Accepted, Annals of Statistics, June 2023

SUPPLEMENT TO “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*

CONTENTS

APPENDIX A: IDENTITIES ON THE LOSS FUNCTION

We recall the following known identities, which we will use to analyze the loss $\ell = \rho_{\mathcal{Y}}^{\alpha}$.

LEMMA A.1. *Let $\alpha \geq 1$. Then, $(a+b)^\alpha \leq 2^{\alpha-1}(a^\alpha + b^\alpha)$ for all $a, b \geq 0$. Let $0 < \epsilon \leq 1$ and $\alpha \geq 1$. There exists some constant $c_\epsilon^\alpha > 0$ such that $(a+b)^\alpha \leq (1+\epsilon)a^\alpha + c_\epsilon^\alpha b^\alpha$ for all $a, b \geq 0$, and $c_\epsilon^\alpha \leq \left(\frac{4\alpha}{\epsilon}\right)^\alpha$.*

PROOF. The first identity is classical. A proof of the second one can be found for example in [4] (Lemma 2.3) where they obtain $c_\epsilon^\alpha = \left(1 + \frac{1}{(1+\epsilon)^{1/\alpha}-1}\right)^\alpha \leq \left(\frac{4\alpha}{\epsilon}\right)^\alpha$. □

APPENDIX B: PROOFS OF SECTION 4

B.1. Proof of Theorem 4.1. In this section, we prove that for any $\delta > 0$, the $(1+\delta)$C1NN learning rule is optimistically universal for the noiseless setting. The proof follows the same structure as the proof of the main result in [1] which shows that 2C1NN is optimistically universal. We first focus on the binary classification setting and show that the learning rule $(1+\delta)$C1NN is consistent on functions representing open balls.

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

PROOF. We fix $\bar{x} \in \mathcal{X}$, $r > 0$ and $f^* = \mathbb{1}_{B(\bar{x},r)}$. We reason by the contrapositive and suppose that $(1+\delta)$C1NN is not consistent on f^*. Then, $\eta := \mathbb{P}(\mathcal{L}_{\mathbb{X}}((1+\delta)C1NN, f^*) > 0) > 0$. Therefore, there exists $0 < \epsilon \leq 1$ such that $\mathbb{P}(\mathcal{L}_{\mathbb{X}}((1+\delta)C1NN, f^*) > \epsilon) > \frac{\eta}{2}$. Denote by $\mathcal{A} := \{\mathcal{L}_{\mathbb{X}}((1+\delta)C1NN, f^*) > \epsilon\}$. this event of probability at least $\frac{\eta}{2}$. Because $\mathcal{X}$ is separable, let $(x^i)_{i\geq 1}$ a dense sequence of $\mathcal{X}$. We consider the same partition $(P_i)_{i\geq 1}$ of $B(\bar{x},r)$ and the partition $(A_i)_{i\geq 0}$ of $\mathcal{X}$ as in the original proof of [1], but with the constant $c_\epsilon := \frac{1}{2\cdot 2^{2^8/(\epsilon\delta)}}$ and changing the construction of the sequence $(n_l)_{l\geq 1}$ so that for all $l \geq 1$

$$\mathbb{P}\left[\forall n \geq n_l,\ |\{i,\ P_i(\tau_l) \cap \mathbb{X}_{<n} \neq \emptyset\}| \leq \frac{\epsilon\delta}{2^{10}} n\right] \geq 1 - \frac{\delta}{2\cdot 2^{l+2}} \quad \text{and} \quad n_{l+1} \geq \frac{2^9}{\epsilon\delta} n_l.$$

Last, consider the product partition of $(P_i)_{i\geq 1}$ and $(A_i)_{i\geq 0}$ which we denote $\mathcal{Q}$. Similarly, we define the same events $\mathcal{E}_l, \mathcal{F}_l$ for $l \geq 1$. We aim to show that with nonzero probability, $\mathbb{X}$ does not visit a sublinear number of sets of $\mathcal{Q}$.

