{1:n}, y_{1:n}, \cdot)$ as

$$\operatorname*{argmin}_{f \in \mathcal{G}_{i_n}} \max_{m_{i_n} \leq s \leq n} \frac{1}{s} \sum_{t=1}^{s} \ell(f(x_t), \tilde{f}_{n,s}(x_{1:n}, y_{1:n}, x_t)). \tag{78}$$

We can specify a self-adaptive learning rule analogously, as follows. Let $\{\mathcal{F}_i\}_{i=1}^{\infty}$, $\{\gamma_i\}_{i=1}^{\infty}$, and $\{u_i\}_{i=1}^{\infty}$ be as in (33), with u_i strictly increasing here, and note that without loss of generality we can suppose $|\mathcal{F}_i| = i$. Define $\hat{i}_{n,m}$ as in (33). Then for any $n, m \in \mathbb{N}$ ($m \geq n$), $x_{1:m} \in \mathcal{X}^m$, and $y_{1:n} \in \mathcal{Y}^n$, for each $s \in \{u_{\hat{i}_{n,m}(x_{1:m})}, \ldots, n\}$ define the function $\tilde{f}_{n,m,s}(x_{1:m}, y_{1:n}, \cdot)$ as

$$\operatorname*{argmin}_{f \in \mathcal{F}_{\hat{i}_{n,m}(x_{1:m})}} \frac{1}{s} \sum_{t=1}^{s} \ell(f(x_t), y_t).$$

Then define the function $\hat{f}_{n,m}(x_{1:m}, y_{1:n}, \cdot)$ as

$$\operatorname*{argmin}_{f \in \mathcal{F}_{\hat{i}_{n,m}(x_{1:m})}} \max_{u_{\hat{i}_{n,m}(x_{1:m})} \leq s \leq n} \frac{1}{s} \sum_{t=1}^{s} \ell(f(x_t), \tilde{f}_{n,m,s}(x_{1:m}, y_{1:n}, x_t)). \tag{79}$$

One can easily verify that the choices in these optimizations can be made in such a way that the functions $\hat{f}_n$ and $\hat{f}_{n,m}$ are measurable (e.g., by breaking ties based on a fixed predefined ordering of each $\mathcal{F}_i$ set); for simplicity, we will suppose ties in the argmin are broken deterministically, so that the learning rules are deterministic functions.

We have the following theorem, which reveals that a process $\mathbb{X}$ admits strong universal (inductive/self-adaptive) learning with independent noise if and only if it admits strong universal (inductive/self-adaptive) learning *without* noise. Furthermore, the above learning rules witness the sufficiency of these conditions, which further implies that the self-adaptive learning rule (79) is optimistically universal with independent noise.

Theorem 65 *If ℓ satisfies* (77)*, then Condition 1 is necessary and sufficient for a process $\mathbb{X}$ to admit strong universal (inductive/self-adaptive) learning with independent noise. Moreover, the self-adaptive learning rule $\hat{f}_{n,m}$ defined by* (79) *is optimistically universal with independent noise.*

The restrictions to ℓ guaranteeing existence of $\underline{\phi}$ and $\overline{\phi}$ provide an important convenience for us: namely, under these conditions, we can still characterize consistency as convergence to $f^\star$. More specifically, we have the following guarantee, which will be a key component in the proof of Theorem 65.

Lemma 66 *If $\mathbb{X} \in \mathcal{C}_1$, $(\mathbb{X}, \mathbb{Y})$ has independent noise, ℓ satisfies* (77)*, and $\hat{f}_n$ and $\hat{f}_{n,m}$ are as in* (78) *and* (79) *respectively, then*

$$\limsup_{n\to\infty} \hat{\mu}_{\mathbb{X}}(\ell(\hat{f}_n(X_{1:n}, Y_{1:n}, \cdot), f^\star(\cdot))) = 0 \text{ (a.s.)}$$

and

$$\limsup_{n\to\infty} \limsup_{t\to\infty} \frac{1}{t+1} \sum_{m=n}^{n+t} \ell(\hat{f}_{n,m}(X_{1:m}, Y_{1:n}, X_{m+1}), f^\star(X_{m+1})) = 0 \text{ (a.s.)}.$$

Proof For brevity, we only give the detailed proof of the claim for the self-adaptive rule $\hat{f}_{n,m}$. The proof of the claim for the inductive learning rule $\hat{f}_n$ follows analogously, merely substituting properties of the sequence i_n established in the proof of Lemma 27, in place of the analogous properties of $\hat{i}_n$ used here. We provide an outline of that analogous argument following the detailed proof for the self-adaptive learning rule $\hat{f}_{n,m}$, which we now turn to.

To simplify notation, let $\hat{g}_{n,m}(\cdot) = \hat{f}_{n,m}(X_{1:m}, Y_{1:n}, \cdot)$ and $\tilde{g}_{n,m,s}(\cdot) = \tilde{f}_{n,m,s}(X_{1:m}, Y_{1:n}, \cdot)$. Let m_n^*, $\hat{i}_n$, $f_i^\star$, α_i, ι_0, and the event K all be as in the proof of Theorem 29 (defined relative to the fixed function $f^\star$ from property Y1), and define $\hat{g}_n = \hat{g}_{n,m_n^*}$.

By the conditional independence property Y2, Hoeffding's inequality (applied under the conditional distribution given $\mathbb{X}$) and the law of total probability imply that, for any $i, s \in \mathbb{N}$ and $f \in \mathcal{F}_i$, with probability at least $1 - \frac{1}{i^3 s^2}$,

$$\left| \frac{1}{s} \sum_{t=1}^{s} (\ell(f(X_t), Y_t) - \ell(f^\star(X_t), Y_t)) - \mathbb{E}[\ell(f(X_t), Y_t) - \ell(f^\star(X_t), Y_t) | X_t] \right| \leq 2\bar{\ell} \sqrt{\frac{\ln(2i^3 s^2)}{2s}}. \tag{80}$$

By the union bound, for any fixed $i \in \mathbb{N}$, the inequality (80) holds simultaneously for all $f \in \mathcal{F}_i$ and $s \in \mathbb{N}$ with probability at least $1 - \sum_{s=1}^{\infty} |\mathcal{F}_i| \frac{1}{i^3 s^2} = 1 - \frac{\pi^2}{6i^2}$ (since $|\mathcal{F}_i| = i$).

Furthermore, since $\sum_{i=1}^{\infty} \frac{\pi^2}{6i^2} < \infty$, the Borel-Cantelli lemma implies that there is an event $\tilde{K}$ of probability one, on which $\exists \iota_1 \in \mathbb{N}$ such that (80) holds simultaneously for every $i \geq \iota_1$ and every $s \in \mathbb{N}$.

Now suppose the event $K \cap \tilde{K}$ occurs. Recalling that $\lim_{n\to\infty} \hat{i}_n = \infty$ by (36), let $\tilde{\nu} \in \mathbb{N}$ be such that $\forall n \geq \tilde{\nu}$ it holds that $\hat{i}_n \geq \max\{\iota_0, \iota_1\}$, so that both (80) (for $i = \hat{i}_n$, and for all s) and (37) hold. Then for any $n \geq \tilde{\nu}$ and any $s \in \{u_{\hat{i}_n}, \ldots, n\}$, for any $m \geq m_n^*$ we have

