Accepted, Annals of Statistics, June 2023

UNIVERSAL REGRESSION WITH ADVERSARIAL RESPONSES

BY MOÏSE BLANCHARD[1,a] AND PATRICK JAILLET[2,b]

[1]*Operations Research Center, Massachusetts Institute of Technology,* [a]*moiseb@mit.edu*

[2]*Department of Electrical Engineering and Computer Science, Laboratory for Information and Decision Systems, Operations Research Center, Massachusetts Institute of Technology,* [b]*jaillet@mit.edu*

We provide algorithms for regression with adversarial responses under large classes of non-i.i.d. instance sequences, on general separable metric spaces, with *provably minimal* assumptions. We also give characterizations of learnability in this regression context. We consider *universal consistency* which asks for strong consistency of a learner without restrictions on the value responses. Our analysis shows that such an objective is achievable for a significantly larger class of instance sequences than stationary processes, and unveils a fundamental dichotomy between value spaces: whether finite-horizon mean estimation is achievable or not. We further provide *optimistically universal* learning rules, i.e., such that if they fail to achieve universal consistency, any other algorithms will fail as well. For unbounded losses, we propose a mild integrability condition under which there exist algorithms for adversarial regression under large classes of non-i.i.d. instance sequences. In addition, our analysis also provides a learning rule for mean estimation in general metric spaces that is consistent under adversarial responses without any moment conditions on the sequence, a result of independent interest.

1. Introduction.

1.1. *Motivation and background.* We study the classical statistical problem of metric-valued regression. Given an instance metric space $(\mathcal{X}, \rho_{\mathcal{X}})$ and a value metric space $(\mathcal{Y}, \rho_{\mathcal{Y}})$ with a loss ℓ, one observes instances in $\mathcal{X}$ and aims to predict the corresponding values in $\mathcal{Y}$. The learning procedure follows an iterative process where successively, the learner is given an instance X_t and predicts the value Y_t based on the historical samples and the new instance. The learner's goal is to minimize the loss of its predictions $\hat{Y}_t$ compared to the true value Y_t. In particular, $\mathcal{Y} = \{0, 1\}$ (resp. $\mathcal{Y} = \{0, \ldots, k\}$) with 0-1 loss corresponds to binary (resp. multiclass) classification while $\mathcal{Y} = \mathbb{R}$ corresponds to the classical regression setting. Motivated by the increase of new types of data in numerous data analysis applications—e.g., data lying on spherical spaces [8, 29], manifolds [34, 10, 14], Hilbert spaces [38], Hadamard spaces [26]—we will study the case where both instances and value spaces are general separable metric spaces. This general setting adopted in the recent literature on universal learning [21, 9, 2] includes and extends the specific classification and regression settings mentioned above. In this context, we model the stream of data as a general stochastic process $(\mathbb{X}, \mathbb{Y}) := (X_t, Y_t)_{t\geq 1}$, and are interested in *consistent* predictions that have vanishing average *excess* loss compared to any fixed measurable predictor functions $f : \mathcal{X} \to \mathcal{Y}$, i.e., $\frac{1}{T}\sum_{t=1}^{T} \ell(\hat{Y}_t, Y_t) - \ell(f(X_t), Y_t) \to 0$ $(a.s.)$. Naturally, one would hope that the algorithm converges for a large class of value functions. Thus, we are interested in *universally consistent* learning rules that are consistent irrespective of the value process $\mathbb{Y}$.

The i.i.d. version of this problem where one assumes that the sequence $(\mathbb{X}, \mathbb{Y})$ is i.i.d. has been extensively studied. A classical result is that for binary classification in Euclidean

MSC2020 subject classifications: Primary 62C05, 62J02, 97K80, 60G20; secondary 54E70.

Keywords and phrases: statistical learning theory, consistency, nonparametric estimation, generalization, stochastic processes, online learning, metric spaces.

spaces, $k-$nearest neighbor (kNN) with $k/\ln T \to \infty$ and $k/T \to 0$ is universally consistent under mild assumptions on the distribution of (X_1, Y_1) [37, 11, 12]. These results were then extended to a broader class of spaces [12, 17] and more recently, [24, 20, 9] provided universally consistent algorithms for any essentially separable metric space $\mathcal{X}$ which are precisely those for which universal consistency is achievable for i.i.d. pairs $(X_t, Y_t)_{t\geq 1}$ of instances and responses. In parallel, a significant line of work aimed to obtain such results in non-i.i.d. settings, notably relaxations of the i.i.d. assumptions such as stationary ergodic processes [31, 18, 17] or processes satisfying the law of large numbers [30, 16, 36].

1.2. *Optimistic universal learning.* In this work, we aim to understand which are the minimal assumptions on the data sequences for which universal consistency is still achievable. As such, we follow the *optimistic decision theory* [22] which formalizes the paradigm of "learning whenever learning is possible". Precisely, the *provably minimal* assumption for a given objective is that this task is achievable, or in other words that learning is possible. The goal then becomes to 1. characterize for which settings this objective is achievable and 2. if possible, provide learning rules that achieve this objective whenever it is achievable. These are called *optimistically universal* learning rules and enjoy the convenient property that if they failed the objective, any other algorithms would fail as well.

1.3. *Related works in universal learning.* This paradigm was recently used to study minimal assumptions for the noiseless (realizable) case where there exists an unknown underlying function $f^*: \mathcal{X} \to \mathcal{Y}$ such that $Y_t = f^*(X_t)$ [22]. In this setting, the two questions described above were recently settled. For bounded losses, a simple variant of the nearest neighbor algorithm is optimistically universal [3, 2] and learnable processes are significantly larger than stationary processes. On the other hand, for unbounded losses, universal regression is extremely restrictive since the only learnable processes are those which visit a finite number of points almost surely [4]. Yet, the general non-realizable setting was not characterized. As an initial result, for bounded losses, [23] proposed an algorithm that achieves universal consistency for a large class of processes $\mathbb{X}$, which intuitively asks that the sub-measure induced by empirical visits of the input sequence be continuous. There is however a significant gap between the proposed condition and the learnable processes in the bounded noiseless setting. [23] then left open the question of identifying the precise provably-minimal conditions to achieve consistency, and whether there exists an optimistically universal learning rule.

1.4. *Adversarial responses and related works in learning with experts.* The consistency results in [23] hold for arbitrary value processes $\mathbb{Y}$, arbitrarily correlated to the instance process $\mathbb{X}$. We consider the slightly more general *adversarial* responses and show that we can obtain the same results as for adversarial processes, without any generalizability cost. Formally, adversarial responses can not only arbitrarily depend on the instance sequence $\mathbb{X}$, but may also depend on past predictions and past randomness used by the learner. This is a non-trivial generalization for randomized algorithms—note that randomization is necessary to obtain guarantees for general online learning problems [5, 35]. There is a rich theory for arbitrary or adversarial responses $\mathcal{Y}$ when the reference functions $f^*: \mathcal{X} \to \mathcal{Y}$ are restricted to specific function classes $\mathcal{F}$. As a classical example, for the noiseless binary classification setting, there exist learning rules which guarantee a finite number of mistakes for arbitrary sequences $\mathbb{X}$, if and only if the class $\mathcal{F}$ has finite Littlestone dimension [27]. Other restrictions on the function class have been considered [7, 1, 32]. Universal learning diverges from this line of work by imposing no restrictions on function classes—namely *all* measurable functions—but instead restricting instance processes $\mathbb{X}$ to the optimistic set where universal consistency is achievable. Nevertheless, the algorithms we introduce for adversarial responses use as subroutine the traditional exponentially weighted forecaster for learning with expert advice from the online learning literature, also known as the Hedge algorithm [28, 6, 15].

1.5. *Contributions.* In this paper, we provide answers to two fundamental questions in universal regression. First, we exactly characterize the set of processes we call *learnable*. These are instance processes $\mathbb{X}$ for which universal learning is possible, i.e., consistency is achieved for every process $(X_t, Y_t)_{t\geq 1}$ with covariate sequence $\mathbb{X}$. Second, we provide optimistically universal learning rules, i.e., a unique algorithm that achieves universal consistency for all processes $\mathbb{X}$ for which this is achievable by some learning rule. The specific answers to these questions depend on the value space and loss $(\mathcal{Y}, \ell)$ as detailed below.

1.5.1. *Universal learning with empirically integrable responses.* We introduce a mild moment-type assumption on the responses $\mathbb{Y}$, namely *empirical integrability*, that roughly asks that one can bound the tails of the empirical first moment of $\mathbb{Y}$. We then proceed to analyze the processes for which learning adversarial responses guaranteed to satisfy this assumption, is achievable. The answer depends on a property of the value space and loss $(\mathcal{Y}, \ell)$ which we denote F-TiME.

- If every ball $B_\ell(y, r)$ of $(\mathcal{Y}, \ell)$ satisfies the F-TiME property, the class of processes $\mathbb{X}$ for which universal consistency under adversarial empirically integrable responses may be achieved is the so-called Sublinear Measurable Visits (SMV) class. This coincides with the class of processes that admits universal learning for bounded losses in the realizable setting (noiseless responses) [2]. In particular, this shows that for value spaces with bounded losses satisfying F-TiME, one can extend consistency results from the realizable setting to the adversarial one at no generalizability cost.
- Otherwise, the classes of processes $\mathbb{X}$ for which one can achieve universal consistency for empirically integrable responses is a smaller class called Continuous Submeasure (CS). This is a condition that was already considered by [23], which showed that for bounded metric losses, one can achieve universal learning under CS processes. Our results show that whenever the F-TiME condition is not satisfied for bounded losses, CS is also a necessary condition for universal learning.

Also, in both cases, we give an optimistically universal learning rule, that is implicit for the first case—it uses as subroutine the learning rule for mean-estimation—and explicit for the second. These results resolve an open question from [23].

Intuitively, the property F-TiME asks that, for any fixed tolerance $\epsilon > 0$, there is a learning rule that solves the analogous prediction problem without covariates $\mathbb{X}$—*mean-estimation*—in finite time within the tolerance ϵ. This property is satisfied for "reasonable" value spaces, e.g., totally-bounded spaces or countably-many-classes classification $(\mathbb{N}, \ell_{01})$, but we also provide an explicit example of bounded metric space that does not satisfy this condition.

To motivate the introduction of the empirical integrability condition we show that a weaker moment-type assumption on responses—that $\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^T \ell(y_0, Y_t) < \infty$ $(a.s.)$ for some $y_0 \in \mathcal{Y}$—is not sufficient to extend the results from the bounded loss case to unbounded losses, resolving an open question from [4]. Further, empirical integrability is essentially necessary to obtain consistency results: it is automatically satisfied if the loss is bounded and for the i.i.d. setting it exactly asks that responses Y have finite first moment.

As a direct implication of this work, finite second moment $\mathbb{E}[Y^2]$ is sufficient to achieve consistency for stationary ergodic processes. This result relaxes the conditions of all past works to the best of our knowledge, which required finite fourth moment $\mathbb{E}[Y^4]$ [19].

1.5.2. *Universal learning with unrestricted responses.* For completeness, we also characterize the set of learnable processes without assuming empirical integrability on responses. Since the two notions coincide for bounded losses, we focus on unbounded losses. While there always exists an optimistically universal learning rule, the precise class of universally learnable processes depends on an alternative involving the mean-estimation problem. Either mean-estimation on $(\mathcal{Y}, \ell)$ is impossible and universal learning is never achievable, or universal learning is achievable for processes that only visit a finite number of distinct points, a property called Finite Support (FS). Along the way, we show that mean-estimation with adversarial responses is always possible for metric losses, a result of independent interest.

1.6. *Organization of the paper.* After presenting the learning framework and definitions in Section 2, we describe in Section 3 our main results. Although these are stated for general value spaces under the empirical integrability constraint, the proofs build upon the bounded loss case. We follow this proof structure: in Section 4 we consider totally-bounded value spaces for which we can give explicit optimistically universal learning rules, in Section 5 we consider general bounded loss spaces. We then turn to unbounded and mean estimation in Section 6. Last, in Section 7 we introduce the empirical integrability and prove our general results for unbounded losses. We discuss open directions in Section 8.

2. Formal setup.

We provide the necessary definitions, concepts and conditions.

2.1. *Instance and value spaces.* Consider a separable metric *instance space* $(\mathcal{X}, \rho_{\mathcal{X}})$ equipped with its Borel $\sigma-$algebra $\mathcal{B}$, and a separable metric *value space* $(\mathcal{Y}, \rho_{\mathcal{Y}})$ given with a loss ℓ. We recall that a metric space is *separable* if it contains a dense countable set. Unless mentioned otherwise, we suppose that the loss is a power of a metric, i.e., there exists $\alpha \geq 1$ such that the loss is $\ell = (\rho_{\mathcal{Y}})^{\alpha}$. As a remark, all of the results in this work can be generalized to *essentially separable* metric instance spaces, a condition introduced by [24] which was shown to be the largest class of metric spaces for which learning possible. However, for the sake of exposition, we restrict ourselves to separable metric spaces. We denote $\bar{\ell} := \sup_{y_1, y_2 \in \mathcal{Y}} \ell(y_1, y_2)$. In the first Sections 4 and 5 of this work, we suppose that the loss ℓ is *bounded*, i.e., $\bar{\ell} < \infty$. The case of *unbounded* losses is addressed in the next sections 6 and 7. We also introduce the notion of near-metrics for which we will provide some results. We say that ℓ is a near-metric on $\mathcal{Y}$ if it is symmetric, satisfies $\ell(y, y) = 0$ for all $y \in \mathcal{Y}$, for any $y' \neq y \in \mathcal{Y}$ we have $\ell(y, y') > 0$, and it satisfies a relaxed triangle inequality $\ell(y_1, y_2) \leq c_\ell(\ell(y_1, y_3) + \ell(y_2, y_3))$ where c_ℓ is a finite constant.