We now denote by $(t_k)_{k\geq 1}$ the increasing sequence of all (random) times when $(1+\delta)$C1NN makes an error in the prediction of $f^*(X_t)$. Because the event $\mathcal{A}$ is satisfied, $\mathcal{L}_{\boldsymbol{x}}((1+\delta)C1NN, f^*) > \epsilon$, we can construct an increasing sequence of indices $(k_l)_{l\geq 1}$ such that $t_{k_l} < \frac{2k_l}{\epsilon}$. For any $t \geq 2$, we will denote by $\phi(t)$ the (random) index of the representative chosen by the $(1+\delta)$C1NN learning rule. Now let $l \geq 1$. Consider the tree $\mathcal{G}$ where nodes are times $\mathcal{T} := \{t \leq t_{k_l}\}$ within horizon t_{k_l}, where the parent relations are given by $(t, \phi(t))$ for $t \in \mathcal{T} \setminus \{1\}$. In other words, we construct the tree in which the parent of each new input is its representative. Note that by construction of the $(1+\delta)$C1NN learning rule, each node has at most 2 children.

B.1.1. *Step 1.* In this step, we consider the case when the majority of input points on which $(1+\delta)$C1NN made a mistake belong to $B(\bar{x},r)$, i.e., $|\{k \leq k_l,\ X_{t_k} \in B(\bar{x},r)\}| \geq \frac{k_l}{2}$. We denote $\mathcal{H}_1$ this event. Let us now consider the subgraph $\tilde{\mathcal{G}}$ given by restricting $\mathcal{G}$ only to nodes in the ball $B(\bar{x},r)$—which are mapped to the true value 1—i.e., on times $\mathcal{T} := \{t \leq t_{k_l},\ X_t \in B(\bar{x},r)\}$. In this subgraph, the only times with no parent are times t_k with $k \leq k_l$ and $X_{t_k} \in B(\bar{x},r)$, and possibly time $t = 1$. Therefore, $\tilde{\mathcal{G}}$ is a collection of disjoint trees with roots times $\{t_k,\ k \leq k_l,\ x_{t_k} \in B(\bar{x},r)\}$, and possibly $t = 1$ if $X_1 \in B(\bar{x},r)$. For a given time t_k with $k \leq k_l$ and $X_{t_k} \in B(\bar{x},r)$, we denote by $\mathcal{T}_k$ the corresponding tree in $\tilde{\mathcal{G}}$ with root t_k. We now introduce the notion of *good* trees. We say that $\mathcal{T}_k$ is a good tree if $\mathcal{T}_k \cap \mathcal{D}_{t_{k_l}+1} \neq \emptyset$, i.e., the tree survived until the last dataset. Conversely a tree is *bad* if all its nodes were deleted before time $t_{k_l}+1$. We denote the set of good and bad trees by $G = \{k : \mathcal{T}_k \text{ good}\}$ and $B = \{k : \mathcal{T}_k \text{ bad}\}$. In particular, we have $|G| + |B| = |\{k \leq k_l, X_{t_k} \in B(\bar{x},r)\}| \geq k_l/2$. We aim to upper bound the number of bad trees. We now focus on trees $\mathcal{T}_k$ which induced a future first mistake, i.e., such that $\{l \in \mathcal{T}_k | \exists u \leq t_{k_l} : \phi(u) = l, \rho_{\mathcal{X}}(X_l, \bar{x}) \geq r \text{ and } \forall v <$

$u, \phi(v) \neq l\} \neq \emptyset$. We denote the corresponding minimum time $l_k = \min\{l \in \mathcal{T}_k \mid \exists u \leq t_{k_l} : \phi(u) = l, \rho_{\mathcal{X}}(X_l, \bar{x}) \geq r, \forall v < u, \phi(v) \neq l\}$. The terminology first mistake refers to the fact that the first time which used l as representative corresponded to a mistake, as opposed to l already having a children $X_u \in B(\bar{x}, r)$ which continues descendents of l within the tree $\mathcal{T}_k$. Note that bad trees necessarily induce a future first mistake—otherwise, this tree would survive. For each of these times l_k two scenarios are possible.

1. The value U_{l_k} was never revealed within horizon t_{k_l}: as a result $l_k \in \mathcal{D}_{t_{k_l}+1}$.
2. The value U_{l_k} was revealed within horizon t_{k_l}. Then, U_{l_k} we revealed using a time t for which l_k was a potential representative. This scenario has two cases:
   a) $\rho_{\mathcal{X}}(X_t, \bar{x}) < r$. If used as representative $\phi(t) = l_k$, then l_k would not have induced a mistake in the prediction of Y_t.
   b) $\rho_{\mathcal{X}}(X_t, \bar{x}) \geq r$. If used as representative $\phi(t) = l_k$, then l_k would have induced a mistake in the prediction of Y_t.