$$\begin{aligned}
&\frac{1}{s}\sum_{t=1}^{s} \ell(\tilde{g}_{n,m,s}(X_t), Y_t) - \ell(f^\star(X_t), Y_t) \\
&\leq \frac{1}{s}\sum_{t=1}^{s} \ell(f^\star_{\hat{i}_n}(X_t), Y_t) - \ell(f^\star(X_t), Y_t) \\
&\leq 2\bar{\ell}\sqrt{\frac{\ln(2\hat{i}_n^3 s^2)}{2s}} + \frac{1}{s}\sum_{t=1}^{s} \mathbb{E}\Big[\ell(f^\star_{\hat{i}_n}(X_t), Y_t) - \ell(f^\star(X_t), Y_t)\Big| X_t, \hat{i}_n\Big] \\
&\leq 2\bar{\ell}\sqrt{\frac{\ln(2\hat{i}_n^3 s^2)}{2s}} + \frac{1}{s}\sum_{t=1}^{s} \overline{\phi}\Big(\ell(f^\star_{\hat{i}_n}(X_t), f^\star(X_t))\Big) \\
&\leq 2\bar{\ell}\sqrt{\frac{\ln(2\hat{i}_n^3 s^2)}{2s}} + \overline{\phi}\left(\frac{1}{s}\sum_{t=1}^{s} \ell(f^\star_{\hat{i}_n}(X_t), f^\star(X_t))\right) \\
&\leq 2\bar{\ell}\sqrt{\frac{\ln(2\hat{i}_n^3 s^2)}{2s}} + \overline{\phi}(\alpha_{\hat{i}_n}) \leq 2\bar{\ell}\sqrt{\frac{\ln(2\hat{i}_n^5)}{2\hat{i}_n}} + \overline{\phi}(\alpha_{\hat{i}_n}),
\end{aligned}$$

where the last four inequalities are due to (77), Jensen's inequality (due to concavity of $\overline{\phi}$), (37) and monotonicity of $\overline{\phi}$, and the fact that u_i is strictly increasing (so that $s \geq u_{\hat{i}_n} \geq \hat{i}_n$).

Now denote by $\{Y'_t\}_{t=1}^{\infty}$ a sequence with the same conditional distribution given $\mathbb{X}$ as $\mathbb{Y}$, but conditionally independent of $\mathbb{Y}$ given $\mathbb{X}$. Again by (80), if $n \geq \tilde{\nu}$ and $m \geq m_n^*$, every $s \in \{u_{\hat{i}_n}, \ldots, n\}$ has

$$\begin{aligned}
&\frac{1}{s}\sum_{t=1}^{s} \ell(\tilde{g}_{n,m,s}(X_t), Y_t) - \ell(f^\star(X_t), Y_t) \\
&\geq -2\bar{\ell}\sqrt{\frac{\ln(2\hat{i}_n^3 s^2)}{2s}} + \frac{1}{s}\sum_{t=1}^{s} \mathbb{E}\big[\ell(\tilde{g}_{n,m,s}(X_t), Y'_t) - \ell(f^\star(X_t), Y'_t)\big| X_t, \tilde{g}_{n,m,s}\big] \\
&\geq -2\bar{\ell}\sqrt{\frac{\ln(2\hat{i}_n^5)}{2\hat{i}_n}} + \frac{1}{s}\sum_{t=1}^{s} \underline{\phi}(\ell(\tilde{g}_{n,m,s}(X_t), f^\star(X_t))) \\
&\geq -2\bar{\ell}\sqrt{\frac{\ln(2\hat{i}_n^5)}{2\hat{i}_n}} + \underline{\phi}\left(\frac{1}{s}\sum_{t=1}^{s} \ell(\tilde{g}_{n,m,s}(X_t), f^\star(X_t))\right),
\end{aligned}$$

where the last two inequalities use that u_i is strictly increasing (so that $s \geq u_{\hat{i}_n} \geq \hat{i}_n$) and (77) along with Jensen's inequality (due to convexity of $\underline{\phi}$).

Define

$$\beta_i = \underline{\phi}^{-1}\left(4\bar{\ell}\sqrt{\frac{\ln(2i^5)}{2i}} + \overline{\phi}(\alpha_i)\right),$$

noting that the inverse function $\underline{\phi}^{-1}$ is well-defined due to $\underline{\phi}$ being continuous and strictly increasing. Note that $\lim_{n\to\infty} \beta_{\hat{i}_n} = 0$ by the facts that $\underline{\phi}$ and $\overline{\phi}$ are continuous with $\underline{\phi}(0) = \overline{\phi}(0) = 0$, together with the facts that $\alpha_i \to 0$ and $\hat{i}_n \to \infty$ (as established in the proof of Theorem 29). Altogether, we have established that, on $K \cap \tilde{K}$, every $n \geq \tilde{\nu}$, $m \geq m_n^*$, and $s \in \{u_{\hat{i}_n}, \ldots, n\}$ satisfy

$$\frac{1}{s}\sum_{t=1}^{s} \ell(\tilde{g}_{n,m,s}(X_t), f^\star(X_t)) \leq \beta_{\hat{i}_n}. \tag{81}$$

Now to relate this performance guarantee for these $\tilde{g}_{n,m,s}$ functions on the first n data points to performance of $\hat{g}_{n,m}$ on the full sequence, note that (just as in the proof of Theorem 29) since $\hat{g}_{n,m} = \hat{g}_n$ for all $m \geq m_n^*$, we have

$$\begin{aligned}
&\operatorname*{limsup}_{s\to\infty} \frac{1}{s+1}\sum_{m=n}^{n+s} \ell(\hat{g}_{n,m}(X_{m+1}), f^\star(X_{m+1})) \\
&= \operatorname*{limsup}_{s\to\infty} \frac{1}{s}\sum_{t=1}^{s} \ell(\hat{g}_n(X_t), f^\star(X_t)) \leq \sup_{u_{\hat{i}_n} \leq s < \infty} \frac{1}{s}\sum_{t=1}^{s} \ell(\hat{g}_n(X_t), f^\star(X_t)).
\end{aligned}$$

Again supposing $K \cap \tilde{K}$ holds and that $n \geq \tilde{\nu}$, by the relaxed triangle inequality we have

$$\begin{aligned}
&\sup_{u_{\hat{i}_n} \leq s < \infty} \frac{1}{s}\sum_{t=1}^{s} \ell(\hat{g}_n(X_t), f^\star(X_t)) \\
&\leq c_\ell\left(\sup_{u_{\hat{i}_n} \leq s < \infty} \frac{1}{s}\sum_{t=1}^{s} \ell(\hat{g}_n(X_t), f^\star_{\hat{i}_n}(X_t))\right) + c_\ell\left(\sup_{u_{\hat{i}_n} \leq s < \infty} \frac{1}{s}\sum_{t=1}^{s} \ell(f^\star_{\hat{i}_n}(X_t), f^\star(X_t))\right) \\
&\leq c_\ell\left(\max_{u_{\hat{i}_n} \leq s \leq n} \frac{1}{s}\sum_{t=1}^{s} \ell(\hat{g}_n(X_t), f^\star_{\hat{i}_n}(X_t))\right) + c_\ell(\gamma_{\hat{i}_n} + \alpha_{\hat{i}_n}),
\end{aligned}$$

where the inequality in the last line is due to (35) and (37). By the relaxed triangle inequality again, we have

$$\begin{aligned}
&\max_{u_{\hat{i}_n} \leq s \leq n} \frac{1}{s}\sum_{t=1}^{s} \ell(\hat{g}_n(X_t), f^\star_{\hat{i}_n}(X_t)) \\
&\leq c_\ell\left(\max_{u_{\hat{i}_n} \leq s \leq n} \frac{1}{s}\sum_{t=1}^{s} \ell(f^\star_{\hat{i}_n}(X_t), \tilde{g}_{n,m,s}(X_t))\right) + c_\ell\left(\max_{u_{\hat{i}_n} \leq s \leq n} \frac{1}{s}\sum_{t=1}^{s} \ell(\hat{g}_n(X_t), \tilde{g}_{n,m,s}(X_t))\right),
\end{aligned}$$

and since $\hat{i}_{n,m}(X_{1:m}) = \hat{i}_n$ and $\hat{g}_{n,m} = \hat{g}_n$ for $m \geq m_n^*$, the definition of $\hat{f}_{n,m}$ from (79) implies that in this case this last line is at most

$$2c_\ell\left(\max_{u_{\hat{i}_n} \leq s \leq n} \frac{1}{s}\sum_{t=1}^{s} \ell(f^\star_{\hat{i}_n}(X_t), \tilde{g}_{n,m,s}(X_t))\right).$$