2.2. *Online learning on adversarial responses.* We consider the *online learning* framework where at step $t \geq 1$, one observes a new instance $X_t \in \mathcal{X}$ and predicts a value $\hat{Y}_t \in \mathcal{Y}$ based on the past history $(X_u, Y_u)_{u \leq t-1}$ and the new instance X_t only. The learning rule may be randomized, where the private randomness used at each iteration t is drawn from a fixed probability space $\mathcal{R}$ and independent of the data generation process used to generate Y_t.

DEFINITION 2.1. An *online learning rule* is a sequence $f_\cdot := \{f_t, R_t\}_{t \geq 1}$ of measurable functions $f_t : \mathcal{R} \times \mathcal{X}^{t-1} \times \mathcal{Y}^{t-1} \times \mathcal{X} \to \mathcal{Y}$ together with a distribution R_t on $\mathcal{R}$.

The prediction at time t of $f_\cdot$ is $f_t(r_t; (X_u)_{\leq t-1}, (Y_u)_{\leq t-1}, X_t)$ where $r_t \sim R_t$ is independent of the new value X_t and the past history $(X_u, Y_u)_{\leq t}$. For simplicity, we may omit the internal randomness r_t and write directly $f_t : \mathcal{X}^{t-1} \times \mathcal{Y}^{t-1} \times \mathcal{X} \to \mathcal{Y}$. We are interested in general data-generating processes. To this means, a possible very general choice of instances and values are general stochastic processes $(\mathbb{X}, \mathbb{Y}) := \{(X_t, Y_t)\}_{t \geq 1}$ on the product space $\mathcal{X} \times \mathcal{Y}$. This corresponds to the arbitrarily dependent responses under instance processes $\mathbb{X}$ [23]. In this work, we consider the slightly more general *adversarial responses* where the value Y_t is also allowed to depend on the past private randomness $(r_u)_{u \leq t-1}$ used by the learning rule $f_\cdot$.

DEFINITION 2.2. Let $\mathbb{X}=(X_t)_{t\geq 1}$ be a stochastic process on $\mathcal{X}$. An *adversarial response mechanism* on $\mathbb{X}$ is a stochastic process $\{(\tilde{X}_t, \boldsymbol{Y}_t)\}_{t\geq 1}$ where $\tilde{X}_t \in \mathcal{X}$, $\boldsymbol{Y}_t = \boldsymbol{Y}_t(\cdot \mid \cdot)$ is a Markov kernel from $\mathcal{R}^{t-1}$ to $\mathcal{Y}$, and $(\tilde{X}_t)_{t\geq 1}$ has same distribution as $\mathbb{X}$.

For a given learning rule $f_\cdot$, having observed the sampled randomness $r_1,\ldots,r_{t-1}\in\mathcal{R}$ used by the learning rule before time t, the target value at time t is $Y_t = \boldsymbol{Y}_t(r_1,\ldots,r_{t-1})$. Again, for simplicity, we will refer to the adversarial response mechanism as $\mathbb{Y}$, which allows us to view the data generating process as a usual stochastic process on $\mathcal{X}\times\mathcal{Y}$. Of course, if the learning rule is *deterministic*, adversarial responses are equivalent to arbitrary dependent responses as in [23], but this is not necessarily the case for general *randomized* algorithms.

2.3. *Empirically integrable responses.* We introduce a novel assumption on the responses, namely *empirical integrability*.

DEFINITION 2.3. A process $(Y_t)_{t\geq 1}$ is *empirically integrable* if there exists $y_0\in\mathcal{Y}$ such that for any $\epsilon>0$, almost surely there exists $M\geq 0$ for which

$$\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^T \ell(y_0, Y_t)\mathbb{1}_{\ell(y_0,Y_t)\geq M} \leq \epsilon.$$

Unless mentioned otherwise, we will focus on the case where responses satisfy this property. This is a mild assumption on the responses. Indeed, it is worth noting that this condition is always satisfied if the loss ℓ is bounded. Further, if for some $y_0\in\mathcal{Y}$, $\ell(y_0,Y_t)$ admits moments of order $p>1$, the empirical integrability condition is also satisfied.

2.4. *Universal consistency.* In this general setting, we are interested in online learning rules which achieve low long-run average loss compared to any fixed prediction function for general adversarial mechanisms. Given a learning rule $f_\cdot$ and an adversarial process $(\mathbb{X},\mathbb{Y})$, for any measurable function $f^*:\mathcal{X}\to\mathcal{Y}$, we denote the long-run average excess loss as

$$\mathcal{L}_{(\mathbb{X},\mathbb{Y})}(f_\cdot, f^*) := \limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^T \left(\ell(f_t(\mathbb{X}_{\leq t-1}, \mathbb{Y}_{\leq t-1}, X_t), Y_t) - \ell(f^*(X_t), Y_t)\right).$$

We can then define the notion of consistency which asks that the excess loss compared to any measurable function vanishes to zero.

DEFINITION 2.4. Let $(\mathbb{X},\mathbb{Y})$ be an adversarial process and $f_\cdot$ a learning rule. $f_\cdot$ is consistent under $(\mathbb{X},\mathbb{Y})$ if for any measurable function $f^*:\mathcal{X}\to\mathcal{Y}$, we have $\mathcal{L}_{(\mathbb{X},\mathbb{Y})}(f_\cdot,f^*)\leq 0, \quad (a.s.)$.

For example, if $(\mathbb{X},\mathbb{Y})$ is an i.i.d. process on $\mathcal{X}\times\mathcal{Y}$ following a distribution μ where μ has a finite first-order moment, achieving consistency is equivalent to reaching the optimal risk $R^* := \inf_{f^*}\mathbb{E}_{(X,Y)\sim\mu}\left[\ell(f^*(X),Y)\right]$, where the infimum is taken over all measurable functions $f^*:\mathcal{X}\to\mathcal{Y}$. As introduced in [22, 23], consistency against all measurable function is the natural extension of consistency for i.i.d. processes $(\mathbb{X},\mathbb{Y})$ to non-i.i.d. settings. The goal of universal learning is to design learning rules that are consistent for any adversarial process $\mathbb{Y}$ that is empirically integrable.

DEFINITION 2.5. Let $\mathbb{X}$ be a stochastic process on $\mathcal{X}$ and $f_\cdot$ a learning rule. $f_\cdot$ is *universally consistent* under $\mathbb{X}$ for empirically integrable adversarial responses if for any adversarial process $(\tilde{\mathbb{X}},\mathbb{Y})$ with $\tilde{\mathbb{X}}\sim\mathbb{X}$ and such that $\mathbb{Y}$ is empirically integrable, $f_\cdot$ is consistent.

2.5. *Optimistic universal learning.* Given this regression setup, we define SOLAR (Strong universal Online Learning with Adversarial Responses) as the set of processes $\mathbb{X}$ for which universal consistency with adversarial responses is *achievable*,

$$\begin{aligned}\text{SOLAR} = \{\mathbb{X} : \exists f_\cdot \text{ universally consistent learning rule under } \mathbb{X}\\ \text{for empirically integrable adversarial responses}\}.\end{aligned}$$

Note that this learning rule is allowed to depend on the process $\mathbb{X}$. Similarly, in the realizable (noiseless) setting, one can define the set SOUL (Strong Online Universal Learning) of processes for which there exists a learning rule that is universally consistent for realizable responses when the loss is bounded (and hence, the empirical integrability condition is always satisfied). Of course, SOLAR $\subset$ SOUL. We are then interested in learning rules that would achieve universal consistency whenever possible.

DEFINITION 2.6. A learning rule $f_\cdot$ is *optimistically universal* for adversarial regression with empirically integrable responses if it is universally consistent under all $\mathbb{X} \in$ SOLAR for adversarial empirically integrable responses.

Similarly, we say that a learning rule is optimistically universal for noiseless regression if it is universally consistent under all $\mathbb{X} \in$ SOUL for noiseless responses when the loss is bounded. In this general framework, the main interests of optimistic learning are 1. identifying the set of learnable processes with adversarial responses SOLAR, 2. determining whether there exists an optimistically universal learning rule, and 3. constructing one if it exists.

REMARK 2.7. Except for Section 6.1 in which we assume that the loss is a metric $\alpha = 1$, one can generalize our results to any symmetric and discernible losses ℓ such that for any $0 < \epsilon \leq 1$, there exists a constant c_ϵ such that for all $y_1, y_2, y_3 \in \mathcal{Y}$, $\ell(y_1, y_2) \leq (1+\epsilon)\ell(y_1, y_3) + c_\epsilon \ell(y_2, y_3)$. Without loss of generality, we can further assume that c_ϵ is non-increasing in ϵ. This is a stronger assumption than having a near-metric ℓ, for which we also give some results in Section 4 and 7.

3. Main results. We introduce some conditions on stochastic processes. For any process $\mathbb{X}$ on $\mathcal{X}$, given any measurable set $A \in \mathcal{B}$ of $\mathcal{X}$, let $\hat{\mu}_{\mathbb{X}}(A) := \limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_A(X_t)$. We consider the condition CS (Continuous Sub-measure) defined as follows.

Condition CS: *For every decreasing sequence $\{A_k\}_{k=1}^\infty$ of measurable sets in $\mathcal{X}$ with $A_k \downarrow \emptyset$, $\mathbb{E}[\hat{\mu}_{\mathbb{X}}(A_k)] \underset{k\to\infty}{\longrightarrow} 0$.*

It is known that this condition is equivalent to $\mathbb{E}[\hat{\mu}_{\mathbb{X}}(\cdot)]$ being a continuous sub-measure [22], hence the adopted name CS. Importantly, CS processes contain in particular i.i.d., stationary ergodic or stationary processes. We now introduce a second condition SMV (Sublinear Measurable Visits) which asks that for any partition, the process $\mathbb{X}$ visits a sublinear number of sets of the partition.

Condition SMV: *For every disjoint sequence $\{A_k\}_{k=1}^\infty$ of measurable sets of $\mathcal{X}$ with $\bigcup_{k=1}^\infty A_k = \mathcal{X}$, (every countable measurable partition),*

$$|\{k \geq 1 : A_k \cap \mathbb{X}_{\leq T} \neq \emptyset\}| = o(T), \quad (a.s.).$$

This condition is significantly weaker and allows to consider a larger family of processes CS $\subset$ SMV, with CS $\subsetneq$ SMV whenever $\mathcal{X}$ is infinite [22]. Note that these sets depend on the instance space $(\mathcal{X}, \rho_{\mathcal{X}})$. This dependence is omitted for simplicity. We first consider bounded losses. In the *noiseless* case, where there exists some unknown measurable function $f^*: \mathcal{X} \to \mathcal{Y}$ such that the stochastic process $\mathbb{Y}$ is given as $Y_t = f^*(X_t)$ for all $t \geq 1$, [2] showed that learnable processes are exactly SOUL = SMV for bounded losses. [2] also introduced a learning rule 2-Capped-1-Nearest-Neighbor (2C1NN), variant of the classical 1NN algorithm, which is *optimistically universal* in the noiseless case for bounded losses. Interestingly, we show that this same learning rule is universally consistent for unbounded losses in the noiseless setting with empirically integrable responses.

THEOREM 3.1. *Let $(\mathcal{Y}, \ell)$ be a separable near-metric space. Then, 2C1NN is optimistically universal in the noiseless setting with empirically integrable responses, i.e., for all processes $\mathbb{X} \in$ SMV and for all measurable target functions $f^*: \mathcal{X} \to \mathcal{Y}$ such that $(f^*(X_t))_{t\geq 1}$ is empirically integrable, $\mathcal{L}_{(\mathbb{X},(f^*(X_t))_{t\geq 1})}(2C1NN, f^*) = 0$ $(a.s.)$.*

In general, one has SOLAR $\subset$ SMV. It was posed as a question whether we could recover the complete set SMV for learning under adversarial—or arbitrary—processes [23].

Question [23]: For bounded losses, does there exist an online learning rule that is universally consistent for arbitrary responses under all processes $\mathbb{X} \in$ SMV$(=$ SOUL$)$?

We answer this question with an alternative. Depending on the bounded value space $(\mathcal{Y}, \ell)$, either SOLAR = SMV or SOLAR = CS, but in both cases there exists an optimistically universal learning rule. We now introduce the property F-TiME (Finite-Time Mean Estimation) on the value space $(\mathcal{Y}, \ell)$ which characterizes this alternative.

Property F-TiME: *For any $\eta > 0$, there exists a horizon time $T_\eta \geq 1$, an online learning rule $g_{\leq T_\eta}$ such that for any $\boldsymbol{y} := (y_t)_{t=1}^{T_\eta}$ of values in $\mathcal{Y}$ and any value $y \in \mathcal{Y}$, we have*

$$\frac{1}{T_\eta}\mathbb{E}\left[\sum_{t=1}^{T_\eta} \ell(g_t(\boldsymbol{y}_{\leq t-1}), y_t) - \ell(y, y_t)\right] \leq \eta.$$

We are now ready to state our main results for bounded value spaces. The first result shows that if the value space satisfies the above property locally, we can universally learn all the processes in SOUL even under adversarial responses.