In the case 2.a), if the point is used as representative $\phi(t) = l_k$ and if the corresponding tree $\mathcal{T}_k$ was bad, at least another future mistake is induced by $\mathcal{T}_k$—otherwise this tree would survive. We consider times l_k for which the value was revealed, which corresponds to the only possible scenario for bad trees. We denote the corresponding set $K := \{k : U_{l_k} \text{ revealed within horizon } t_{k_l}\}$. We now consider the sequence $k_1^a, \ldots k_\alpha^a$ containing all indices of K for which scenario 2.a) was followed, ordered by chronological order for the reveal of $U_{l_{k_i^a}}$, i.e., $U_{l_{k_1^a}}$ was the first item of scenario 2.a) to be revealed, then $U_{l_{k_2^a}}$ etc. until $U_{l_{k_\alpha^a}}$. Similarly, we construct the sequence $k_1^b, \ldots k_\beta^b$ of indices in K corresponding to scenario 2.b), ordered by order for the reveal of $U_{l_{k_i^b}}$. We now consider the events

$$\mathcal{B} := \left\{\alpha + \beta \leq \frac{k_l}{2} - \frac{k_l \delta}{32}\right\}, \quad \mathcal{C} := \left\{\sum_{i=1}^{\min(\alpha, \lceil k_l/8 \rceil)} U_{l_{k_i^a}} \geq \frac{k_l \delta}{16}\right\},$$

$$\mathcal{D} := \left\{\sum_{i=1}^{\min(\beta, \lceil k_l/8 \rceil)} U_{l_{k_i^b}} \geq \frac{k_l \delta}{16}\right\}.$$

We now show that for $l > 16$, under the event

$$\mathcal{M}_{k_l} := \mathcal{H}_1 \cap \left[\mathcal{B} \cup (\{\alpha \geq \lceil k_l/8 \rceil\} \cap \mathcal{C}) \cup (\{\alpha < \lceil k_l/8 \rceil\} \cap \mathcal{D})\right],$$

we have that $|G| \geq \frac{k_l \delta}{32}$. Suppose that $\mathcal{M}_{k_l}$ is met. First note that because a bad tree can only fall into scenarios 2.a) or 2.b) we have $|B| \leq \alpha + \beta$. Hence $|G| \geq \frac{k_l}{2} - \alpha - \beta$ because of $\mathcal{H}_1$. Thus, the result holds directly if $\mathcal{B}$ is satisfied. We can now suppose that $\mathcal{B}^c$ is satisfied, i.e., $\alpha + \beta > \frac{k_l}{2} - \frac{k_l \delta}{32}$. Now suppose that $\alpha \geq \lceil k_l/8 \rceil$ and $\mathcal{C}$ are also satisfied. For all indices such that $U_{l_{k_i^a}} = 1$, i.e., we fall in case 2.a) and $l_{k_i^a}$ is used as representative, the corresponding tree $\mathcal{T}_{k_i^a}$ would need to induce at least an additional mistake to be bad. Recall that in total at most $k_l/2$ mistakes are induced by points of $\mathcal{T}$. Also, by definition of the set K, $\alpha + \beta$ mistakes are already induced by the times t_k for $k \in K$. These corresponded to the future first mistakes for all times $\{l_k : k \in K\}$. Hence, we obtain

$$|G| \geq \sum_{i=1}^{\alpha} U_{l_{k_i^a}} - \left(\frac{k_l}{2} - \alpha - \beta\right) \geq \frac{k_l \delta}{16} - \frac{k_l \delta}{32} = \frac{k_l \delta}{32}.$$

Now consider the case where $\mathcal{H}_1$, $\mathcal{B}^c$, $\alpha < \lceil k_l/8 \rceil$ and $\mathcal{D}$ are met. In particular, because $l > 16$ we have $k_l > 16$ hence $\frac{k_l}{2} - \frac{k_l \delta}{32} \geq 2\lceil k_l/8 \rceil$. Thus, because of $\mathcal{B}^c$ we have $\beta > \frac{k_l}{2} - \frac{k_l \delta}{32} - \alpha \geq$

$\lceil k_l/8 \rceil$. Now observe that for all indices such that $U_{l_{k_i^b}} = 1$, the time l_k induced two mistakes. Therefore, counting the total number of mistakes we obtain

$$\frac{k_l}{2} \geq \alpha + \beta + \sum_{i=1}^{\beta} U_{l_{k_i^b}} \geq \frac{k_l}{2} - \frac{k_l \delta}{32} + \frac{k_l \delta}{16}$$

which is impossible. This ends the proof that under $\mathcal{M}_{k_l}$ we have $|G| \geq \frac{k_l \delta}{32}$.