Furthermore, the relaxed triangle inequality implies

$$\begin{aligned}
&\max_{u_{\hat{i}_n} \le s \le n} \frac{1}{s} \sum_{t=1}^{s} \ell(f^\star_{\hat{i}_n}(X_t), \tilde{g}_{n,m,s}(X_t)) \\
&\le c_\ell \left(\max_{u_{\hat{i}_n} \le s \le n} \frac{1}{s} \sum_{t=1}^{s} \ell(f^\star_{\hat{i}_n}(X_t), f^\star(X_t)) \right) + c_\ell \left(\max_{u_{\hat{i}_n} \le s \le n} \frac{1}{s} \sum_{t=1}^{s} \ell(\tilde{g}_{n,m,s}(X_t), f^\star(X_t)) \right) \\
&\le c_\ell(\alpha_{\hat{i}_n} + \beta_{\hat{i}_n}),
\end{aligned}$$

where the last inequality is due to (37) and (81). Altogether, on $K \cap \tilde{K}$, any $n \ge \tilde{\nu}$ has

$$\limsup_{s\to\infty} \frac{1}{s+1} \sum_{m=n}^{n+s} \ell(\hat{g}_{n,m}(X_{m+1}), f^\star(X_{m+1})) \le 2c_\ell^3(\alpha_{\hat{i}_n} + \beta_{\hat{i}_n}) + c_\ell(\gamma_{\hat{i}_n} + \alpha_{\hat{i}_n}).$$

Thus, since $\alpha_{\hat{i}_n} \to 0$, $\gamma_{\hat{i}_n} \to 0$, and $\beta_{\hat{i}_n} \to 0$ (as established above), and the event $K \cap \tilde{K}$ holds with probability one (by the union bound), the claim for the self-adaptive learning rule $\hat{f}_{n,m}$ in the statement of the lemma follows.

The claim for the inductive learning rule $\hat{f}_n$ follows by a very similar argument, except replacing $\hat{i}_n$ above with the quantity i_n from the proof of Lemma 27. For brevity, we only give an outline here to illustrate the key steps, leaving the details as an exercise for the interested reader. For this argument, we take the definitions of $f^\star_i$, α_i, and γ_i as in the proof of Lemma 27. Note that an event identical to $\tilde{K}$ now holds for the sets $\mathcal{G}_i$. Defining $\hat{h}_n(\cdot) = \hat{f}_n(X_{1:n}, Y_{1:n}, \cdot)$ and $\tilde{h}_{n,s}(\cdot) = \tilde{f}_{n,s}(X_{1:n}, Y_{1:n}, \cdot)$, and following the same reasoning as above (using the analogous results from the proof of Lemma 27) we have that with probability one, for every sufficiently large n, every $s \in \{m_{i_n}, \ldots, n\}$ satisfies $\frac{1}{s}\sum_{t=1}^{s} \ell(\tilde{h}_{n,s}(X_t), f^\star(X_t)) \le \beta_{i_n}$. Continuing to follow the same arguments as above, but now using (27), we then have with probability one, for all sufficiently large n,

$$\begin{aligned}
\hat{\mu}_{\mathbb{X}}\Big(\ell\Big(\hat{h}_n(\cdot), f^\star(\cdot)\Big)\Big) &\le c_\ell \max_{m_{i_n} \le s \le n} \frac{1}{s} \sum_{t=1}^{s} \ell(\hat{h}_n(X_t), f^\star_{i_n}(X_t)) + c_\ell\left(\sqrt{\gamma_{i_n}} + \alpha_{i_n}\right) \\
&\le 2c_\ell^2 \max_{m_{i_n} \le s \le n} \frac{1}{s} \sum_{t=1}^{s} \ell(f^\star_{i_n}(X_t), \tilde{h}_{n,s}(X_t)) + c_\ell\left(\sqrt{\gamma_{i_n}} + \alpha_{i_n}\right) \\
&\le 2c_\ell^3(\alpha_{i_n} + \beta_{i_n}) + c_\ell\left(\sqrt{\gamma_{i_n}} + \alpha_{i_n}\right),
\end{aligned}$$

which converges to 0 as $n \to \infty$. ∎

To complete the proof of the claims for $\hat{f}_n$ and $\hat{f}_{n,m}$ in Theorem 65, we will compose the above result with the following general lemma.

Lemma 67 *Fix any process $\mathbb{X}$. If ℓ satisfies (77), then for any deterministic self-adaptive learning rule $f_{n,m}$, if, for every $\mathbb{Y}$ such that $(\mathbb{X}, \mathbb{Y})$ has independent noise,*

$$\limsup_{n\to\infty} \limsup_{t\to\infty} \frac{1}{t+1} \sum_{m=n}^{n+t} \ell(f_{n,m}(X_{1:m}, Y_{1:n}, X_{m+1}), f^\star(X_{m+1})) = 0 \text{ (a.s.)},$$

then $f_{n,m}$ is strongly universally consistent with independent noise under $\mathbb{X}$.

Proof For any $n, m \in \mathbb{N}$ with $m \geq n$, define $g_{n,m}(x) = f_{n,m}(X_{1:m}, Y_{1:n}, x)$ and $\Delta_{m+1}^n = \ell(g_{n,m}(X_{m+1}), Y_{m+1}) - \ell(f^\star(X_{m+1}), Y_{m+1})$. By the assumed properties of the loss ℓ and the conditional independence property Y2, every $m \geq n$ has

$$\mathbb{E}\big[\Delta_{m+1}^n \big| X_{m+1}, g_{n,m}(X_{m+1})\big] \leq \overline{\phi}(\ell(g_{n,m}(X_{m+1}), f^\star(X_{m+1}))) \,.$$

Then note that, due to the conditional independence property Y2, for any $n \in \mathbb{N}$,

$$\left\{ \left(\Delta_{m+1}^n - \mathbb{E}\big[\Delta_{m+1}^n \big| X_{m+1}, g_{n,m}(X_{m+1})\big] \right) \right\}_{m=n}^{\infty}$$

is a martingale difference sequence with respect to $\{(X_{1:(m+2)}, Y_{1:(m+1)})\}_{m=n}^{\infty}$. Therefore, Azuma's inequality (e.g., Devroye, Györfi, and Lugosi, 1996, Theorem 9.1) implies that, for any $t \in \mathbb{N} \cup \{0\}$, with probability at least $1 - \frac{1}{n^2(t+1)^2}$,

$$\frac{1}{t+1} \left| \sum_{m=n}^{n+t} \Delta_{m+1}^n - \mathbb{E}\big[\Delta_{m+1}^n \big| X_{m+1}, g_{n,m}(X_{m+1})\big] \right| \leq \bar{\ell} \sqrt{\frac{2\ln(2n^2(t+1)^2)}{(t+1)}}.$$

Since $\sum_{n=1}^{\infty} \sum_{t=0}^{\infty} \frac{1}{n^2(t+1)^2} < \infty$, the Borel-Cantelli lemma implies that, on an event E_1 of probability one,

$$\limsup_{n\to\infty} \limsup_{t\to\infty} \frac{1}{t+1} \left| \sum_{m=n}^{n+t} \Delta_{m+1}^n - \mathbb{E}\big[\Delta_{m+1}^n \big| X_{m+1}, g_{n,m}(X_{m+1})\big] \right| = 0,$$

so that

$$\limsup_{n\to\infty} \limsup_{t\to\infty} \frac{1}{t+1} \sum_{m=n}^{n+t} \Delta_{m+1}^n \leq \limsup_{n\to\infty} \limsup_{t\to\infty} \frac{1}{t+1} \sum_{m=n}^{n+t} \overline{\phi}(\ell(g_{n,m}(X_{m+1}), f^\star(X_{m+1}))) \,.$$