THEOREM 3.2. *Suppose that any ball of $(\mathcal{Y}, \ell)$, $B_\ell(y, r)$ satisfies F-TiME. Then, SOLAR = SMV and there exists an optimistically universal learning rule $f_\cdot$ for adversarial regression with empirically integrable responses., i.e., such that for any stochastic process $(\mathbb{X}, \mathbb{Y})$ on $\mathcal{X} \times \mathcal{Y}$ with $\mathbb{X} \in$ SMV and $\mathbb{Y}$ empirically integrable, for any measurable function $f: \mathcal{X} \to \mathcal{Y}$ we have $\mathcal{L}_{(\mathbb{X},\mathbb{Y})}(f_\cdot, f^*) \leq 0, \quad (a.s.)$.*

F-TiME defines a non-trivial alternative, and an explicit construction of a non-F-TiME bounded metric space $(\mathcal{Y}, \rho_{\mathcal{Y}})$ is given in Section 5.1 with $\mathcal{Y} = \mathbb{N}$. Nevertheless, F-TiME is satisfied by a large class of spaces, e.g., any totally-bounded metric space and countable classification $(\mathcal{Y}, \ell) = (\mathbb{N}, \ell_{01})$ satisfy F-TiME. Hence, we can universally learn all SOUL processes with adversarial responses, for countable classification (the empirical integrability condition is automatically satisfied because the loss is bounded). If F-TiME is not satisfied locally, we have the following result which shows that learning under CS is still possible but universal learning beyond CS processes cannot be achieved.

THEOREM 3.3. *Suppose that there exists a ball $B_\ell(y,r)$ of $(\mathcal{Y},\ell)$ that does not satisfy F-TiME. Then, SOLAR = CS and there exists an optimistically universal learning rule $f_\cdot$ for adversarial regression with empirically integrable responses., i.e., such that for any stochastic process $(\mathbb{X},\mathbb{Y})$ on $(\mathcal{X},\mathcal{Y})$ with $\mathbb{X}\in$ CS and $\mathbb{Y}$ empirically integrable, then, for any measurable function $f:\mathcal{X}\to\mathcal{Y}$ we have $\mathcal{L}_{(\mathbb{X},\mathbb{Y})}(f_\cdot, f^*)\leq 0, \quad (a.s.)$.*

For metric losses $\ell=\rho_{\mathcal{Y}}$, it was already known [23] that universal learning under adversarial responses under all processes in CS is achievable by some learning rule. Hence, Theorem 3.3 implies that this learning rule is automatically optimistically universal for adversarial regression for all metric value spaces with bounded loss which do not satisfy F-TiME. However, our result is stronger in that consistency holds for any power of a metric loss $\ell=\rho_{\mathcal{Y}}^\alpha, \alpha\geq 1$ and unbounded value spaces.

REMARK 3.4. As a direct consequence of Theorems 3.2 and 3.3, for stationary ergodic processes, finite second moment of the values $\mathbb{E}[Y^2]<\infty$ suffices for consistency, in agreement with the known results for the i.i.d. setting. This relaxes the fourth-moment conditions $\mathbb{E}[Y^4]<\infty$ proposed in the literature [19].

We now consider removing the empirical integrability assumption. As mentioned above, for bounded losses this assumption is automatically satisfied, hence Theorem 3.2 and 3.3 apply directly, with a simplified alternative: whether $(\mathcal{Y},\ell)$ satisfies F-TiME.

COROLLARY 3.5. *Suppose that ℓ is bounded.*

- *If $(\mathcal{Y},\ell)$ satisfies F-TiME. Then, SOLAR = SMV(= SOUL).*
- *If $(\mathcal{Y},\ell)$ does not satisfy F-TiME. Then, SOLAR = CS.*

Further, an optimistically universal learning rule for adversarial regression always exists, i.e., achieving universal consistency with adversarial responses under any $\mathbb{X}\in$ SOLAR.

It remains to analyze the case of unbounded losses without empirical integrability assumption on the responses. To avoid confusions, we denote by SOLAR-U the set of processes that admit universal learning with adversarial (unrestricted) responses. Unfortunately, even in the noiseless setting, universal learning is extremely restrictive in that case. Specifically, the set of universally learnable processes SOUL for noiseless responses is reduced to the set FS (Finite Support) of processes that visit a finite number of different points almost surely [4].

Condition FS: The process $\mathbb{X}$ satisfies $|\{x\in\mathcal{X}:\{x\}\cap\mathbb{X}\neq\emptyset\}|<\infty \quad (a.s.)$.

We show that in the adversarial setting we still have SOLAR-U = FS when ℓ is a metric: we can solve the fundamental problem of mean estimation where one sequentially makes predictions of a sequence $\mathbb{Y}$ of values in $(\mathcal{Y},\ell)$ and aims to have a better long-run average loss than any fixed value. If responses $\mathbb{Y}$ are i.i.d. this is the Fréchet means estimation problem [13, 33, 25]. Our main result on mean estimation holds in general spaces and is of independent interest.

THEOREM 3.6. *Let $(\mathcal{Y},\ell)$ be a separable metric space. There exists an online learning rule $f_\cdot$ that is universally consistent for adversarial mean estimation, i.e., for any adversarial process $\mathbb{Y}$ on $\mathcal{Y}$, almost surely, for all $y\in\mathcal{Y}$,*

$$\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^T\left(\ell(f_t(\mathbb{Y}_{\leq t-1}),Y_t)-\ell(y,Y_t)\right)\leq 0.$$

TABLE 1
Characterization of learnable instance processes in universal consistency (ME = Mean Estimation).

Learning setting	Bounded loss	Unbounded loss	Unbounded loss with empirically integrable responses
Noiseless responses	SOUL $=$ SMV [2]	SOUL $=$ FS [4]	Identical to bounded loss [This paper]
Adversarial (or arbitrary) responses	SOLAR $\supset$ CS (metric loss) [23] Does $(\mathcal{Y}, \ell)$ satisfy F-TiME? $\begin{cases}\text{Yes} & \text{SOLAR} = \text{SMV} \\ \text{No} & \text{SOLAR} = \text{CS}\end{cases}$ [This paper]	Is ME achievable? $\begin{cases}\text{Yes} & \text{SOLAR-U} = \text{FS} \\ \text{No} & \text{SOLAR-U} = \emptyset\end{cases}$ [This paper]	Identical to bounded loss [This paper]

TABLE 2
Proposed learning rules for universal consistency (ME = Mean Estimation and EI = Empirical Integrability).[1]

Learning setting	Loss (and response/setting constraints)	Learning rule	Guarantees for which processes $\mathbb{X}$?	Optimist. universal?	Reference
I.i.d. responses	Finite or countable class., 01-loss	OptiNet	i.i.d.	No	[24]
	Real-valued regression + integrable	Proto-NN	i.i.d.	No	[20]
	Metric loss + integrable	MedNet	i.i.d.	No	[9]
Noiseless responses (realizable)	Bounded loss	2C1NN	SMV	Yes	[2]
	Unbounded loss	Memorization	FS	Yes	[4]
	Unbounded + EI	2C1NN	SMV	Yes	[This paper]
Adversarial (or arbitrary) responses	Bounded loss + metric loss	Hedge-variant	CS	Not always	[23]
	Bounded loss + F-TiME	$(1+\delta)$C1NN-hedged	SMV	Yes	[This paper]
	Bounded loss + not F-TiME	Hedge-variant 2	CS	Yes	[This paper]
	Unbounded loss + ME	ME-algorithm	FS	Yes	[This paper]
	Unbounded loss + not ME	N/A	$\emptyset$	N/A	[This paper]
	Unbounded + EI + local F-TiME	EI-$(1+\delta)$C1NN-hedged	SMV	Yes	[This paper]
	Unbounded + EI + not local F-TiME	EI-Hedge-variant	CS	Yes	[This paper]

Further, we show that for powers of metric we may have SOLAR-U $=\emptyset$. Specifically, for real-valued regression with Euclidean norm and loss $|\cdot|^\alpha$ and $\alpha > 1$, adversarial regression or mean estimation are not achievable. We then show that we have an alternative: either mean estimation with adversarial responses is achievable, SOLAR-U $=$ FS and we have an optimistically universal learning rule; or mean estimation is not achievable and SOLAR-U $=\emptyset$. Thus, even in the best case scenario for unbounded losses, SOLAR-U $=$ FS, which is already extremely restrictive. [4] asked whether imposing moment conditions on the responses would allow recovering the large set SMV as learnable processes instead. Specifically, they formulated the following question.

Question [4]: For unbounded losses ℓ, does there exist an online learning rule $f.$ which is consistent under every $\mathbb{X} \in$ SMV, for every measurable function $f^* : \mathcal{X} \to \mathcal{Y}$ such that there exists $y_0 \in \mathcal{Y}$ with $\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^T \ell(y_0, f^*(X_t)) < \infty$ (a.s.), i.e., such that we have $\mathcal{L}_{\mathbb{X}}(f., f^*) = 0$ (a.s.)?

We answer negatively to this question. Under this first-moment condition, universal learning under all SMV processes is not achievable even in this noiseless case. We show the stronger statement that noiseless universal learning under all processes having pointwise convergent relative frequencies—which are included in CS—is not achievable. However, under the empirical integrability condition introduced above we are able to recover all positive results from bounded losses.

[1]In our paper, an algorithm is optimistically universal if it is universally consistent for all processes under which universal learning is possible in the considered setting. OptiNet, Proto-NN, and MedNet are optimistically universal in another sense, their guarantees hold in all metric spaces for which universal learning with i.i.d. pairs

Table 1 and 2 summarize known results in the literature and our contributions. As a reminder, FS $\subset$ CS $\subset$ SMV in general, and FS $\subsetneq$ CS $\subsetneq$ SMV whenever $\mathcal{X}$ is infinite [22].

4. An optimistically universal learning rule for totally-bounded value spaces. We start our analysis of universal learning under adversarial responses with *totally-bounded* value spaces, for which we can give simple and explicit algorithms. Hence, we suppose in this section that the value space $(\mathcal{Y}, \ell)$ is totally-bounded, i.e., for any $\epsilon > 0$ there exists a finite $\epsilon-$net $\mathcal{Y}_\epsilon$ of $\mathcal{Y}$ such that for any $y \in \mathcal{Y}$, there exists $y' \in \mathcal{Y}_\epsilon$ with $\ell(y, y') < \epsilon$. In particular, a totally-bounded space is necessarily bounded and separable. The goal of this section is to show that for such value spaces, adversarial universal regression is achievable for all processes in SMV as in the noiseless setting (the empirical integrability assumption is automatically satisfied in this context). Further, we explicitly construct an optimistically universal learning rule for adversarial responses.

We recall that in the noiseless setting, the 2C1NN learning rule achieves universal consistency for all SMV processes [2]. At each iteration t, This rule performs the nearest neighbor rule over a restricted dataset instead of the complete history $\mathbb{X}_{\leq t-1}$. The dataset is updated by keeping track of the number of times each point X_u was used as nearest neighbor. This number is then capped at 2 by deleting from the current dataset any point which has been used twice as representative. Unfortunately, this learning rule is not optimistically universal for adversarial responses. More generally, [9] noted that any learning rule which only outputs observed historical values cannot be consistent, even in the simplest case of $\mathcal{X} = \{0\}$ and i.i.d. responses $\mathbb{Y}$. For instance, take $\mathcal{Y} = \bar{B}(0, 1)$ the closed ball of radius 1 in the plane $\mathbb{R}^2$ with the euclidean loss, consider the points $A, B, C \in \mathcal{Y}$ representing the equilateral triangle $e^{2ik\pi/3}$ for $k = 0, 1, 2$, and let $\mathbb{Y}$ be an i.i.d. process following the distribution which visits A, B or C with probability $\frac{1}{3}$. Predictions within observed values, i.e., A, B or C, incur an average loss of $\frac{2}{3}\sqrt{3} > 1$ where 1 is the loss obtained with the fixed value $(0, 0)$.

To construct an optimistically universal learning rule for adversarial responses, we first generalize a result from [2]. Instead of the 2C1NN learning rule, we use $(1+\delta)$C1NN rules for $\delta > 0$ arbitrarily small. Similarly as in 2C1NN, each new input X_t is associated to a representative $\phi(t)$ used for the prediction $\hat{Y}_t = Y_{\phi(t)}$. In the $(1+\delta)$C1NN rule, each point is used as a representative at most twice with probability δ and at most once with probability $1-\delta$. In order to have this behavior irrespective of the process $\mathbb{X}$, which can be thought of been chosen by a (limited) adversary within the SOUL processes, the information of whether a point can allow for 1 or 2 children is only revealed when necessary. Specifically, at any step $t \geq 1$, the algorithm initiates a search for a representative $\phi(t)$. It successively tries to use the nearest neighbor of X_t within the current dataset and uses it as a representative if allowed by the maximum number of children that this point can have. However, the information whether a potential representative u can have at most 1 or 2 children is revealed only when u already has one child.

- If u allows for 2 children, it will be used as final representative $\phi(t)$.
- Otherwise, u is deleted from the dataset and the search for a representative continues.