We now aim to lower bound the probability of this event. To do so, we first upper bound the probability of the event $\{\alpha \geq \lceil k_l/8 \rceil\} \cap \mathcal{C}^c$. We introduce a process $(Z_i)_{i=1}^{\lceil k_l/8 \rceil}$ such that for all $i \leq \max(\alpha, \lceil k_l/8 \rceil)$, $Z_i = U_{l_{k_i^a}} - \delta$ and $Z_i = 0$ for $\alpha < i \leq \lceil k_l/8 \rceil$. Because of the specific ordering chosen $k_1^a, \ldots, k_\alpha^a$, this process is a sequence of martingale differences, with values bounded by 1 in absolute value. Therefore, for $l > 16$ the Azuma-Hoeffing inequality yields

$$\mathbb{P}\left[\sum_{i=1}^{\lceil k_l/8 \rceil} Z_i \leq -\frac{k_l \delta}{16}\right] \leq e^{-\frac{k_l^2 \delta^2}{2 \cdot 16^2 (k_l/8+1)}} \leq e^{-\frac{k_l \delta^2}{2^7}}.$$

But on the event $\{\alpha \geq \lceil k_l/8 \rceil\} \cap \mathcal{C}^c$ we have precisely

$$\sum_{i=1}^{\lceil k_l/8 \rceil} Z_i = \sum_{i=1}^{\min(\alpha, \lceil k_l/8 \rceil)} U_{l_{k_i^a}} - \lceil k_l/8 \rceil \delta \leq \frac{k_l \delta}{16} - \lceil k_l/8 \rceil \delta \leq -\frac{k_l \delta}{16}.$$

Therefore $\mathbb{P}[\mathcal{C}^c \cap \{\alpha \geq \lceil k_l/8 \rceil\}] \leq \mathbb{P}\left[\sum_{i=1}^{\lceil k_l/8 \rceil} Z_i \leq -\frac{k_l \delta}{16}\right] \leq e^{-k_l \delta^2/2^7}$. Similarly we obtain $\mathbb{P}[D^c \cap \{\beta \geq \lceil k_l/8 \rceil\}] \leq e^{-k_l \delta^2/2^7}$. Finally we write for any $l > 16$,

$$\begin{aligned}
\mathbb{P}[\mathcal{H}_1 \setminus \mathcal{M}_{k_l}] &= \mathbb{P}[\mathcal{H}_1 \cap \mathcal{B}^c \cap (\{\alpha < \lceil k_l/8 \rceil\} \cup \mathcal{C}^c) \cap (\{\alpha \geq \lceil k_l/8 \rceil\} \cup \mathcal{D}^c)] \\
&= \mathbb{P}[\mathcal{H}_1 \cap \mathcal{B}^c \cap [(\{\alpha < \lceil k_l/8 \rceil\} \cap \mathcal{D}^c) \cup (\{\alpha \geq \lceil k_l/8 \rceil\} \cap \mathcal{C}^c)]] \\
&\leq \mathbb{P}[\mathcal{C}^c \cap \{\alpha \geq \lceil k_l/8 \rceil\}] + \mathbb{P}[\mathcal{D}^c \cap \{\alpha < \lceil k_l/8 \rceil\} \cap \mathcal{B}^c] \\
&\leq \mathbb{P}[\mathcal{C}^c \cap \{\alpha \geq \lceil k_l/8 \rceil\}] + \mathbb{P}[\mathcal{D}^c \cap \{\beta \geq \lceil k_l/8 \rceil\}] \\
&\leq 2e^{-\frac{k_l \delta^2}{2^7}}.
\end{aligned}$$

In particular, we obtain

$$\mathbb{P}\left[\left\{|G| \geq \frac{k_l \delta}{32}\right\} \cap \mathcal{H}_1\right] \geq \mathbb{P}[\mathcal{M}_{k_l}] \geq \mathbb{P}[\mathcal{H}_1] - 2e^{-\frac{k_l \delta^2}{2^7}}.$$