By Jensen's inequality, the right hand side above is at most

$$\limsup_{n\to\infty} \limsup_{t\to\infty} \overline{\phi}\left(\frac{1}{t+1} \sum_{m=n}^{n+t} \ell(g_{n,m}(X_{m+1}), f^\star(X_{m+1})) \right),$$

and since $\overline{\phi}$ is continuous and strictly increasing, and its argument is bounded, this expression equals

$$\overline{\phi}\left(\limsup_{n\to\infty} \limsup_{t\to\infty} \frac{1}{t+1} \sum_{m=n}^{n+t} \ell(g_{n,m}(X_{m+1}), f^\star(X_{m+1})) \right).$$

By the assumed property of $f_{n,m}$ in the statement of the lemma, and the fact that $\overline{\phi}(0) = 0$, this last expression equals 0 on an event E_2 of probability one. Thus, on the event $E_1 \cap E_2$, we have $\limsup_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(f_{n,\cdot}, \mathbb{Y}; n, f^\star) \leq 0$. Also recall that Lemma 64 implies that, on an event E_3 of probability one, $\liminf_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(f_{n,\cdot}, \mathbb{Y}; n, f^\star) \geq 0$. Altogether, $\hat{\mathcal{L}}_{\mathbb{X}}(f_{n,\cdot}, \mathbb{Y}; n, f^\star) \to 0$ on the event $E_1 \cap E_2 \cap E_3$, which has probability one by the union bound. ■

We may also note that, since any inductive learning rule f_n can be interpreted as a self-adaptive learning rule that simply ignores the additional data $X_{(n+1):m}$, Lemma 67 has the further implication that, if ℓ satisfies (77), then for any deterministic inductive learning rule f_n, if, for every $\mathbb{Y}$ such that $(\mathbb{X}, \mathbb{Y})$ has independent noise,

$$\limsup_{n\to\infty} \hat{\mu}_{\mathbb{X}}(\ell(f_n(X_{1:n}, Y_{1:n}, \cdot), f^\star(\cdot))) = 0 \text{ (a.s.)},$$

then f_n is strongly universally consistent with independent noise under $\mathbb{X}$.

With the above two lemmas in hand, we are ready for the proof of Theorem 65.

Proof of Theorem 65 It follows immediately from Theorem 7 that Condition 1 is a necessary condition for $\mathbb{X}$ to admit strong universal learning (either inductive or self-adaptive) with independent noise, since the noise-free case is a special case of the stated conditions. Furthermore, Lemmas 66 and 67 together imply that Condition 1 is sufficient for the rules $\hat{f}_n$ and $\hat{f}_{n,m}$ to be strongly universally consistent with independent noise, and therefore also sufficient for $\mathbb{X}$ to admit strong universal (inductive/self-adaptive) learning with independent noise. Finally, note that the self-adaptive learning rule $\hat{f}_{n,m}$ has no direct dependence on the distribution of $\mathbb{X}$, aside from the data supplied as its arguments, and yet (as just established) is strongly universally consistent with independent noise under every $\mathbb{X}$ satisfying Condition 1. Together with the fact (also just established) that Condition 1 is necessary and sufficient for $\mathbb{X}$ to admit strong universal self-adaptive learning with independent noise, this also establishes the claim that $\hat{f}_{n,m}$ is optimistically universal with independent noise. ■

In particular, this implies (79) is optimistically universal with independent noise for the special case of *regression*, where $\mathcal{Y}$ is any bounded interval of $\mathbb{R}$ and ℓ is the squared loss: $\ell_{\text{sq}}(y, y') = (y - y')^2$. However, since not every $(\mathcal{Y}, \ell)$ satisfies (77), the questions of whether Condition 1 is sufficient for universal learning with independent noise, and whether there exist self-adaptive learning rules that are optimistically universal with independent noise, for general (bounded, separable) losses ℓ, remain open.

Open Problem 5 *Is Condition 1 sufficient for a process $\mathbb{X}$ to admit strong universal inductive and self-adaptive learning with independent noise, for* every *separable near-metric space $(\mathcal{Y}, \ell)$ with $\bar{\ell} < \infty$?*

Open Problem 6 *Is it true that, for* every *separable near-metric space $(\mathcal{Y}, \ell)$ with $\bar{\ell} < \infty$, there exists a self-adaptive learning rule that is optimistically universal with independent noise?*

9.3 Learning Noisy Functions

While the above results for learning with independent noise are quite general, it turns out the important problem of *classification* with the 0-1 loss is not directly covered by these results, since it does not guarantee the existence of functions $\underline{\phi}$, $\overline{\phi}$ satisfying (77). Fortunately, we can extend the above theory to a result on classification for *noisy functions* via the well-known *plug-in* classifier technique (see e.g., Devroye, Györfi, and Lugosi, 1996, Theorem 2.2).

Specifically, let $\hat{f}^{sq}_n$ and $\hat{f}^{sq}_{n,m}$ be the inductive and self-adaptive learning rules $\hat{f}_n$ and $\hat{f}_{n,m}$ from (78) and (79), respectively, but defined for the setting $(\mathcal{Y}, \ell) = ([0,1], \ell_{\text{sq}})$ (where $\ell_{\text{sq}}(a,b) = (a-b)^2$).[12] In the finite classification problem, we consider a setting with $\mathcal{Y}$ a finite set with $|\mathcal{Y}| \geq 2$, and $\ell = \ell_{01}$ (where $\ell_{01}(a,b) = \mathbb{1}[a \neq b]$). For any $n \in \mathbb{N}$, $x_{1:n} \in \mathcal{X}^n$, $y_{1:n} \in \mathcal{Y}^n$, and $x \in \mathcal{X}$, define an inductive learning rule

$$\hat{h}_n(x_{1:n}, y_{1:n}, x) = \operatorname*{argmax}_{y \in \mathcal{Y}} \hat{f}^{sq}_n(x_{1:n}, \mathbb{1}_{\{y\}}(y_{1:n}), x). \tag{82}$$

Similarly, for any $n, m \in \mathbb{N}$ $(m \geq n)$, any $x_{1:m} \in \mathcal{X}^m$, $y_{1:n} \in \mathcal{Y}^n$, and $x \in \mathcal{X}$, define a self-adaptive learning rule

$$\hat{h}_{n,m}(x_{1:m}, y_{1:n}, x) = \operatorname*{argmax}_{y \in \mathcal{Y}} \hat{f}^{sq}_{n,m}(x_{1:m}, \mathbb{1}_{\{y\}}(y_{1:n}), x). \tag{83}$$

Since $\hat{f}^{sq}_n$ and $\hat{f}^{sq}_{n,m}$ are measurable functions, the functions $\hat{h}_n$ and $\hat{h}_{n,m}$ can also be defined as measurable functions (e.g., by breaking ties in the argmax based on a pre-specified preference order on the finite set $\mathcal{Y}$); for simplicity, let us suppose ties in the argmax are broken deterministically, so that $\hat{h}_n$ and $\hat{h}_{n,m}$ are deterministic functions.

Note that when $\mathbb{Y}$ is a noisy function of $\mathbb{X}$, for every $y \in \mathcal{Y}$ the conditional probability $\mathbb{P}(Y_t = y | X_t)$ is a t-invariant function of X_t: that is, there is a function $\eta(\cdot; y)$ such that $\eta(X_t; y) = \mathbb{P}(Y_t = y | X_t)$ for every t. Moreover, the value $\eta(X_t; y)$ minimizes $\mathbb{E}\big[(\eta(X_t; y) - \mathbb{1}_{\{y\}}(Y_t))^2 \big| X_t\big]$ almost surely. Thus, for $\mathbb{X} \in \mathcal{C}_1$, Lemma 66 implies that for each $y \in \mathcal{Y}$, the estimators $\hat{f}^{sq}_n(X_{1:n}, \mathbb{1}_{\{y\}}(Y_{1:n}), X_{m+1})$ and $\hat{f}^{sq}_{n,m}(X_{1:m}, \mathbb{1}_{\{y\}}(Y_{1:n}), X_{m+1})$ will (on average, in the limit) be very close to $\mathbb{P}(Y_{m+1} = y | X_{m+1})$ for all large n. By this fact, we see that the learning rules $\hat{h}_n$ and $\hat{h}_{n,m}$ will (on average, in the limit) predict with a value y that nearly maximizes $\mathbb{P}(Y_{m+1} = y | X_{m+1})$, and therefore achieves a nearly-minimal 0-1 loss for all large n. The following theorem formalizes these claims, and summarizes our results on learning for noisy functions.