The rule is formally described in Algorithm 1, where $\bar{y} \in \mathcal{Y}$ is an arbitrary value, and the maximum number of children that a point X_t can have is represented by $1 + U_t$. In this formulation, all Bernouilli $\mathcal{B}(\delta)$ samples are drawn independently of the past history. Note that if $\delta = 1$, the $(1+\delta)$C1NN learning rule coincides with the 2C1NN rule of [2].

of instances and responses is achievable: *essentially separable* spaces $(\mathcal{X}, \rho_{\mathcal{X}})$ [24]. Our learning rules also enjoy this second optimistic property.

Input: Historical samples $(X_t, Y_t)_{t<T}$ and new input point X_T
Output: Predictions $\hat{Y}_t = (1+\delta)C1NN_t(\boldsymbol{X}_{<t}, \boldsymbol{Y}_{<t}, X_t)$ for $t \le T$
$\hat{Y}_1 := \bar{y}$ // Arbitrary prediction at $t=1$
$\mathcal{D}_2 \leftarrow \{1\}$; $n_1 \leftarrow 0$; // Initialisation
for $t = 2, \dots, T$ **do**
  if *exists* $u < t$ *such that* $X_u = X_t$ **then**
    $\hat{Y}_t := Y_u$
  else
    continue $\leftarrow True$ // Begin search for available representative $\phi(t)$
    while *continue* **do**
      $\phi(t) \leftarrow \min\left\{l \in \arg\min_{u \in \mathcal{D}_t} \rho_{\mathcal{X}}(X_t, X_u)\right\}$
      if $n_{\phi(t)} = 0$ **then** // Candidate representative has no children
        $\mathcal{D}_{t+1} \leftarrow \mathcal{D}_t \cup \{t\}$
        continue $\leftarrow False$
      else // Candidate representative has one child
        $U_{\phi(t)} \sim \mathcal{B}(\delta)$
        if $U_{\phi(t)} = 0$ **then**
          $\mathcal{D}_t \leftarrow \mathcal{D}_t \setminus \{\phi(t)\}$
        else
          $\mathcal{D}_{t+1} \leftarrow (\mathcal{D}_t \setminus \{\phi(t)\}) \cup \{t\}$
          continue $\leftarrow False$
    end
  end
  $\hat{Y}_t := Y_{\phi(t)}$
  $n_{\phi(t)} \leftarrow n_{\phi(t)} + 1$
  $n_t \leftarrow 0$
end

Algorithm 1: The $(1+\delta)$C1NN learning rule

THEOREM 4.1. *Fix $\delta > 0$. For any separable Borel space $(\mathcal{X}, \mathcal{B})$ and any separable near-metric output setting $(\mathcal{Y}, \ell)$ with bounded loss, in the noiseless setting, $(1+\delta)$C1NN is optimistically universal.*

We now construct our algorithm. This learning rule uses a collection of algorithms $f^{\epsilon}_{\cdot}$ which each yield an asymptotic error at most a constant factor from $\epsilon^{\frac{1}{\alpha+1}}$. Now fix $\epsilon > 0$ and let $\mathcal{Y}_\epsilon$ be a finite $\epsilon-$net of $\mathcal{Y}$ for ℓ. Recall that we denote by $\bar{\ell}$ the supremum loss. We pose

$$T_\epsilon := \left\lceil \frac{\bar{\ell}^2 \ln |\mathcal{Y}_\epsilon|}{2\epsilon^2} \right\rceil \quad \text{and} \quad \delta_\epsilon := \frac{\epsilon}{2T_\epsilon}.$$

The quantity T_ϵ will be the horizon window used by our learning rule to make its prediction using the $(1+\delta_\epsilon)$C1NN learning rule. Precisely, let ϕ be the representative function from the $(1+\delta_\epsilon)$C1NN learning rule. Note that this representative function $\phi(t)$ is defined only for times t where a new instance X_t is revealed, otherwise $(1+\delta_\epsilon)$C1NN uses simple memorization $\hat{Y}_t = Y_u$. For simplicity, we will denote by $\mathcal{N} = \{t : \forall u < t, X_u \neq X_t\}$ these times of new instances. For $t \in \mathcal{N}$, we denote by $d(t)$ the depth of time t within the graph constructed by $(1+\delta_\epsilon)$C1NN, and define the horizon $L_t = d(t) \mod T_\epsilon$. Intuitively, the learning rule $f^{\epsilon}_{\cdot}$ performs the classical Hedge algorithm [7] on clusters of times that are close within the graph ϕ. Precisely, we define the equivalence relation between times as follows:

$$t_1 \overset{\phi}{\sim} t_2 \quad \iff \quad \begin{cases} \phi^{L_{u_1}}(u_1) = \phi^{L_{u_2}}(u_2) & \text{and } |\{u < t_i : X_u = X_{t_i}\}| \le \frac{T_\epsilon}{\epsilon},\ i = 1,2 \\ \text{or} & \\ X_{t_1} = X_{t_2} & \text{and } |\{u < t_i : X_u = X_{t_1}\}| > \frac{T_\epsilon}{\epsilon},\ i = 1,2, \end{cases}$$

Input: Historical samples $(X_t, Y_t)_{t<T}$ and new input point X_T,
Representatives $\phi_\epsilon(\cdot)$ and depths $d_\epsilon(\cdot)$ constructed iteratively within $(1+\delta_\epsilon)$C1NN.
Output: Predictions $\hat{Y}_t(\epsilon) = f_t^\epsilon(\boldsymbol{X}_{<t}, \boldsymbol{Y}_{<t}, X_t)$ for $t \leq T$
$\mathcal{Y}_\epsilon$ an $\epsilon-$net of $\mathcal{Y}$
$T_\epsilon := \left\lceil \frac{\bar{\ell}^2 \ln|\mathcal{Y}_\epsilon|}{2\epsilon^2} \right\rceil, \quad \eta_\epsilon := \sqrt{\frac{8\ln|\mathcal{Y}_\epsilon|}{\bar{\ell}^2 T_\epsilon}}$
for $t = 1, \ldots, T$ **do**
$L_y^t = \sum_{u<t: u \overset{\phi_\epsilon}{\sim} t} \ell(Y_u, y), \quad y \in \mathcal{Y}_\epsilon$ `// Losses on the cluster given by` ϕ_ϵ
$p^t(y) = \frac{\exp(-\eta_\epsilon L_y^t)}{\sum_{z \in \mathcal{Y}_\epsilon} \exp(-\eta_\epsilon L_z^t)}, \quad y \in \mathcal{Y}_\epsilon$
$\hat{Y}_t \sim p^t$
end

Algorithm 2: The $f^\epsilon_\cdot$ learning rule

where $u_i = \min\{u : X_u = X_{t_i}\}$ is the first occurrence of the considered instance point X_{t_i}. Hence, multiple occurrences of the same instance value fall in the same cluster and for new instance points times $t \in \mathcal{N}$, all times of a given cluster share the same ancestor up to generation at most $T_\epsilon - 1$. Additionally, a cluster is dedicated to instance points that have a significant number of duplicates. To make its prediction at time t, $f^\epsilon_\cdot$ performs the Hedge algorithm based on values observed on its current cluster $\{u \leq t : u \overset{\phi}{\sim} t\}$. Let $\eta_\epsilon := \sqrt{\frac{8\ln|\mathcal{Y}_\epsilon|}{\bar{\ell}^2 T_\epsilon}}$ and define the losses $L_y^t = \sum_{u<t: u \overset{\phi}{\sim} t} \ell(Y_u, y)$. The learning rule $f_t^\epsilon(\mathbb{X}_{\leq t-1}, \mathbb{Y}_{\leq t-1}, X_t)$ outputs a random value in $\mathcal{Y}_\epsilon$ independently from the past history with

$$\mathbb{P}(\hat{Y}_t(\epsilon) = y) = \frac{e^{-\eta_\epsilon L_y^t}}{\sum_{z \in \mathcal{Y}_\epsilon} e^{-\eta_\epsilon L_z^t}}, \quad y \in \mathcal{Y}_\epsilon,$$

where, for simplicity, we denoted $\hat{Y}_t(\epsilon)$ the prediction given by the learning rule $f^\epsilon_\cdot$ at time t.

Having constructed the learning rules $f^\epsilon_\cdot$, we are now ready to define our final learning rule $f_\cdot$. Let $\epsilon_i = 2^{-i}$ for all $i \geq 0$. Intuitively, it aims to select the best prediction within the rules $f^{\epsilon_i}_\cdot$. If there were a finite number of such predictors, we could directly use the algorithms for learning with experts from the literature [7]. Instead, we introduce these predictors one at a time: at step $t \geq 1$ we only consider the indices $I_t := \{i \leq \ln t\}$. We then compute an estimate $\hat{L}_{t-1,i}$ of the loss incurred by each predictor $f^{\epsilon_i}_\cdot$ for $i \in I_t$ and select a random index $\hat{i}_t$ independent from the past history from an exponentially-weighted distribution based on the estimates $\hat{L}_{t-1,i}$. The final output of our learning rule is $\hat{Y}_t := \hat{Y}_t(\epsilon_{\hat{i}})$. The complete algorithm is formally described in Algorithm 3. The following lemma quantifies the loss of the rule $f_\cdot$ compared to the best rule $f^{\epsilon_i}_\cdot$.

LEMMA 4.2. *Almost surely, there exists $\hat{t} \geq 0$ such that*

$$\forall t \geq \hat{t}, \forall i \in I_t, \quad \sum_{s=t_i}^{t} \ell(\hat{Y}_t, Y_t) \leq \sum_{s=t_i}^{t} \ell(\hat{Y}_t(\epsilon_i), Y_t) + (2 + \bar{\ell} + \bar{\ell}^2)\sqrt{t \ln t}.$$

We are now ready to show that Algorithm 3 is universally consistent under SMV processes.

THEOREM 4.3. *Suppose that $(\mathcal{Y}, \ell)$ is totally-bounded. There exists an online learning rule $f_\cdot$ which is universally consistent for adversarial responses under any process $\mathbb{X} \in$ SMV(= SOUL), i.e., for any process $(\mathbb{X}, \mathbb{Y})$ on $(\mathcal{X}, \mathcal{Y})$ with adversarial response, such that $\mathbb{X} \in$ SMV, then for any measurable function $f : \mathcal{X} \to \mathcal{Y}$, we have $\mathcal{L}_{(\mathbb{X},\mathbb{Y})}(f_\cdot, f) \leq 0, \quad (a.s.)$.*

Input: Historical samples $(X_t, Y_t)_{t<T}$ and new input point X_T,
Predictions $\hat{Y}_\cdot(\epsilon_i)$ from the learning rules $f_\cdot^{\epsilon_i}$.
Output: Predictions $\hat{Y}_t$ for $t \leq T$
$w_{0,0} = 1, t_i := \lceil e^i \rceil, \quad i \geq 0$
$I_t = \{i \leq \ln t\}$, $\eta_t = \sqrt{\frac{\ln t}{t}}, \quad t \geq 1$
for $t = 1, \ldots, T$ **do**
$L_{t-1,i} := \sum_{s=t_i}^{t-1} \ell(\hat{Y}_s(\epsilon_i), Y_s)$, $\hat{L}_{t-1,i} := \sum_{s=t_i}^{t-1} \hat{\ell}_s, \quad i \in I_t$
$w_{t-1,i} = e^{\eta_t(\hat{L}_{t-1,i} - L_{t-1,i})}$
$p_t(i) = \dfrac{w_{t-1,i}}{\sum_{j \in I_t} w_{t-1,j}}$
$\hat{i}_t \sim p_t(\cdot)$ `// model selection`
$\hat{Y}_t = \hat{Y}_t(\epsilon_i)$
$\hat{\ell}_t := \dfrac{\sum_{i \in I_t} w_{t-1,i} \ell(\hat{Y}_t(\epsilon_i), Y_t)}{\sum_{i \in I_t} w_{t-1,i}}$.
end

Algorithm 3: An optimistically universal learning rule for totally bounded spaces

Proof sketch. First observe that Lemma 4.2 allows us to combine predictors $f_\cdot^\epsilon$: if individually they perform well, Algorithm 3 achieves the best long-term average excess loss among them. We then proceed to show that $f_\cdot^\epsilon$ has low average error in the long run. First, $(1+\delta_\epsilon)$C1NN is universally consistent on SMV processes in the noiseless setting by Theorem 4.1. This intuitively shows that for noiseless functions, the value at time $\phi_\epsilon(t)$ provides a good representative for the value at time t. Extrapolating this argument, we show that if two times are close (for the graph metric) within the graph formed by ϕ_ϵ, they will have close values for any fixed function in the long run. As a result, times in the same cluster defined by $\overset{\phi_\epsilon}{\sim}$ share similar values in the long run. The $f_\cdot^\epsilon$ rule precisely aims to learn the best predictor by cluster using the classical Hedge algorithm. Because it can only ensure low regret compared to a finite number of options, we use ϵ-nets of the value space $\mathcal{Y}$. The reason why we need to have $(1+\delta_\epsilon)$C1NN instead of the known 2C1NN algorithm is that for a given time T, we need to ensure low excess error even though some clusters might not be completed. Because the tree formed by ϕ_ϵ resembles a $(1+\delta_\epsilon)$-branching process, the fraction of times which belong to unfinished clusters is only a small fraction ϵT of the T times, hence does not affect the average long-term excess error significantly. Altogether, we show that $f_\cdot^\epsilon$ has $\mathcal{O}(\epsilon^{\frac{1}{\alpha+1}})$ long-term average excess error compared to any fixed function for any SMV process, which ends the proof.