B.1.2. *Step 2.* We now consider the opposite case, when a majority of mistakes are made outside $B(\bar{x}, r)$, i.e., $|\{k \leq k_l,\ X_{t_k} \in B(\bar{x}, r)\}| < \frac{k_l}{2}$, which corresponds to the event $\mathcal{H}_1^c$. Similarly, we consider the subgraph $\tilde{\mathcal{G}}$ given by restricting $\mathcal{G}$ only to nodes outside the ball $B(\bar{x}, r)$, i.e., on times $\mathcal{T} := \{t \leq t_{k_l},\ \rho_{\mathcal{X}}(X_t, \bar{x}) \geq r)\}$. Again, $\tilde{\mathcal{G}}$ is a collection of disjoint trees with roots times $\{t_k,\ k \leq k_l,\ \rho_{\mathcal{X}}(X_{t_k}, \bar{x}) \geq r)\}$—and possibly $t = 1$. For a given time t_k with $k \leq k_l$ and $\rho_{\mathcal{X}}(X_{t_k}, \bar{x}) \geq r$, we denote by $\mathcal{T}_k$ the corresponding tree in $\tilde{\mathcal{G}}$ with root t_k. Similarly to the previous case, $\mathcal{T}_k$ is a *good* tree if $\mathcal{T}_k \cap \mathcal{D}_{t_{k_l}+1} \neq \emptyset$ and *bad* otherwise. We denote the set of good and bad trees by $G = \{k : \mathcal{T}_k \text{ good}\}$. We can again focus on trees $\mathcal{T}_k$ which induced a future first mistake, i.e., such that $\{l \in \mathcal{T}_k | \exists u \leq t_{k_l} : \phi(u) = l, \rho_{\mathcal{X}}(X_l, \bar{x}) < r \text{ and } \forall v < u, \phi(v) \neq l\} \neq \emptyset$ and more specifically their minimum time $l_k = \min\{l \in \mathcal{T}_k \mid \exists u \leq t_{k_l} : \phi(u) = l, \rho_{\mathcal{X}}(X_l, \bar{x}) < r, \forall v < u, \phi(v) \neq l\}$. The same analysis as above shows that

$$\mathbb{P}\left[\left\{|G| \geq \frac{k_l \delta}{32}\right\} \cap \mathcal{H}_1^c\right] \geq \mathbb{P}[\mathcal{H}_1^c] - 2e^{-\frac{k_l \delta^2}{2^7}}.$$

Therefore, if G denotes more generally the set of good trees (where we follow the corresponding case 1 or 2) we finally obtain that for any $l > 16$,

$$\mathbb{P}\left[|G| \geq \frac{k_l\delta}{32}\right] \geq 1 - 4e^{-\frac{k_l\delta^2}{2^7}}.$$

We denote by $\tilde{\mathcal{M}}_{k_l}$ this event. By Borel-Cantelli lemma, almost surely, there exists $\hat{l}$ such that for any $l \geq \hat{l}$, the event $\tilde{\mathcal{M}}_{k_l}$ is satisfied. We denote $\mathcal{M} := \bigcup_{l\geq 1}\bigcap_{l'\geq l}\tilde{\mathcal{M}}_{k_l}$ this event of probability one. The aim is to show that on the event $\mathcal{A}\cap\mathcal{M}\cap\bigcap_{l\geq 1}(\mathcal{E}_l\cap\mathcal{F}_l)$, which has probability at least $\frac{\eta}{4}$, $\mathbb{X}$ disproves the SMV condition. In the following, we consider a specific realization $\boldsymbol{x}$ of the process $\mathbb{X}$ falling in the event $\mathcal{A}\cap\mathcal{M}\cap\bigcap_{l\geq 1}(\mathcal{E}_l\cap\mathcal{F}_l)$—$\boldsymbol{x}$ is not random anymore. Let $\hat{l}$ be the index given by the event $\mathcal{M}$ such that for any $l\geq\hat{l}$, $\mathcal{M}_{k_l}$ holds. We consider $l \geq \hat{l}$ and successively consider different cases in which the realization $\boldsymbol{x}$ may fall.