Theorem 68 *For finite $\mathcal{Y}$ with $|\mathcal{Y}| \geq 2$, and $\ell = \ell_{01}$, Condition 1 is necessary and sufficient for a process $\mathbb{X}$ to admit strong universal (inductive/self-adaptive) learning for noisy functions. Moreover, the self-adaptive learning rule $\hat{h}_{n,m}$ defined by* (83) *is optimistically universal for noisy functions.*

Proof The fact that Condition 1 is necessary for $\mathbb{X}$ to admit strong universal inductive or self-adaptive learning for noisy functions is immediate from Theorem 7 since the noise-free case is a special case of a noisy function: that is, defining $Y_t = f^\star(X_t)$ for a t-invariant measurable function $f^\star$ always satisfies the property of $\mathbb{Y}$ being a noisy function of $\mathbb{X}$.

For the sufficiency claim, for brevity, we only present the details for the self-adaptive learning rule $\hat{h}_{n,m}$. The proof for the inductive learning rule $\hat{h}_n$ is essentially identical,

12. In fact, it is not hard to show that, in the arguments below, it suffices to take $\hat{f}^{sq}_n$ and $\hat{f}^{sq}_{n,m}$ as *any* learning rules that are strongly universally consistent for noisy functions with respect to $([0,1], \ell_{\text{sq}})$ under the given $\mathbb{X} \in \mathcal{C}_1$, and the resulting "plug-in" learning rules $\hat{h}_n$ and $\hat{h}_{n,m}$ will then be strongly universally consistent for noisy functions under $\mathbb{X}$ with respect to $(\mathcal{Y}, \ell_{01})$ for finite $\mathcal{Y}$. As this more general reduction requires a few extra steps in the proof, we present the result specialized to (78) and (79) for simplicity.

except replacing every occurence of $\hat{h}_{n,m}(X_{1:m}, Y_{1:n}, X_{m+1})$ with $\hat{h}_n(X_{1:n}, Y_{1:n}, X_{m+1})$ and every occurence of $\hat{f}^{sq}_{n,m}(X_{1:m}, \mathbb{1}_{\{y\}}(Y_{1:n}), X_{m+1})$ with $\hat{f}^{sq}_n(X_{1:n}, \mathbb{1}_{\{y\}}(Y_{1:n}), X_{m+1})$.

We now proceed with the proof for the self-adaptive rule $\hat{h}_{n,m}$. Suppose $\mathbb{X}$ satisfies Condition 1, and that $\mathbb{Y}$ is a noisy function of $\mathbb{X}$. As mentioned above, since $\mathbb{Y}$ is a noisy function of $\mathbb{X}$, for every $y \in \mathcal{Y}$ the conditional probability $\mathbb{P}(Y_t = y|X_t)$ is a t-invariant function of X_t. Also, as is well known, for any $p \in [0,1]$, it holds that $\mathbb{E}\big[(p - \mathbb{1}_{\{y\}}(Y_t))^2 \big| X_t\big] = \mathbb{E}\big[(p - \mathbb{P}(Y_t = y|X_t))^2 \big| X_t\big] + \mathbb{E}\big[(\mathbb{P}(Y_t = y|X_t) - \mathbb{1}_{\{y\}}(Y_t))^2 \big| X_t\big]$, which is minimized at $p = \mathbb{P}(Y_t = y|X_t)$ almost surely. Therefore, the process $\{(X_t, \mathbb{1}_{\{y\}}(Y_t))\}_{t=1}^{\infty}$ satisfies property Y1 for the squared loss ℓ_{sq} on $[0,1]$, with the function $x \mapsto \eta(x;y) := \mathbb{P}(Y_t = y|X_t = x)$ being the function realizing the minimum value of $\mathbb{E}[\ell_{\text{sq}}(\eta(X_t;y), \mathbb{1}_{\{y\}}(Y_t))|X_t]$ (a.s.). Furthermore, since $(\mathbb{X}, \mathbb{Y})$ has independent noise, it follows immediately that, $\forall t \in \mathbb{N}$, the variable $\mathbb{1}_{\{y\}}(Y_t)$ is conditionally independent of $\{(X_{t'}, \mathbb{1}_{\{y\}}(Y_{t'}))\}_{t' \neq t}$ given X_t: that is, the process $\{(X_t, \mathbb{1}_{\{y\}}(Y_t))\}_{t=1}^{\infty}$ also satisfies property Y2. Therefore, the process $\{(X_t, \mathbb{1}_{\{y\}}(Y_t))\}_{t=1}^{\infty}$ has independent noise (under the loss ℓ_{sq} on $[0,1]$). Furthermore, as discussed above, the loss ℓ_{sq} on $[0,1]$ satisfies (77) with $\underline{\phi}(x) = \overline{\phi}(x) = x$. Therefore, Lemma 66 and the union bound imply that, on an event E of probability one, we have

$$\max_{y \in \mathcal{Y}} \limsup_{n \to \infty} \limsup_{t \to \infty} \frac{1}{t+1} \sum_{m=n}^{n+t} \Big(\hat{f}^{sq}_{n,m}(X_{1:m}, \mathbb{1}_{\{y\}}(Y_{1:n}), X_{m+1}) - \eta(X_{m+1};y) \Big)^2 = 0.$$

Furthermore, since the maximum of a finite number of values is continuous and nondecreasing in those values, this implies that on the event E,

$$\limsup_{n \to \infty} \limsup_{t \to \infty} \max_{y \in \mathcal{Y}} \frac{1}{t+1} \sum_{m=n}^{n+t} \Big(\hat{f}^{sq}_{n,m}(X_{1:m}, \mathbb{1}_{\{y\}}(Y_{1:n}), X_{m+1}) - \eta(X_{m+1};y) \Big)^2 = 0,$$

and again because $\mathcal{Y}$ is a finite set, this implies that

$$\begin{aligned}
&\limsup_{n \to \infty} \limsup_{t \to \infty} \frac{1}{t+1} \sum_{m=n}^{n+t} \max_{y \in \mathcal{Y}} \Big(\hat{f}^{sq}_{n,m}(X_{1:m}, \mathbb{1}_{\{y\}}(Y_{1:n}), X_{m+1}) - \eta(X_{m+1};y) \Big)^2 \\
&\leq \limsup_{n \to \infty} \limsup_{t \to \infty} \frac{1}{t+1} \sum_{m=n}^{n+t} \sum_{y \in \mathcal{Y}} \Big(\hat{f}^{sq}_{n,m}(X_{1:m}, \mathbb{1}_{\{y\}}(Y_{1:n}), X_{m+1}) - \eta(X_{m+1};y) \Big)^2 \\
&\leq \limsup_{n \to \infty} \limsup_{t \to \infty} |\mathcal{Y}| \max_{y \in \mathcal{Y}} \frac{1}{t+1} \sum_{m=n}^{n+t} \Big(\hat{f}^{sq}_{n,m}(X_{1:m}, \mathbb{1}_{\{y\}}(Y_{1:n}), X_{m+1}) - \eta(X_{m+1};y) \Big)^2 = 0.
\end{aligned}$$

By Jensen's inequality, this further implies that on the event E,

$$\limsup_{n \to \infty} \limsup_{t \to \infty} \left(\frac{1}{t+1} \sum_{m=n}^{n+t} \max_{y \in \mathcal{Y}} \Big| \hat{f}^{sq}_{n,m}(X_{1:m}, \mathbb{1}_{\{y\}}(Y_{1:n}), X_{m+1}) - \eta(X_{m+1};y) \Big| \right)^2 = 0,$$

which (since $x \mapsto x^2$ is continuous and nondecreasing on $[0,1]$) implies

$$\limsup_{n \to \infty} \limsup_{t \to \infty} \frac{1}{t+1} \sum_{m=n}^{n+t} \max_{y \in \mathcal{Y}} \Big| \hat{f}^{sq}_{n,m}(X_{1:m}, \mathbb{1}_{\{y\}}(Y_{1:n}), X_{m+1}) - \eta(X_{m+1};y) \Big| = 0. \tag{84}$$