As a result, SMV $\subset$ SOLAR for totally-bounded value spaces. Recalling that for bounded values SMV $=$ SOUL [2], i.e., processes $\mathbb{X} \notin$ SMV are not universally learnable even in the noiseless setting, we have SOLAR $\subset$ SMV. Thus we obtain a complete characterization of the processes which admit universal learning with adversarial responses: SOLAR $=$ SMV. Further, the proposed learning rule is optimistically universal for adversarial regression.

COROLLARY 4.4. *Suppose that $(\mathcal{Y}, \ell)$ is totally-bounded. Then, SOLAR $=$ SMV, and there exists an optimistically universal learning rule for adversarial regression, i.e., which achieves universal consistency with adversarial responses under any process $\mathbb{X} \in$ SOLAR.*

This is a first step towards the more general Corollary 5.5. Indeed, one can note that F-TiME is satisfied by any totally-bounded value space: given a fixed error tolerance $\eta > 0$, consider a finite $\frac{\eta}{2}-$net $\mathcal{Y}_{\eta/2}$ of $\mathcal{Y}$. Because this is a finite set, we can perform the classical Hedge algorithm [7] to have $\Theta(\sqrt{T \ln |\mathcal{Y}_{\eta/2}|})$ regret compared to the best fixed value of $\mathcal{Y}_{\eta/2}$. For example, if $\alpha = 1$, posing $T_\eta = \Theta(\frac{4}{\eta^2} \ln |\mathcal{Y}_{\eta/2}|)$ enables to have a regret at most $\frac{\eta}{2} T_\eta$ compared to any fixed value of $\mathcal{Y}_{\eta/2}$, hence regret at most ηT_η compared to any value of $\mathcal{Y}$. This achieves F-TiME, taking a deterministic time $\tau_\eta := T_\eta$.

5. Characterization of learnable processes for bounded losses. While Section 4 focused on totally-bounded value spaces, the goal of this section is to give a full characterization of the set SOLAR of processes for which adversarial regression is achievable and provide optimistically universal algorithms, for any *bounded* value space.

5.1. *Negative result for non-totally-bounded spaces.* Although for all bounded value spaces $(\mathcal{Y}, \ell)$, noiseless universal learning is achievable on all SMV($=$ SOUL) processes, this is not the case for adversarial regression in non-totally-bounded spaces. We show in this section that extending Corollary 4.4 to any bounded value space is impossible: the set of learnable processes for adversarial regression may be reduced to CS only, instead of SMV.

THEOREM 5.1. *Let $(\mathcal{X}, \mathcal{B})$ a separable Borel metrizable space. There exists a separable metric value space $(\mathcal{Y}, \ell)$ with bounded loss such that the following holds: for any process $\mathbb{X} \notin$ CS, universal learning under $\mathbb{X}$ for arbitrary responses is not achievable. Precisely, for any learning rule $f_\cdot$, there exists a process $\mathbb{Y}$ on $\mathcal{Y}$, a measurable function $f^* : \mathcal{X} \to \mathcal{Y}$ and $\epsilon > 0$ such that with non-zero probability $\mathcal{L}_{(\mathbb{X},\mathbb{Y})}(f_\cdot, f^*) \geq \epsilon$.*

In the proof, we explicitly construct a bounded metric space that does not satisfy F-TiME. More precisely, we choose $\mathcal{Y} = \mathbb{N} = \{i \geq 0\}$ and a specific metric loss ℓ with values in $\{0, \frac{1}{2}, 1\}$. For any $k \geq 1$, we pose $n_k := 2k(k-1) + 2^k - 1$ and define the sets

$$I_k := \{n_k, n_k + 1, \ldots, n_k + 4k - 1\} \quad \text{and} \quad J_k := \{n_k + 4k, n_k + 4k + 1, \ldots, n_{k+1} - 1\}.$$

These sets are constructed so that $|I_k| = 4k$, $|J_k| = 2^k$ for all $k \geq 1$, and together with $\{0\}$, they form a partition of $\mathbb{N}$. We now construct the loss ℓ. We pose $\ell(i,j) = \mathbb{1}_{i=j}$ for all $i, j \in \mathbb{N}$ unless there is $k \geq 1$ such that $(i,j) \in I_k \times J_k$ or $(j,i) \in I_k \times J_k$. It now remains to define the loss $\ell(i,j)$ for all $i \in I_k$ and $j \in J_k$. Note that for any $j \in J_k$, we have that $j - n_k - 4k \in \{0, \ldots, 2^k - 1\}$. Hence we will use their binary representation which we write as $j - n_k - 4k = \{b_j^{k-1} \ldots b_j^1 b_j^0\}_2 = \sum_{u=0}^{k-1} b_j^u 2^u$ where $b_j^0, b_j^1, \ldots, b_j^{k-1} \in \{0, 1\}$ are binary digits. Finally, we pose

$$\ell(n_k + 4u, j) = \ell(n_k + 4u + 1, j) = \frac{1 + b_j^u}{2},$$
$$\ell(n_k + 4u + 2, j) = \ell(n_k + 4u + 3, j) = \frac{2 - b_j^u}{2},$$

for all $u \in \{0, 1, \ldots, k-1\}$ and $j \in J_k$.

Proof sketch. This value space does not belong to F-TiME because for any algorithm and horizon time k, there is a sequence of length k of elements in I_k with $y_u = n_k + 4(u-1) + 2b_u + c_u$ for $1 \leq u \leq k$ and $b_u, c_u \in \{0, 1\}$, such that the algorithm incurs an average excess loss $\frac{1}{4}$ per iteration compared to some fixed element of J_k. To find such a sequence, we sample randomly and independently Bernoulli variables $b_u, c_u \sim \mathcal{B}(\frac{1}{2})$. In hindsight, the best predictor of the sequence is $n_k + 4k + j$, where $j = b_1 \cdots b_k$ in binary representation. However, the algorithm only observes these bits in an online fashion: at time t it incurs an excess loss cost if it guesses an element of I_k because it has probability at most $\frac{1}{4}$ of finding y_t. And if it predicts an element of J_k, it cannot know in advance the correct t-th bit to choose in their binary representation.

We then proceed to show that for this space SOLAR = CS $\subsetneq$ SOUL. To do so, we show that for processes $\mathbb{X} \notin$ CS there exists a sequence of disjoint measurable sets $\{B_p\}_{p\geq 1}$ and increasing times $(t_p)_{p\geq 1}$ and $\epsilon > 0$ such that with non-zero probability,

$$\forall p \geq 1, \quad \mathbb{X}_{\leq t_{p-1}} \cap B_p = \emptyset \text{ and } \exists t_{p-1} < t \leq t_p : \frac{1}{t}\sum_{t'=1}^{t} \mathbb{1}_{B_p}(X_{t'}) \geq \epsilon.$$

On this event, an online algorithm does not receive any information for instances in B_p before time t_{p-1}. We then construct responses by $(t_{p-1}, t_p]$. During this period and for contexts in B_p, we choose the same difficult-to-predict sequence of values as above for $k = t_p - t_{p-1}$. On the other hand, because the sets B_p are disjoint, there exists a measurable function f^* that selects the best action in hindsight for each set B_p. Intuitively, within horizon t_p, the algorithm cannot gather enough information to achieve lower average excess error than $\frac{\epsilon}{4}$ compared to f^*, which shows that it is not universally consistent.

Although learning beyond CS is impossible in this case, there still exists an optimistically universal learning rule for adversarial responses. Indeed, the main result of [23] shows that for any bounded value space, there exists a learning rule which is consistent under all CS processes for arbitrary responses (when ℓ is a metric, i.e., $\alpha = 1$).

THEOREM 5.2 ([23]). *Suppose that $(\mathcal{Y}, \ell)$ is metric and ℓ is bounded. Then, there exists an online learning rule $f_\cdot$ which is universally consistent for arbitrary responses under any process $\mathbb{X} \in$ CS, i.e., such that for any stochastic process $(\mathbb{X}, \mathbb{Y})$ on $(\mathcal{X}, \mathcal{Y})$ with $\mathbb{X} \in$ CS, then for any measurable function $f : \mathcal{X} \to \mathcal{Y}$, we have $\mathcal{L}_{(\mathbb{X},\mathbb{Y})}(f_\cdot, f) \leq 0, \quad (a.s.)$.*

The proof of this theorem given in [23] extends to adversarial responses. However, we defer the argument because we will later prove Theorem 3.3 which also holds for any loss $\ell = \rho_{\mathcal{Y}}^{\alpha}$ for $\alpha \geq 1$ and unbounded losses in Section 7. This shows that for any separable metric space $(\mathcal{X}, \rho_{\mathcal{X}})$, there exists a metric value space for which the learning rule proposed in [23] was already optimistically universal.

5.2. *Adversarial regression for classification with a countable number of classes.* Although we showed in the last section that adversarial regression under all SMV processes is not achievable for some non-totally-bounded spaces, we will show that there exist non-totally-bounded value spaces for which we can recover SOLAR = SMV. Precisely, we consider the case of classification with countable number of classes $(\mathbb{N}, \ell_{01})$, with $0-1$ loss $\ell_{01}(i,j) = \mathbb{1}_{i\neq j}$. The goal of this section is to prove that in this case, we can learn arbitrary responses under any SOUL process. The main difficulty with non-totally-bounded classification is that we cannot apply traditional online learning tools because $\epsilon-$nets may be infinite. Hence, we first show a result that allows us to perform online learning with an infinite number of experts in the context of countable classification.

LEMMA 5.3. *Let $t_0 \geq 1$. There exists an online learning rule $f_\cdot$ such that for any sequence $\boldsymbol{y} := (y_i)_{i\geq 1}^T$ of values in $\mathbb{N}$, we have that for $T \geq t_0$*

$$\sum_{t=1}^{T} \mathbb{E}[\ell_{01}(f_t(\boldsymbol{y}_{\leq t-1}), y_t)] \leq \min_{y\in\mathbb{N}} \sum_{t=1}^{T} \ell_{01}(y, y_t) + 1 + \ln 2\sqrt{\frac{t_0}{2\ln t_0}} + \sqrt{\frac{\ln t_0}{2t_0}}(t_0 + T),$$

and with probability $1-\delta$,

$$\sum_{t=1}^{T}\mathbb{E}[\mathbb{1}_{f_t(\boldsymbol{y}_{\leq t-1})=y_t}] \geq \max_{y\in\mathbb{N}}\sum_{t=1}^{T}\mathbb{1}_{y=y_t} - 1 - \ln 2\sqrt{\frac{t_0}{2\ln t_0}} - \sqrt{\frac{\ln t_0}{2t_0}}(t_0+T) - \sqrt{2T\ln\frac{1}{\delta}}.$$

Proof sketch. We adapt the classical Hedge algorithm, which in its standard form can only ensure sublinear regret compared to a fixed set of values. Instead, we only consider a small subset of candidate values that is enlarged occasionally with previously observed values $y \in \mathbb{Y}_{\leq t}$. This formalizes the intuition that even though there are a priori an infinite number of candidate values ($\mathbb{N}$), it is reasonable to only focus on values with high frequency in the observed sequence $\mathbb{Y}_{\leq t}$: if the next value y_{t+1} is not in this set, the algorithm incurs a loss 1, which would also be incurred by the best fixed predictor until time $t+1$ in hindsight.

We can therefore adapt the learning rules $f^\epsilon_\cdot$ from Section 4 by replacing the Hedge algorithm with the algorithm from Lemma 5.3. Further adapting parameters, we obtain our main result for countable classification.

THEOREM 5.4. *Let $(\mathcal{X},\mathcal{B})$ be a separable Borel metrizable space. There exists an online learning rule $f_\cdot$ which is universally consistent for adversarial responses under any process $\mathbb{X}\in$ SMV for countable classification, i.e., such that for any adversarial process $(\mathbb{X},\mathbb{Y})$ on $(\mathcal{X},\mathbb{N})$ with $\mathbb{X}\in$ SMV, for any measurable function $f^*:\mathcal{X}\to\mathbb{N}$, we have that $\mathcal{L}_{(\mathbb{X},\mathbb{Y})}(f_\cdot,f^*)\leq 0, \quad (a.s.)$.*

5.3. *A complete characterization of universal regression on bounded spaces.* The last two Sections 5.1 and 5.2 gave examples of non-totally-bounded value spaces for which we obtain respectively SOLAR $=$ CS or SOLAR $=$ SMV. In this section, we prove that there is an underlying alternative, defined by F-TiME, which enables us to precisely characterize the set SOLAR of learnable processes for adversarial regression.

When F-TiME is satisfied by the value space, similarly to the case of countable classification, we recover SOLAR $=$ SMV and there exists an optimistically universal rule. The corresponding algorithm follows the same general structure as the learning rule provided in Section 4 for totally-bounded-spaces, however, the learning rules $f^\epsilon_\cdot$ need to be significantly modified. First, the Hedge algorithm should be replaced by the learning rule $g_{\leq t_\epsilon}$ provided by the F-TiME property. Second, as the horizon time t_ϵ of this learning rule is bounded, the clusters of points on which it is applied have to be adapted: we cannot simply use clusters by distance in the graph defined by the $(1+\delta_\epsilon)C1NN$ algorithm. Instead, we construct clusters of smaller size t_ϵ among these larger graph-based clusters.