For each $n, m \in \mathbb{N}$ with $m \geq n$, define $\hat{Y}_{n,m+1} = \hat{h}_{n,m}(X_{1:m}, Y_{1:n}, X_{m+1})$ and $\Delta^n_{m+1} = \mathbb{1}\Big[\hat{Y}_{n,m+1} \neq Y_{m+1}\Big] - \mathbb{1}[f^\star(X_{m+1}) \neq Y_{m+1}]$. Then note that, due to the conditional independence property Y2, for each $n \in \mathbb{N}$, the sequence

$$\left\{\Delta^n_{m+1} - \mathbb{E}\Big[\Delta^n_{m+1}\Big| X_{m+1}, \hat{Y}_{n,m+1}\Big]\right\}_{m=n}^{\infty}$$

is a martingale difference sequence with respect to $\{(X_{1:(m+2)}, Y_{1:(m+1)})\}_{m=n}^{\infty}$. Therefore, Azuma's inequality (e.g., Devroye, Györfi, and Lugosi, 1996, Theorem 9.1) implies that, for any $t \in \mathbb{N} \cup \{0\}$, with probability at least $1 - \frac{1}{n^2(t+1)^2}$,

$$\frac{1}{t+1}\left|\sum_{m=n}^{n+t} \Delta^n_{m+1} - \mathbb{E}\Big[\Delta^n_{m+1}\Big| X_{m+1}, \hat{Y}_{n,m+1}\Big]\right| \leq \sqrt{\frac{2\ln(2n^2(t+1)^2)}{t+1}}.$$

Since $\sum_{n=1}^{\infty}\sum_{t=0}^{\infty} \frac{1}{n^2(t+1)^2} < \infty$, the Borel-Cantelli lemma implies that, on an event E' of probability one,

$$\limsup_{n\to\infty} \limsup_{t\to\infty} \frac{1}{t+1}\left|\sum_{m=n}^{n+t} \Delta^n_{m+1} - \mathbb{E}\Big[\Delta^n_{m+1}\Big| X_{m+1}, \hat{Y}_{n,m+1}\Big]\right| = 0,$$

which implies

$$\limsup_{n\to\infty} \hat{\mathcal{L}}_{\mathbb{X}}(\hat{h}_{n,\cdot}, \mathbb{Y}; n, f^\star) = \limsup_{n\to\infty} \limsup_{t\to\infty} \frac{1}{t+1}\sum_{m=n}^{n+t} \mathbb{E}\Big[\Delta^n_{m+1}\Big| X_{m+1}, \hat{Y}_{n,m+1}\Big].$$

Next, note that for any $n, m \in \mathbb{N}$ with $m \geq n$, by the conditional independence property Y2, it holds that

$$\begin{aligned}
&\mathbb{E}\Big[\Delta^n_{m+1}\Big| X_{m+1}, \hat{Y}_{n,m+1}\Big] \\
&= \mathbb{P}\Big(Y_{m+1} \neq \hat{Y}_{n,m+1}\Big| X_{m+1}, \hat{Y}_{n,m+1}\Big) - \mathbb{P}(Y_{m+1} \neq f^\star(X_{m+1})|X_{m+1}) \\
&= \mathbb{P}(Y_{m+1} = f^\star(X_{m+1})|X_{m+1}) - \mathbb{P}\Big(Y_{m+1} = \hat{Y}_{n,m+1}\Big| X_{m+1}, \hat{Y}_{n,m+1}\Big).
\end{aligned}$$

Recalling that $\eta(X_{m+1}; y) = \mathbb{P}(Y_{m+1} = y|X_{m+1})$, the conditional independence property Y2 implies the last expression above equals

$$\begin{aligned}
&\eta(X_{m+1}; f^\star(X_{m+1})) - \eta(X_{m+1}; \hat{Y}_{n,m+1}) \\
&\leq \eta(X_{m+1}; f^\star(X_{m+1})) - \hat{f}^{sq}_{n,m}(X_{1:m}, \mathbb{1}_{\{f^\star(X_{m+1})\}}(Y_{1:n}), X_{m+1}) \\
&\quad + \max_{y\in\mathcal{Y}} \hat{f}^{sq}_{n,m}(X_{1:m}, \mathbb{1}_{\{y\}}(Y_{1:n}), X_{m+1}) - \eta(X_{m+1}; \hat{Y}_{n,m+1}) \\
&= \eta(X_{m+1}; f^\star(X_{m+1})) - \hat{f}^{sq}_{n,m}(X_{1:m}, \mathbb{1}_{\{f^\star(X_{m+1})\}}(Y_{1:n}), X_{m+1}) \\
&\quad + \hat{f}^{sq}_{n,m}(X_{1:m}, \mathbb{1}_{\{\hat{Y}_{n,m+1}\}}(Y_{1:n}), X_{m+1}) - \eta(X_{m+1}; \hat{Y}_{n,m+1}) \\
&\leq 2\max_{y\in\mathcal{Y}} \left|\hat{f}^{sq}_{n,m}(X_{1:m}, \mathbb{1}_{\{y\}}(Y_{1:n}), X_{m+1}) - \eta(X_{m+1}; y)\right|.
\end{aligned}$$

Therefore, on the event E, (84) implies we have

$$\limsup_{n\to\infty}\limsup_{t\to\infty}\frac{1}{t+1}\sum_{m=n}^{n+t}\mathbb{E}\left[\Delta_{m+1}^{n}\middle|X_{m+1},\hat{Y}_{n,m+1}\right]$$
$$\leq 2\limsup_{n\to\infty}\limsup_{t\to\infty}\frac{1}{t+1}\sum_{m=n}^{n+t}\max_{y\in\mathcal{Y}}\left|\hat{f}_{n,m}^{sq}(X_{1:m},\mathbb{1}_{\{y\}}(Y_{1:n}),X_{m+1})-\eta(X_{m+1};y)\right|=0.$$

Altogether, on the event $E\cap E'$, it holds that $\limsup_{n\to\infty}\hat{\mathcal{L}}_{\mathbb{X}}(\hat{h}_{n,\cdot},\mathbb{Y};n,f^\star)\leq 0$. Also, Lemma 64 implies that, on an event E'' of probability one, $\liminf_{n\to\infty}\hat{\mathcal{L}}_{\mathbb{X}}(\hat{h}_{n,\cdot},\mathbb{Y};n,f^\star)\geq 0$. Thus, on the event $E\cap E'\cap E''$ of probability one (by the union bound), we have $\hat{\mathcal{L}}_{\mathbb{X}}(\hat{h}_{n,\cdot},\mathbb{Y};n,f^\star)\to 0$.

Since this holds for any $\mathbb{X}\in\mathcal{C}_1$, it immediately follows that Condition 1 is sufficient for $\mathbb{X}$ to admit strong universal self-adaptive learning for noisy functions. Moreover, note that the definition of $\hat{h}_{n,m}$ has no dependence on the distributions of $\mathbb{X}$ or $\mathbb{Y}$ beyond the data supplied as its arguments, and we have shown that $\hat{h}_{n,m}$ is strongly universally consistent for noisy functions under every $\mathbb{X}$ satisfying Condition 1. Since we have just established Condition 1 is necessary and sufficient for $\mathbb{X}$ to admit strong universal self-adaptive learning for noisy functions, this also completes the proof that $\hat{h}_{n,m}$ is optimistically universal for noisy functions. ■

We leave open the question of whether the above result for classification can be extended to general (bounded, separable) losses ℓ, as stated in the following open problems.