More precisely, we take the horizon time t_ϵ and the learning rule $g^\epsilon_{\leq t_\epsilon}$ satisfying the condition imposed by the assumption on $(\mathcal{Y},\ell)$. Then, let $T_\epsilon = \lceil \frac{t_\epsilon}{\epsilon}\rceil$. Similarly as before, we then define $\delta_\epsilon := \frac{\epsilon}{2T_\epsilon}$ and let ϕ be the representative function from the $(1+\delta_\epsilon)$C1NN learning rule. Then, we introduce the same equivalence relation between times $\overset{\phi}{\sim}$, which induces clusters of times. We define a sequence of i.i.d. copies $g^{\epsilon,t}_\cdot$ of the learning rule $g^\epsilon_\cdot$ for all $t\geq 1$. This means that the randomness used within these learning rules is i.i.d, and the copy $g^{\epsilon,t}_\cdot$ should be sampled only at time t, independently of the past history. Predictions are then made by blocks of size t_ϵ within the same cluster: at time t, let $u_1<\ldots<u_{L_t}<t$ be the elements of the current block. If the block does not contain t_ϵ elements yet, we use $g^{\epsilon,u_1}_{L_t+1}$ for the prediction at time t. Otherwise, we start a new block and use $g^{\epsilon,t}_1$. Hence, letting $\psi(t)=\max\mathcal{C}(t)$ be the last time in the same cluster as t (as defined by ϕ_ϵ) and L_t the size of the current block of t without counting t, we now define the learning rule $f^\epsilon_\cdot$ such that for any sequence $\boldsymbol{x}$, $\boldsymbol{y}$,

$$f^\epsilon_t(\boldsymbol{x}_{\leq t-1},\boldsymbol{y}_{\leq t-1},x_t) := g^{\epsilon,\psi^{L_t}(t)}_{L_t+1}\left(\{y_{\psi^{L_t+1-u}(t)}\}_{u=1}^{L_t}\right).$$

Input: Historical samples $(X_t, Y_t)_{t<T}$ and new input point X_T,
Learning rule for finite-time mean estimation $g^{\epsilon}_{\leq t_\epsilon}$, $T_\epsilon = \lceil \frac{t_\epsilon}{\epsilon} \rceil$, $\delta_\epsilon := \frac{\epsilon}{2T_\epsilon}$.
Representatives $\phi_\epsilon(\cdot)$ constructed iteratively within $(1+\delta_\epsilon)$C1NN.
Output: Predictions $\hat{Y}_t(\epsilon) = f^\epsilon_t(\boldsymbol{X}_{<t}, \boldsymbol{Y}_{<t}, X_t)$ for $t \leq T$
for $t = 1, \ldots, T$ **do**
  $\mathcal{C}(t) = \{u < t : u \overset{\phi_\epsilon}{\sim} t\}$
  if $\mathcal{C}(t) = \emptyset$ **then** $L_t = 0$ and initialize learner $g^{\epsilon,t}_\cdot$;
  else
    $\psi(t) = \max \mathcal{C}(t)$
    if $L_{\psi(t)} < t_\epsilon - 1$ **then** $L_t = L_{\psi(t)} + 1$;
    else $L_t = 0$ and initialize learner $g^{\epsilon,t}_\cdot$;
  end
  $\hat{Y}_t = g^{\epsilon, \psi^{L_t}(t)}_{L_t+1}\left(\{y_{\psi^{L_t+1-u}(t)}\}_{u=1}^{L_t}\right)$
end

Algorithm 4: The modified $f^\epsilon_\cdot$ learning rule for value spaces $(\mathcal{Y}, \ell)$ satisfying F-TiME. When initializing a learner $g^{\epsilon,t}_\cdot$ for finite-time mean estimation, its internal randomness is sampled independently from the past.

The complete learning rule is given in Algorithm 4. The learning rules $f^\epsilon_\cdot$ are then combined into a single learning rule as in the original algorithm for totally-bounded spaces, following the same procedure given in Algorithm 3. We then show that it is universally consistent under SMV processes using same arguments as for Theorem 4.3.

THEOREM 5.5. *Suppose that ℓ is bounded and $(\mathcal{Y}, \ell)$ satisfies F-TiME. Then, SOLAR = SMV(= SOUL) and there exists an optimistically universal learning rule for adversarial regression, i.e., which achieves universal consistency with adversarial responses under any process $\mathbb{X} \in$ SMV.*

We are now interested in value spaces $(\mathcal{Y}, \ell)$ which do not satisfy F-TiME. We will show that in this case, SOLAR is reduced to the processes CS. We first introduce a second property on value spaces as follows.

Property 2: *For any $\eta > 0$, there exists a horizon time $T_\eta \geq 1$ and an online learning rule $g_{\leq \tau}$ where τ is a random time with $1 \leq \tau \leq T_\eta$ such that for any $\boldsymbol{y} := (y_t)_{t=1}^{T_\eta}$ of values in $\mathcal{Y}$ and any value $y \in \mathcal{Y}$, we have*

$$\mathbb{E}\left[\frac{1}{\tau}\sum_{t=1}^{\tau}\left(\ell(g_t(\boldsymbol{y}_{\leq t-1}), y_t) - \ell(y, y_t)\right)\right] \leq \eta.$$

REMARK 5.6. The random time τ may depend on the possible randomness of the learning rule $g_\cdot$, but it does not depend on any of the values $y_1, y_2, \ldots$ on which the learning rule $g_\cdot$ may be tested. Intuitively, the learning rule uses some randomness which is first privately sampled and may be used by τ. This randomness is never explicitly revealed to the adversary choosing the values $\boldsymbol{y}$, but only implicitly through the realizations of the predictions.

LEMMA 5.7. *Property F-TiME is equivalent to Property 2.*

Using this second property, we can then show that when F-TiME is not satisfied, universal consistency outside CS under adversarial responses is not achievable. In the proof, we only use stochastic processes $(\mathbb{X}, \mathbb{Y})$, hence the same result holds if we only considered universal consistency under arbitrary responses.

THEOREM 5.8. *Suppose that ℓ is bounded and $(\mathcal{Y}, \ell)$ does not satisfy F-TiME. Then, SOLAR = CS and there exists an optimistically universal learning rule for adversarial regression, i.e., which achieves universal consistency with adversarial responses under any process $\mathbb{X} \in$ CS.*

Proof sketch. First, from Theorem 5.2 we already have CS $\subset$ SOLAR. The main difficulty is to prove that one cannot universally learn any process $\mathbb{X} \notin$ CS. To do so, we re-use the property derived in the proof of Theorem 5.1 that for non-CS processes, one can find a disjoint sequence of sets $\{B_p\}_{p\geq 1}$, an increasing times $(t_p)_{p\geq 1}$ and $\epsilon > 0$ such that with non-zero probability for all $p \geq 1$, the process $\mathbb{X}$ never visits B_p before time t_{p-1} and at some point between times $t_{p-1}+1$ and t_p, the set B_p has been visited a proportion ϵ of times. Now $(\mathcal{Y}, \ell)$ does not satisfy F-TiME, hence does not satisfy Property 2 by Lemma 5.7 for some constant $\eta > 0$. Then, for $p \geq 1$, during period $(t_{p-1}, t_p]$, we define the values $\mathbb{Y}_{t_{p-1} < \cdot \leq t_p}$ when the instance process visits B_p as a sequence $\boldsymbol{y}_{t_{p-1} < \cdot \leq t_p}$ such that the algorithm has average excess loss at least η whenever $\mathbb{X}$ visits B_p, compared to a fixed value $y_p^* \in \mathcal{Y}$. We note that the randomized version of F-TiME given by Lemma 5.7 is important because we do not know in advance when, between t_{p-1} and t_p, B_p has been visited a fraction ϵ of times: potentially, this time is random and there is a huge gap (exponential or more) between t_{p-1} and t_p. On the constructed stochastic process $\mathbb{Y}$, the algorithm does not have vanishing average excess loss compared to the function equal to y_p^* on B_p. This proves that no algorithm is universally consistent on $\mathbb{X}$.

This completes the proof of Corollary 3.5 and closes our study of universal learning with adversarial responses for bounded value spaces. Notably, there always exists an optimistically universal learning rule, however, this rule highly depends on the value space.

- If $(\mathcal{Y}, \ell)$ satisfies F-TiME, we can learn all SMV = SOUL processes. The proposed learning rule of Theorem 5.5 is *implicit* in general. Indeed, to construct it one first needs to find an online learning rule for mean estimation with finite horizon as described by property F-TiME, which is then used as a subroutine in the optimistically universal learning rule for adversarial regression. We showed however that for totally-bounded value spaces, this learning rule can be *explicited* using $\epsilon-$nets.
- If the value space does not satisfy F-TiME, we can only learn CS processes and there is an inherent gap between noiseless online learning and regression. We propose a learning rule in Section 7 which is optimistically universal—see Theorem 3.3. This rule is inspired by the proposed algorithm of [23] which is optimistically universal for metric losses $\alpha = 1$.

These two classes of learning rules use very different techniques. Specifically, under processes $\mathbb{X} \in$ CS, [22] showed that there exists a countable set $\mathcal{F}$ of measurable functions $f : \mathcal{X} \to \mathcal{Y}$ which is "dense" within the space of all measurable functions along the realizations $f(X_t)$. We refer to Section 7 for a precise description of this density notion. Hence, under process $\mathbb{X}$, we can approximate f^* by functions in $\mathcal{F}$ with arbitrary long-run average precision. However, such property is impossible to obtain for any process $\mathbb{X} \in$ SMV $\setminus$ CS: no process $\mathbb{X} \notin$ CS admits a "dense" countable sequence of measurable functions. Thus, to learn processes SMV for value spaces satisfying F-TiME, a fundamentally different learning rule than that proposed by [22] or [23] was needed.

Input: Historical samples $(Y_t)_{t<T}$
Output: Predictions $\hat{Y}_t$ for $t \leq T$
$(y^i)_{i\geq 0}$ dense sequence in $\mathcal{Y}$
$I_t := \{i \leq \ln t : \ell(y^0, y^i) \leq \ln t\}, \eta_t := \frac{1}{4\sqrt{t}}, t \geq 1; \quad t_i = \lceil \max(e^i, e^{\ell(y^0,y^i)}) \rceil, i \geq 0$
$w_{0,0} := 1, \quad \hat{Y}_1 = y^0$ // Initialisation
for $t = 2, \ldots, T$ **do**
  $L_{t-1,i} = \sum_{s=t_i}^{t-1} \ell(y^i, Y_s), \quad \hat{L}_{t-1,i} = \sum_{s=t_i}^{t-1} \hat{\ell}_s, \quad i \in I_t$
  $w_{t-1,i} := \exp(\eta_t(\hat{L}_{t-1,i} - L_{t-1,i})), \quad i \in I_t$
  $p_t(i) = \frac{w_{t-1,i}}{\sum_{j\in I_t} w_{t-1,j}}, \quad i \in I_t$
  $\hat{Y}_t \sim p_t(\cdot)$ // Prediction
  $\hat{\ell}_t := \frac{\sum_{j\in I_t} w_{t-1,j}\ell(y^j, Y_t)}{\sum_{j\in I_t} w_{t-1,j}}$
end

Algorithm 5: The mean estimation algorithm.

6. Adversarial universal learning for unbounded losses. We now turn to the case of unbounded losses, i.e., value spaces $(\mathcal{Y}, \ell)$ with $\bar{\ell} = \infty$. In this section, we consider universal learning without empirical integrability constraints, for which we introduced the notation SOLAR-U as the set of processes that admit universal learning (we recall that for bounded losses such distinction was unnecessary). In this case, and for more general near-metrics, [4] showed that SOUL $=$ FS. In other terms, for unbounded losses, the learnable processes in the noiseless setting necessarily visit a finite number of distinct instance points of $\mathcal{X}$ almost surely. Thus, universal learning on unbounded value spaces is very restrictive and in particular, SOLAR-U $\subset$ FS. We will show that either SOLAR-U $=$ FS or SOLAR-U $= \emptyset$.

6.1. *Adversarial regression for metric losses.* In this section, we focus on metric losses ℓ, i.e., $\alpha = 1$. In this case, we show that we always have the equality SOLAR-U $=$ FS and that we can provide an optimistically universal learning rule. To do so, we first consider the fundamental estimation problem where one observes values $\mathbb{Y}$ from a general separable metric value space and aims to sequentially predict a value $\hat{Y}_t$ in order to minimize the long-run average loss. We refer to this problem as the mean estimation problem, which is equivalent to regression for the instance space $\mathcal{X} = \{0\}$. For instance, in the specific case of i.i.d. processes $\mathbb{Y}$, mean estimation is exactly the problem of Fréchet mean estimation for distributions on $\mathcal{Y}$. We show that even for adversarial processes $\mathbb{Y}$, we can achieve sublinear regret compared to the best single value prediction, even for unbounded value spaces $(\mathcal{Y}, \ell)$.

If the space were finite, then we could use traditional Hedge algorithms [7]. Instead, given a separable value space, we have access to a dense countable sequence of values. We then select the best prediction among this dense sequence by introducing the values of the sequence one at a time, similarly to the argument we used in Lemma 4.2. The learning rule for mean estimation is described in Algorithm 5.