Open Problem 7 *Is Condition 1 sufficient for a process $\mathbb{X}$ to admit strong universal inductive and self-adaptive learning for noisy functions, for* every *separable near-metric space $(\mathcal{Y},\ell)$ with $\bar{\ell}<\infty$?*

Open Problem 8 *Is it true that, for* every *separable near-metric space $(\mathcal{Y},\ell)$ with $\bar{\ell}<\infty$, there exists a self-adaptive learning rule that is optimistically universal for noisy functions?*

Of course, Open Problem 7 would also be resolved by a positive resolution of Open Problem 5, which represents a strictly stronger result. Moreover, a positive resolution of both Open Problems 5 and 6 together would also positively resolve Open Problem 8.

10. Extensions

Here we briefly mention two simple extensions of the above theory. First, we present a straightforward extension to losses ℓ beyond near-metrics, admitting any loss dominated by a nondecreasing function of a near-metric loss. Second, we present an extension of the results to *weak* universal consistency. In this latter case, we find that all of the results for inductive and self-adaptive learning above hold without modification for weak consistency as well. However, interestingly, this is not true for online learning, as we find that the set of processes admitting weak universal online learning is a *strict superset* of the set SUOL of processes admitting strong universal online learning (if $\mathcal{X}$ is infinite and $\bar{\ell}<\infty$).

10.1 More-General Loss Functions

For simplicity, we have chosen to restrict the loss function ℓ to be a *near-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 near-metric ℓ_o, in the sense that $\forall y, y' \in \mathcal{Y}, \ell(y,y') \leq \chi(\ell_o(y,y'))$ for some continuous nondecreasing unbounded function $\chi : [0,\infty) \to [0,\infty)$ with $\chi(0) = 0$, and that also satisfy a non-triviality condition: $\sup_{y_0,y_1 \in \mathcal{Y}} \inf_{y \in \mathcal{Y}} \max\{\ell(y,y_0), \ell(y,y_1)\} > 0$. 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 asymmetric losses, such as in discrete classification with asymmetric misclassification costs.

Here we briefly elaborate on the (minor) changes to the above theory yielding this generalization. For any $z \in [0,\infty)$, define $\chi^{-1}(z) = \inf\{x \in [0,\infty) : \chi(x) \geq z\}$; this always exists since the conditions on χ guarantee that its range is $[0,\infty)$, and moreover by continuity of χ we have $\chi(\chi^{-1}(z)) = z$. Still defining $\bar{\ell} = \sup_{y,y' \in \mathcal{Y}} \ell(y,y')$, 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 $\chi^{-1}(\bar{\ell})$, since the near-metric $(y,y') \mapsto \ell_o(y,y') \wedge \chi^{-1}(\bar{\ell})$ still satisfies the requirement $\ell(y,y') \leq \chi(\ell_o(y,y') \wedge \chi^{-1}(\bar{\ell}))$. Then we can simply replace ℓ with ℓ_o in the learning rules proposed in (12) and (34), 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

$$\ell(\hat{y},y^\star) \leq \chi(\ell_o(\hat{y},y^\star)) \leq \varepsilon + \bar{\ell}\mathbb{1}\big[\ell_o(\hat{y},y^\star) > \chi^{-1}(\varepsilon)\big] \leq \varepsilon + \frac{\bar{\ell}}{\chi^{-1}(\varepsilon)}\ell_o(\hat{y},y^\star)\,,$$

noting that $\chi^{-1}(\varepsilon) > 0$. 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 20, Theorems 6 and 37), 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 near-metric, any distinct $y_0, y_1 \in \mathcal{Y}$ have $\inf_{y \in \mathcal{Y}} (\ell(y,y_0) + \ell(y,y_1)) \geq \frac{1}{c_\ell}\ell(y_0,y_1) > 0$, but the arguments would hold just as well for these more-general losses ℓ by replacing $\frac{1}{c_\ell}\ell(y_0,y_1)$ with $\inf_{y \in \mathcal{Y}} \max\{\ell(y,y_0), \ell(y,y_1)\}$ 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 (68) again works for

establishing universal consistency with ℓ under $\mathbb{X} \in \mathcal{C}_3$ in Lemma 57, but in this case it follows from the stronger guarantee (76) for ℓ_o (together with continuity and monotonicity of χ, and $\chi(0) = 0$) 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 (76) for the loss ℓ_o under $\mathbb{X} \in \mathcal{C}_3$, and the properties of χ imply that it remains true for $\chi(\ell_o(\cdot,\cdot))$, and hence also for the loss ℓ. However, the complementary result in Lemma 54 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 54 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 (66)), 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.

10.2 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 19 and 20) 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 20 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, 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)

$$\lim_{n\to\infty} \mathbb{E}\left[\hat{\mu}_{\mathbb{X}}\left(\bigcup\{A_i : X_{1:n} \cap A_i = \emptyset\}\right)\right] = \mathbb{E}\left[\lim_{n\to\infty} \hat{\mu}_{\mathbb{X}}\left(\bigcup\{A_i : X_{1:n} \cap A_i = \emptyset\}\right)\right] > 0.$$

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 27 and 19), as well as the result on optimistically universal self-adaptive learning (Theorem 29), continue to hold for the *weak* universal consistency criterion in place of *strong* universal consistency. In particular, this means that the set of processes WUIL (or WUAL) 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 (WUOL) 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 $B_k \sim \text{Bernoulli}(\frac{1}{k})$. 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 $\sum_{k=1}^{\infty} \frac{1}{k} = \infty$, 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 37, $\mathbb{X} \notin$ SUOL. However, if we take f_n as the simple memorization-based online learning rule (from the proof of Theorem 39), then for any $n \in \mathbb{N}$ and measurable $f^\star : \mathcal{X} \to \mathcal{Y}$, we have $\mathbb{E}\left[\hat{\mathcal{L}}_{\mathbb{X}}(f_\cdot, f^\star; n)\right] \leq \frac{\bar{\ell}}{n}\mathbb{E}\left[|\{j \in \mathbb{N} \cup \{0\} : X_{1:n} \cap \{z_j\} \neq \emptyset\}|\right] \leq \frac{\bar{\ell}}{n}\left(1 + \sum_{k=1}^{\lfloor \log_2(2n) \rfloor} \frac{1}{k} 2^{k-1}\right) \leq \frac{\bar{\ell}}{n}\left(1 + \int_1^{\lfloor \log_2(4n) \rfloor} \frac{1}{x} 2^{x-1} \mathrm{d}x\right)$. Since $\int_1^t \frac{1}{x} 2^{x-1} \mathrm{d}x = o\left(\int_1^t 2^x \mathrm{d}x\right)$ 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 $\int_1^{\lfloor \log_2(4n) \rfloor} \frac{1}{x} 2^{x-1} \mathrm{d}x = o(n)$, so that $\mathbb{E}\left[\hat{\mathcal{L}}_{\mathbb{X}}(f_\cdot, f^\star; n)\right] \to 0$, 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 37, 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 39, 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 50 and 51 extend to weak universal consistency without modification. Specifically, since almost sure convergence implies convergence in probability, Theorem 50 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 54 can be used to show that Condition 3 is also necessary for weak universal learning (again in all three settings) when $\bar{\ell} = \infty$. Briefly, for any $\mathbb{X} \notin \mathcal{C}_3$, in the notation defined in the proof of Lemma 54, we would have that for any online learning rule h_n, every $j \in \mathbb{N}$ has $\mathbb{P}\left(\hat{\mathcal{L}}_{\mathbb{X}}(h_\cdot, f_K^\star; T_j) > \frac{1}{2c_\ell}\right) \geq \frac{1}{2}\mathbb{P}(0 < \tau_j \leq T_j) > \frac{1}{2}(\mathbb{P}(E) - 2^{-j})$, 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 $\limsup_{n\to\infty} \mathbb{P}\left(\hat{\mathcal{L}}_{\mathbb{X}}(h_\cdot, f^\star_\kappa; n) > \frac{1}{2c_\ell}\right) > 0$, 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 $n \in \mathbb{N}$, $\mathbb{P}\left(\hat{\mathcal{L}}_{\mathbb{X}}(g_{n,\cdot}, f^\star_K; n) \geq \frac{1}{2c_\ell}\right) \geq \mathbb{P}(E \cap E') > 0$, which implies $\exists \kappa \in [0,1)$ such that $\limsup_{n\to\infty} \mathbb{P}\left(\hat{\mathcal{L}}_{\mathbb{X}}(g_{n,\cdot}, f^\star_\kappa; n) \geq \frac{1}{2c_\ell}\right) > 0$, 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 50 and 51 and the fact that almost sure convergence implies convergence in probability, this also implies that (when $\bar{\ell} = \infty$) there exists an optimistically universal learning rule (in all three settings) under this weak consistency criterion as well.