THEOREM 3.6. *Let $(\mathcal{Y}, \ell)$ be a separable metric space. There exists an online learning rule $f_\cdot$ that is universally consistent for adversarial mean estimation, i.e., for any adversarial process $\mathbb{Y}$ on $\mathcal{Y}$, almost surely, for all $y \in \mathcal{Y}$,*

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \left(\ell(f_t(\mathbb{Y}_{\leq t-1}), Y_t) - \ell(y, Y_t)\right) \leq 0.$$

REMARK 6.1. The above result guarantees that on the same event of probability one, the proposed learning rule achieves sublinear regret compared to any fixed value prediction. This was not the case for universal regression where, instead, for every fixed measurable function $f : \mathcal{X} \to \mathcal{Y}$, with probability one our learning rules achieved sublinear regret. This stems essentially from the fact that there exists a dense countable set of values $\mathcal{Y}$, but in general, there does not exist a countable set of measurable functions which are dense within all measurable functions in infinity norm.

We now return to the general regression problem on unbounded spaces. A simple learning rule would be to run in parallel the learning rule g_x for mean estimation on each distinct observed $x \in \mathcal{X}$, i.e., on the sub-process $\mathbb{Y}_{\{t:X_t=x\}}$. As a consequence of Theorem 3.6 we can show that this learning rule is universally consistent on FS processes.

COROLLARY 6.2. *Suppose that $(\mathcal{Y}, \ell)$ is an unbounded metric space. Then, SOLAR-U $=$ FS$(=$ SOUL$)$ and there exists an optimistically universal learning rule for adversarial regression, i.e., which achieves universal consistency with adversarial responses under any process $\mathbb{X} \in$ FS.*

6.2. *Negative result for real-valued adversarial regression with loss $\ell = |\cdot|^\alpha$ with $\alpha > 1$.* Unfortunately, one cannot extend Corollary 6.2 to losses that are powers of metrics in general. Even in the classical setting of real-valued regression $\mathcal{Y} = \mathbb{R}$ with Euclidean norm, we show that adversarial regression with any loss $\ell = |\cdot|^\alpha$ for $\alpha > 1$ is not achievable, i.e., SOLAR-U $= \emptyset$.

THEOREM 6.3. *Let $\alpha > 1$. For the Euclidean value space $(\mathbb{R}, |\cdot|)$ and loss $\ell = |\cdot|^\alpha$ we obtain SOLAR-U $= \emptyset$. In particular, there does not exist a consistent learning rule for mean estimation on $\mathbb{R}$ with squared loss for adversarial responses.*

Proof sketch. The reason why mean estimation with adversarial responses is impossible for $\alpha > 1$ but possible for $\alpha = 1$ is that for $\alpha > 1$, predicting a value off by 1 unit of the best value in hindsight can yield unbounded excess loss for that specific prediction. In particular, we consider a sequence of values of the form $Y_t^{\boldsymbol{b}} = M_t b_t$ where $(M_t)_{t\geq 1}$ is a fixed sequence growing super-exponentially in t, and $\boldsymbol{b} = (b_t)$ is an i.i.d. Rademacher random variables in $\{\pm 1\}$. The sequence $(M_t)_{t\geq 1}$ is constructed so that if the prediction $\hat{Y}_t$ and true value Y_t have different signs $\hat{Y}_t \cdot Y_t \leq 0$, the excess loss of the algorithm compared to the value $sign(Y_t^{\boldsymbol{b}}) = sign(b_t)$ is (super-)linear in t. Because the algorithm cannot know in advance the sign of b_t, there is a realization in which it makes an infinite number of mistakes and as a result has non-zero long-term excess loss compared to the value 1 or -1.

The above of this result also shows that the same negative result holds more generally for unbounded metric value spaces which have some "symmetry". The main ingredients for this negative result were having a point from which there exist arbitrary far values from symmetric directions. In particular, this holds for a discretized value space $(\mathbb{N}, |\cdot|)$ with Euclidean metric, and any Euclidean space $\mathbb{R}^d$ with $d \geq 1$.

6.3. *An alternative for adversarial regression with unbounded losses.* In the two previous sections, we gave examples of losses for which SOLAR-U $= \emptyset$ or SOLAR-U $=$ FS. The following simple result is that this is the only alternative and that SOLAR-U $=$ FS is equivalent to achieving consistency for mean estimation with adversarial responses.

PROPOSITION 6.4. *Let $(\mathcal{Y},\rho_{\mathcal{Y}})$ be a separable metric value space. Suppose that there exists an online learning rule $g.$ which is consistent for mean estimation with adversarial responses for the loss $\ell=\rho_{\mathcal{Y}}^{\alpha}$, where $\alpha\geq 1$, i.e., for any adversarial process $\mathbb{Y}$ on $(\mathcal{Y},\ell)$, we have for any $y^*\in\mathcal{Y}$,*

$$\limsup \frac{1}{T}\sum_{t=1}^{T}\left(\ell(f_t(\mathbb{Y}_{\leq t-1}),Y_t)-\ell(y^*,Y_t)\right)\leq 0,\quad (a.s),$$

then SOLAR-U = FS and there exists an optimistically universal learning rule for adversarial regression. Otherwise, SOLAR-U $=\emptyset$.

REMARK 6.5. There exists separable metric value spaces $(\mathcal{Y},\rho_{\mathcal{Y}})$ for which powers of metrics losses still yield SOLAR-U = FS. For instance, consider $(\mathcal{Y},\rho_{\mathcal{Y}})=(\mathbb{R},\sqrt{|\cdot|_2})$, where $|\cdot|_2$ denotes the Euclidean metric. One can check that this defines a metric on $\mathcal{Y}$ and for any loss $\ell=\rho_{\mathcal{Y}}^{\alpha}$ with $\alpha\leq 2$, we have SOLAR-U = FS. However, for $\alpha>2$, SOLAR-U $=\emptyset$.

7. Adversarial universal learning with moment constraint. In the previous section, we showed that learnable processes for adversarial regression are only in FS, i.e., visit a finite number of instance points. This shows that universal learning without restrictions on the adversarial responses $\mathbb{Y}$ is extremely restrictive. For instance, it does not contain i.i.d. processes. A natural question is whether adding mild constraints on the process $\mathbb{Y}$ would allow recovering the same results for unbounded losses as for bounded losses from Section 4 and 5. This question also arises in noiseless regression since the set of learnable processes is reduced from SOUL = SMV for bounded losses to SOUL = FS for unbounded losses. Hence, [4] posed as question whether having finite long-run empirical first-order moments would be sufficient to recover learnability in SMV. Precisely, they introduced the following constraint on noiseless processes $\mathbb{Y}=f^*(\mathbb{X})$: there exists $y_0\in\mathcal{Y}$ with

$$\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\ell(y_0,f^*(X_t))<\infty\quad (a.s.).$$

The question now becomes whether there exists an online learning rule which would be consistent under all $\mathbb{X}\in$ SMV processes for any noiseless responses $\mathbb{Y}=f^*(\mathbb{X})$ with f^* satisfying the above first-moment condition. We show that such an objective is not achievable whenever $\mathcal{X}$ is infinite—if $\mathcal{X}$ is finite, any process $\mathbb{X}$ on $\mathcal{X}$ is automatically FS and hence learnable in a noiseless or adversarial setting. In fact, under this first-order moment condition, we show the stronger statement that learning under all processes $\mathbb{X}$ which admit pointwise convergent relative frequencies (CRF) is impossible even in this noiseless setting.

Condition CRF: For any measurable set $A\in\mathcal{B}$, $\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{1}_A(X_t)$ exists almost surely.

[22] showed that CRF $\subset$ CS. In particular, CRF $\subset$ SMV. We show the following negative result on learning under CRF processes for noiseless regression under first-order moment constraint, which holds for unbounded near-metric spaces $(\mathcal{Y},\ell)$.

THEOREM 7.1. *Suppose that $\mathcal{X}$ is infinite and that $(\mathcal{Y},\ell)$ is an unbounded separable near-metric space. There does not exist an online learning rule which would be consistent under all processes $\mathbb{X}\in$ CRF for all measurable target functions $f^*:\mathcal{X}\to\mathcal{Y}$ such that there exists $y_0\in\mathcal{Y}$ with*

$$\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\ell(y_0,f^*(X_t))<\infty\quad (a.s.).$$

Proof sketch. We consider a sequence of values $(y_k)_{k\geq 0}$ such that $\ell(y_0,y_k)$ diverges as $k\to\infty$, then let $(t_k)_{k\geq 1}$ be a sequence of times such that $t_k \approx \sum_{k'\leq k}\ell(y_0,y_k)$. Next, let $(x_k)_{k\geq 0}$ be a sequence of distinct points. We construct a process $\mathbb{X}$ such that $X_t = x_0$ except at sparse times $(t_k)_{k\geq 1}$ for which $X_{t_k} = x_k$. Because t_k has a super-linear growth, $\mathbb{X}$ visits a sublinear number of distinct points and we can show that it satisfies the CRF property. Now for a random binary sequence $\boldsymbol{b} = (b_k)_{kk\geq 1}$ we consider the function $f^*_{\boldsymbol{b}}$ which is equal to y_0 except at points x_k for $k\geq 1$ where $f^*_{\boldsymbol{b}}(x_k) = y_0\mathbb{1}[b_k=0] + y_k\mathbb{1}[b_k=1]$. With these classes of functions, the algorithm cannot know in advance at time t_k whether to predict y_0 or y_k and incurs a loss $\mathcal{O}(\ell(y_0,y_k))$ in average as a result. Therefore, at time t_k, a total loss $\mathcal{O}(\sum_{k'\leq k}\ell(y_0,y_k)) = \mathcal{O}(t_k)$ is incurred compared to $f^*_{\boldsymbol{b}}$. On the other hand, by the construction of the sequence $(t_k)_{k\geq 1}$, $\frac{1}{T}\sum_{t=1}^T \ell(y_0, f^*_{\boldsymbol{b}}(X_t)) \leq \frac{1}{T}\sum_{t_k\leq T}\ell(y_0,y_k)$ stays bounded. Thus the learning rule is not consistent under all target functions satisfying the specified moment constraint.

Theorem 7.1 answers negatively to the question posed in [4]. A natural question is whether another meaningful constraint on responses can be applied to obtain positive results under large classes of processes on $\mathcal{X}$. To this means, we introduced the slightly stronger *empirical integrability* condition. We recall that an (adversarial) process $\mathbb{Y}$ is *empirically integrable* if and only if there exists $y_0\in\mathcal{Y}$ such that for any $\epsilon>0$, almost surely there exists $M\geq 0$ with

$$\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^T \ell(y_0,Y_t)\mathbb{1}_{\ell(y_0,Y_t)\geq M}\leq \epsilon.$$

Note that the threshold M may be *dependent* on the adversarial process $\mathbb{Y}$, but the guarantee should hold for any choice of predictions (in the case of adaptive adversaries). This is essentially the mildest condition on the sequence $\mathbb{Y}$ for which we can still obtain results. For example, if the loss is bounded, this constraint is automatically satisfied using $M>\bar{\ell}$. More importantly, note that any process $\mathbb{Y}$ which has bounded higher-than-first moments, i.e., such that there exists $p>1$ and $y_0\in\mathcal{Y}$ such that $\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^T\ell^p(y_0,Y_t)<\infty,\quad (a.s.)$, is empirically integrable. Further, for stationary processes $\mathbb{Y}$, having bounded first moment $\mathbb{E}[\ell(y_0,Y_1)]<\infty$ is exactly being empirically integrable. Indeed, by the strong law of large numbers, almost surely $\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^T\ell(y_0,Y_t)\mathbb{1}_{\ell(y_0,Y_t)\geq M} = \mathbb{E}[\ell(y_0,Y_1)\mathbb{1}_{\ell(y_0,Y_1)\geq M}]$. Therefore, empirical integrability is a direct consequence of the dominated convergence theorem.

LEMMA 7.2. *Let $\mathbb{Y}$ an stationary process on $\mathcal{Y}$ which has bounded first moment, i.e., there exists $y_0\in\mathcal{Y}$ such that $\mathbb{E}[\ell(y_0,Y_1)]<\infty$. Then, $\mathbb{Y}$ is empirically integrable.*

PROOF. Let $\mathbb{Y}$ an stationary process and $y_0\in\mathcal{Y}$ with $\mathbb{E}[\ell(y_0,Y_1)]<\infty$. Then, by the dominated convergence theorem we have $\mathbb{E}[\ell(y_0,Y_1)\mathbb{1}_{\ell(y_0,Y_1)\geq M}]\to 0$ as $M\to\infty$. Hence, for $\epsilon>0$, there exists M_ϵ such that $\mathbb{E}[\ell(y_0,Y_1)\mathbb{1}_{\ell(y_0,Y_1)\geq M}]\leq\epsilon$. Then, the sequence $(\ell(y_0,Y_t)\mathbb{1}_{\ell(y_0,Y_t)\geq M})_t$ is still stationary. hence, by the law of large numbers, almost surely,

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^T\ell(y_0,Y_t)\mathbb{1}_{\ell(y_0,Y_t)\geq M_\epsilon} = \mathbb{E}[\ell(y_0,Y_1)\mathbb{1}_{\ell(y_0,Y_1)\geq M_\epsilon}]\leq\epsilon.$$

This ends the proof that $\mathbb{Y}$ is empirically integrable. $\square$

The goal of this section is to show that under this moment constraint, we can recover all results from [2], [23] and this work in Sections 4 and 5, even for unbounded value spaces, leading up to Theorems 3.2 and 3.3. We will use the following simple equivalent formulation for empirical integrability.

LEMMA 7.3. *A process $\mathbb{Y}$ is empirically integrable if and only if there exists $y_0 \in \mathcal{Y}$ such that almost surely, for any $\epsilon > 0$ there exists $M > 0$ with*

$$\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \ell(y_0, Y_t)\mathbb{1}_{\ell(y_0,Y_t)\geq M} \leq \epsilon.$$

General strategy. First, the empirical integrability condition holds for some $y_0 \in \mathcal{Y}$ if and only if it holds for all $y_0 \in \mathcal{Y}$. Thus, we can fix $y_0 \in \mathcal{Y}$ independently of the instance or value process. Next, we define the restriction function $\phi_M : \mathcal{Y} \to \mathcal{Y}$ such that $\phi_M(y) = y$ if $\ell(y_0, y) < M$ and $\phi_M(y) = y_0$ otherwise. This function has values in the bounded set $B_\ell(y_0, M)$. Thus, we can apply our learning rules for the bounded loss case to learn the restricted values $\mathbb{Y}^M = (\phi_M(Y_t))_{t\geq 1}$. If we use these predictions to learn $\mathbb{Y}$, the excess loss compared to a fixed function mostly results from the restriction $\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \ell(Y_t, \phi_M(Y_t)) = \limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \ell(y_0, Y_t)\mathbb{1}_{\ell(y_0,Y_t)\geq M}$. This excess can then be bounded with the empirical integrability condition at y_0. We then combine the resulting predictors for $M \geq 1$ using Lemma 4.2. While this general strategy allows to use learning rules for the bounded loss case as subroutine to solve the unbounded loss case with empirical integrability constraint, we can adapt it to each case to simplify the algorithms.

7.1. *Noiseless universal learning with moment condition.* We first apply this strategy to the noiseless case. The main result from [2] showed that the 2C1NN learning rule achieves universal consistency on all SMV processes for bounded value spaces. Instead of using the 2C1NN learning rule as subroutine as described in the strategy above, we show that we can readily use 2C1NN for empirically integrable noiseless responses in unbounded value spaces, as stated in Theorem 3.1.

To prove this result, we first observe that 2C1NN trained on the responses $\mathbb{Y} = (f^*(X_t))_{t\geq 1}$ or the restricted responses $(\phi_M \circ f^*(X_t))_{t\geq 1}$ gives the same prediction at time t provided that the representative $\phi(t)$ satisfied $\ell(y_0, Y_{\phi(t)}) < M$. By construction of the 2C1NN learning rule, points can be used as representatives at most twice. Hence, up to a factor 2, times when the predictions on unrestricted and restricted responses differ, can be associated with times when $\ell(y_0, Y_t) \geq M$. As a result, we show that the empirical integrability condition can be applied to bound the excess loss resulting from the difference between unrestricted and restricted responses.

7.2. *Adversarial regression with moment condition under CS processes.* We now turn to adversarial regression under CS processes. [23] showed that regression for arbitrary responses under all CS processes is achievable in bounded value spaces. We generalize this result to unbounded losses and to adversarial responses with empirical integrability constraint using the general strategy. In particular, our learning rule is also optimistically universal for adversarial regression for all bounded value spaces which do not satisfy F-TiME. Now consider the general case and suppose that there exists a ball $B_\ell(y, r)$ which does not satisfy F-TiME, Theorem 5.8 shows that universal learning for values falling in $B_\ell(y, r)$ cannot be achieved for processes $\mathbb{X} \notin$ CS. Now because $B_\ell(y, r)$ is bounded, responses restricted to this set satisfy the empirical integrability constraint. In particular, this shows that the condition CS is also necessary for universal learning with adversarial responses with empirical integrability. Altogether, this proves Theorem 3.3.

This generalizes the main results from [23] to unbounded non-metric losses and from [9] to non-metric losses, arbitrary responses and CS instance processes $\mathbb{X}$. Indeed, they consider bounded first moment conditions on i.i.d. responses, which are empirically integrable by Lemma 7.2. Further, as a direct consequence of Theorem 3.3 and Lemma 7.2, we can significantly relax the conditions for universal consistency on stationary ergodic processes found in the literature. Precisely, [19] showed that for regression with squared loss, under the assumption $\mathbb{E}[Y_1^4] < \infty$, consistency on stationary ergodic processes is possible. We can relax this result to bounded second moments, matching the standard results for i.i.d. processes.

COROLLARY 7.4. *Let $(\mathcal{Y}, \ell) = (\mathbb{R}, |\cdot|^2)$. The learning rule of Theorem 3.3 is consistent on any stationary ergodic process $(X_t, Y_t)_{t\geq 1}$ with $\mathbb{E}[Y_1^2] < \infty$.*

7.3. *Adversarial regression with moment condition under SMV processes.* Last, we generalize our result Theorem 5.5 for value spaces satisfying F-TiME, to unbounded value spaces, with the same moment condition on responses using the general strategy. In order to apply Theorem 5.5 to bounded balls of the value space, we now ask that all balls $B_\ell(y, r)$ in the value space $(\mathcal{Y}, \ell)$ satisfy F-TiME. This proves Theorem 3.2.

Theorems 3.3 and 3.2 completely characterize learnability for adversarial regression with moment condition. Namely, if the value space $(\mathcal{Y}, \ell)$ is such that any bounded ball satisfies F-TiME (resp. there exists a ball $B_\ell(y, r)$ that disproves F-TiME), Theorem 3.2 (resp. 3.3) gives an optimistic learning rule which achieves consistency under all processes in SMV (resp. CS). This ends our analysis of adversarial regression for unbounded value spaces.

8. Open research directions. In this work, we provided a characterization of learnability for universal learning in the regression setting, for a class of losses satisfying specific relaxed triangle inequality identities, which contains powers of metrics $\ell = \rho_{\mathcal{Y}}^\alpha$ for $\alpha \geq 1$. A natural question would be whether one can generalize these results to larger classes of losses, e.g. non-symmetric losses which may appear in classical machine learning problems.

The present work could also have some implications for adversarial contextual bandits. Specifically, one may consider the case of a learner who receives partial information on the rewards/losses as opposed to the traditional regression setting where the response is completely revealed at each iteration. In the latter case, the learner can for instance compute the loss of *all* values with respect to the response realization. On the other hand, in the contextual bandits framework, the reward/loss is revealed *only* for the pulled arm—or equivalently the prediction of the learner. In these partial information settings, exploration then becomes necessary. The authors are investigating whether the results presented in this work could have consequences in these related domains.

Funding. This work is being partly funded by ONR grant N00014-18-1-2122.

SUPPLEMENTARY MATERIAL

Supplement to "Universal regression with adversarial responses"
The supplementary material contains all the proofs and auxiliary lemmas.

REFERENCES

[1] BEN-DAVID, S., PÁL, D. and SHALEV-SHWARTZ, S. (2009). Agnostic online learning. In *Conference on Learning Theory* **3** 1.

[2] BLANCHARD, M. (2022). Universal online learning: an optimistically universal learning rule. In *Conference on Learning Theory* 1077-1125. PMLR.

[3] BLANCHARD, M. and COSSON, R. (2022). Universal online learning with bounded loss: reduction to binary classification. In *Conference on Learning Theory* 479–495. PMLR.

[4] BLANCHARD, M., COSSON, R. and HANNEKE, S. (2022). Universal online learning with unbounded losses: memory is all you need. In *International Conference on Algorithmic Learning Theory* 107–127. PMLR.

[5] BUBECK, S. and CESA-BIANCHI, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Machine Learning* **5** 1–122.

[6] CESA-BIANCHI, N., FREUND, Y., HAUSSLER, D., HELMBOLD, D. P., SCHAPIRE, R. E. and WARMUTH, M. K. (1997). How to use expert advice. *Journal of the ACM (JACM)* **44** 427–485.
[7] CESA-BIANCHI, N. and LUGOSI, G. (2006). *Prediction, learning, and games*. Cambridge university press.
[8] CHANG, T. (1989). Spherical regression with errors in variables. *The Annals of Statistics* 293–306.
[9] COHEN, D. T. and KONTOROVICH, A. (2022). Learning with metric losses. In *Conference on Learning Theory* 662–700. PMLR.
[10] DAVIS, B. C., FLETCHER, P. T., BULLITT, E. and JOSHI, S. (2010). Population shape regression from random design data. *International Journal of Computer Vision* **90** 255–266.
[11] DEVROYE, L., GYORFI, L., KRZYZAK, A. and LUGOSI, G. (1994). On the strong universal consistency of nearest neighbor regression function estimates. *The Annals of Statistics* 1371–1385.
[12] DEVROYE, L., GYÖRFI, L. and LUGOSI, G. (2013). *A probabilistic theory of pattern recognition* **31**. Springer Science & Business Media.
[13] EVANS, S. N. and JAFFE, A. Q. (2020). Strong laws of large numbers for Fréchet means. *arXiv preprint arXiv:2012.12859*.
[14] FLETCHER, P. T. (2013). Geodesic regression and the theory of least squares on Riemannian manifolds. *International Journal of Computer Vision* **105** 171–185.
[15] FREUND, Y. and SCHAPIRE, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* **55** 119–139.
[16] GRAY, R. M. and GRAY, R. (2009). *Probability, random processes, and ergodic properties* **1**. Springer.
[17] GYÖRFI, L., KÖHLER, M., KRZYŻAK, A. and WALK, H. (2002). *A distribution-free theory of nonparametric regression* **1**. Springer.
[18] GYORFI, L., LUGOSI, G. and MORVAI, G. (1999). A simple randomized algorithm for sequential prediction of ergodic time series. *IEEE Transactions on Information Theory* **45** 2642–2650.
[19] GYÖRFI, L. and OTTUCSÁK, G. (2007). Sequential prediction of unbounded stationary time series. *IEEE Transactions on Information Theory* **53** 1866–1872.
[20] GYÖRFI, L. and WEISS, R. (2021). Universal consistency and rates of convergence of multiclass prototype algorithms in metric spaces. *Journal of Machine Learning Research* **22** 1–25.
[21] HANNEKE, S. (2021). Open Problem: is there an online learning algorithm that learns whenever online learning is possible? In *Conference on Learning Theory* 4642–4646. PMLR.
[22] HANNEKE, S. (2021). Learning whenever learning is possible: Universal learning under general stochastic processes. *Journal of Machine Learning Research* **22** 1–116.
[23] HANNEKE, S. (2022). Universally consistent online learning with arbitrarily dependent responses. In *International Conference on Algorithmic Learning Theory* 488–497. PMLR.
[24] HANNEKE, S., KONTOROVICH, A., SABATO, S. and WEISS, R. (2021). Universal Bayes consistency in metric spaces. *The Annals of Statistics* **49** 2129 – 2150.
[25] JAFFE, A. Q. (2022). Strong consistency for a class of adaptive clustering procedures. *arXiv preprint arXiv:2202.13423*.
[26] LIN, Z. and MÜLLER, H.-G. (2021). Total variation regularized Fréchet regression for metric-space valued data. *The Annals of Statistics* **49** 3510–3533.
[27] LITTLESTONE, N. (1988). Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning* **2** 285–318.
[28] LITTLESTONE, N. and WARMUTH, M. K. (1994). The weighted majority algorithm. *Information and Computation* **108** 212–261.
[29] MARDIA, K. V., JUPP, P. E. and MARDIA, K. (2000). *Directional statistics* **2**. Wiley Online Library.
[30] MORVAI, G., KULKARNI, S. R. and NOBEL, A. B. (1999). Regression estimation from an individual stable sequence. *Statistics: A Journal of Theoretical and Applied Statistics* **33** 99–118.
[31] MORVAI, G., YAKOWITZ, S. and GYÖRFI, L. (1996). Nonparametric inference for ergodic, stationary time series. *The Annals of Statistics* **24** 370–379.
[32] RAKHLIN, A., SRIDHARAN, K. and TEWARI, A. (2015). Online learning via sequential complexities. *Journal of Machine Learning Research* **16** 155–186.
[33] SCHÖTZ, C. (2022). Strong laws of large numbers for generalizations of Fréchet mean sets. *Statistics* 1–19.
[34] SHI, X., STYNER, M., LIEBERMAN, J., IBRAHIM, J. G., LIN, W. and ZHU, H. (2009). Intrinsic regression models for manifold-valued data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 192–199. Springer.
[35] SLIVKINS, A. (2019). Introduction to multi-armed bandits. *arXiv preprint arXiv:1904.07272*.
[36] STEINWART, I., HUSH, D. and SCOVEL, C. (2009). Learning from dependent observations. *Journal of Multivariate Analysis* **100** 175–194.
[37] STONE, C. J. (1977). Consistent nonparametric regression. *The Annals of Statistics* 595–620.
[38] ZAICHYK, H., BIESS, A., KONTOROVICH, A. and MAKARYCHEV, Y. (2019). Efficient Kirszbraun extension with applications to regression. *arXiv preprint arXiv:1905.11930*.