11. 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 dependent on ℓ regard the case $\bar{\ell} < \infty$, and always restrict to $(\mathcal{Y}, \ell)$ a separable near-metric space.

- 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 $(\mathcal{Y}, \ell)$ 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?
- Open Problem (5 / 7): Does every $\mathbb{X} \in \mathcal{C}_1$ admit strong universal inductive and self-adaptive learning (with independent noise / for noisy functions) for every $(\mathcal{Y}, \ell)$?
- Open Problem (6 / 8): Is it true that, for every $(\mathcal{Y}, \ell)$, there exists a self-adaptive learning rule that is optimistically universal (with independent noise / for noisy functions)?

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

T. M. Adams and A. B. Nobel. Uniform convergence of Vapnik-Chervonenkis classes under ergodic sampling. *Annals of Probability*, 38(4):1345–1367, 2010a. 1.3

T. M. Adams and A. B. Nobel. The gap dimension and uniform laws of large numbers for ergodic processes. *arXiv:1007.2964*, 2010b. 1.3

T. M. Adams and A. B. Nobel. Uniform approximation of Vapnik–Chervonenkis classes. *Bernoulli*, 18(4):1310–1319, 2012. 1.3

P. Algoet. Universal schemes for prediction, gambling and portfolio selection. *The Annals of Probability*, 20(2):901–941, 1992. 1.3

P. Algoet. The strong law of large numbers for sequential decisions under uncertainty. *IEEE Transactions on Information Theory*, 40(3):609–633, 1994. 1.3

P. Algoet. Universal schemes for learning the best nonlinear predictor given the infinite past and side information. *IEEE Transactions on Information Theory*, 45(4):1165–1185, 1999. 1.3

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

D. H. Bailey. *Sequential Schemes for Classifying and Predicting Ergodic Processes*. PhD thesis, Department of Mathematics, Stanford University, 1976. 1.3

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

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006. 1, 1.1, 6

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

G. Collomb. Propriétés de convergence presque complète du prédicteur à noyau. *Zeitschrift f ur Wahrscheinlichkeitstheorie und verwandte Gebiete*, 66:441–460, 1984. 1.3

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

T. M. Cover. Open problems in information theory. In *IEEE USSR Joint Workshop on Information Theory*, 1975. 1.3

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag New York, 1996. 1, 3.2, 9.1, 9.1, 9.2, 9.2, 9.3, 9.3

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

D. H. Fremlin. *Measure Theory, Volume 3*. Torres Fremlin, 2002. 2.2

R. M. Gray. *Probability, Random Processes, and Ergodic Properties*. Springer, second edition, 2009. 1.3, 3, 3.1, 5.2

R. M. Gray and J. C. Kieffer. Asymptotically mean stationary measures. *The Annals of Probability*, 8(5):962–973, 1980. 1.3

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. 1.3, 6.1

L. Györfi, G. Lugosi, and G. Morvai. A simple randomized algorithm for sequential prediction of ergodic time series. *IEEE Transactions on Information Theory*, 45(7):2642–2650, 1999. 1.3

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, 9.1

S. Hanneke and S. Kpotufe. On the value of target data in transfer learning. In *Advances in Neural Information Processing Systems 32*, 2019. 1, 1.1

S. Hanneke and L. Yang. Statistical learning under nonstationary mixing processes. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 2019. 1.3

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. Irle. On consistency in nonparametric estimation under mixing conditions. *Journal of Multivariate Analysis*, 60(1):123–147, 1997. 1.3

R. L. Karandikar and M. Vidyasagar. Rates of uniform convergence of empirical means with mixing processes. *Statistics & Probability Letters*, 58(3):297–307, 2002. 1.3

R. L. Karandikar and M. Vidyasagar. Probably approximately correct learning with beta mixing input sequences, 2004. 1.3

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* 4th *European Conference on Computational Learning Theory*, 1999. 6.1

A. N. Kolmogorov and S. V. Fomin. *Introductory Real Analysis*. Dover, 1975. 5.2

S. R. Kulkarni, S. E. Posner, and S. Sandilya. Data-dependent k_n-nn and kernel estimators consistent for arbitrary processes. *IEEE Transactions on Information Theory*, 48(10), 2002. 1.3

V. Kuznetsov and M. Mohri. Generalization bounds for time series prediction with non-stationary processes. In *Proceedings of The 25$^{\text{th}}$ International Conference on Algorithmic Learning Theory*, 2014. 1.3

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

A. C. Lozano, S. R. Kulkarni, and R. E. Schapire. Convergence and consistency of regularized boosting algorithms with stationary β-mixing observations. In *Advances in Neural Information Processing Systems 18*, 2006. 1.3

D. Maharam. An algebraic characterization of measure algebras. *Annals of Mathematics*, 48(1):154–167, 1947. 2.2

G. Morvai, S. Yakowitz, and L. Györfi. Nonparametric inference for ergodic, stationary time series. *The Annals of Statistics*, 24(1):370–379, 1996. 1.3

G. Morvai, S. R. Kulkarni, and A. B. Nobel. Regression estimation from an individual stable sequence. *Statistics*, 33:99–118, 1999. 1.3

A. B. Nobel. Limits to classification and regression estimation from ergodic processes. *The Annals of Statistics*, 27(1):262–273, 1999. 1.3, 5.2

A. B. Nobel. On optimal sequential prediction for general processes. *IEEE Transactions on Information Theory*, 49(1):83–98, 2003. 1.3

G. L. O'Brien and W. Vervaat. How subadditive are subadditive capacities? *Commentationes Mathematicae Universitatis Carolinae*, 35(2):311–324, 1994. 2.2

D. S. Ornstein. Guessing the next output of a stationary process. *Israel Journal of Mathematics*, 30(3):292–296, 1978. 1.3

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

G. G. Roussas. Nonparametric estimation in mixing sequences of random variables. *Journal of Statistical Planning and Inference*, 18:135–149, 1988. 1.3

B. Ryabko. Prediction of random sequences and universal coding. *Problems of Information Theory*, 24(2):87–96, 1988. 1.3

D. Ryabko. Pattern recognition for conditionally independent data. *Journal of Machine Learning Research*, 7(4):645–664, 2006. 1.3

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, 1.3

M. Talagrand. Maharam's problem. *Annals of Mathematics*, 168(3):981–1009, 2008. 2.2

R. van Handel. The universal Glivenko–Cantelli property. *Probability Theory and Related Fields*, 155(3–4):911–934, 2013. 1.3

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

M. Vidyasagar. Convergence of empirical means with alpha-mixing input sequences, and an application to PAC learning. In *Proceedings of the 44$^{\text{th}}$ IEEE Conference on Decision and Control*, pages 560–565, 2005. 1.3

V. Vovk. Aggregating strategies. In *Proceedings of the* 3$^{\text{rd}}$ *Annual Workshop on Computational Learning Theory*, 1990. 6.1

V. Vovk. Universal forecasting algorithms. *Information and Computation*, 96(2):245–277, 1992. 6.1

B. Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 22(1):94–116, 1994. 1.3

B. Zou, L. Li, and Z. Xu. The generalization performance of ERM algorithm with strongly mixing observations. *Machine Learning*, 75(3):275–295, 2009. 1.3