- In the first case, we suppose that a majority of mistakes were made in $B(\bar{x}, r)$, i.e., that we fell into event $\mathcal{H}_1$ similarly to Step 1. Because the event $\tilde{\mathcal{M}}_{k_l}$ is satisfied we have $|G| \geq \frac{k_l\delta}{2^5}$. Now note that trees are disjoint, therefore, $\sum_{k\in G}|\mathcal{T}_k| \leq t_{k_l} < \frac{2k_l}{\epsilon}$. Therefore,

$$\sum_{k\in G}\mathbb{1}_{|\mathcal{T}_k|\leq\frac{2^7}{\epsilon\delta}} = |G| - \sum_{k\in G}\mathbb{1}_{|\mathcal{T}_k|>\frac{2^7}{\epsilon\delta}} > |G| - \frac{\epsilon\delta}{2^7}\sum_{k\in G}|\mathcal{T}_k| \geq \frac{k_l\delta}{2^5} - \frac{k_l\delta}{2^6} = \frac{k_l\delta}{2^6}.$$

  We will say that a tree $|\mathcal{T}_k|$ is *sparse* if it is good and has at most $\frac{2^7}{\epsilon\delta}$ nodes. With $S := \{k\in G,\ |\mathcal{T}_k| \leq \frac{2^7}{\epsilon\delta}\}$ the set of sparse trees, the above equation yields $|S| \geq \frac{k_l\delta}{2^6}$. The same arguments as in [1] give

$$|\{i,\ A_i\cap\boldsymbol{x}_{\leq t_{k_l}} \neq\emptyset\}| \geq |S| \geq \frac{k_l\delta}{2^6} \geq \frac{\epsilon\delta}{2^7}t_{k_l}.$$

  The only difference is that we chose c_ϵ so that $2^{2\cdot\frac{2^7}{\epsilon\delta}-1} \leq \frac{1}{4c_\epsilon}$ as needed in the original proof.
- We now turn to the case when the majority of input points on which $(1+\delta)$C1NN made a mistake are not in the ball $B(\bar{x}, r)$, similarly to Step 2. Using the same notion of sparse tree $S := \{k\in G,\ |\mathcal{T}_k|\leq\frac{2^7}{\epsilon\delta}\}$, we have again $|S|\geq\frac{k_l\delta}{2^6}$. We use the same arguments as in the original proof. Suppose $|\{k\in S,\ \rho_{\mathcal{X}}(x_{p^k_{d(k)}}, \bar{x}) > r\}| \geq \frac{|S|}{2}$, then we have

$$|\{i,\ A_i\cap\boldsymbol{x}_{\leq t_{k_l}}\neq\emptyset\}| \geq |\{k\in S,\ \rho_{\mathcal{X}}(x_{p^k_{d(k)}},\bar{x}) > r\}| \geq \frac{|S|}{2} \geq \frac{k_l\delta}{2^7} \geq \frac{\epsilon\delta}{2^8}t_{k_l}.$$

B.1.3. *Step 3.* In this last step, we suppose again that the majority of input points on which $(1+\delta)$C1NN made a mistake are not in the ball $B(\bar{x}, r)$ but that $|\{k\in S,\ \rho_{\mathcal{X}}(x_{p^k_{d(k)}},\bar{x}) > r\}| < \frac{|S|}{2}$. Therefore, we obtain

$$|\{k\in S,\ \rho_{\mathcal{X}}(x_{p^k_{d(k)}},\bar{x}) = r\}| = |S| - |\{k\in S,\ \rho_{\mathcal{X}}(x_{p^k_{d(k)}},\bar{x}) > r\}| \geq \frac{|S|}{2} \geq \frac{k_l\delta}{2^7} \geq \frac{\epsilon\delta}{2^8}t_{k_l}.$$

We will now make use of the partition $(P_i)_{i\geq 1}$. Because $(n_u)_{u\geq 1}$ is an increasing sequence, let $u\geq 1$ such that $n_{u+1}\leq t_{k_l}\leq n_{u+2}$ (we can suppose without loss of generality that $t_{k_0} > n_2$). Note that we have $n_u \leq \frac{\epsilon\delta}{2^9}n_{u+1} \leq \frac{\epsilon\delta}{2^9}t_{k_l}$. Let us now analyze the process between times n_u and t_{k_l}. In particular, we are interested in the indices $T = \{k\in S,\ \rho_{\mathcal{X}}(x_{p^k_{d(k)}},\bar{x}) = r\}$ and times $\mathcal{U}_u = \{p^k_{d(k)} :\ n_u < p^k_{d(k)} \leq k_l,\ k\in T\}$. In particular, we have

$$|\mathcal{U}_u| \geq |\{k\in S,\ \rho_{\mathcal{X}}(x_{p^k_{d(k)}},\bar{x}) = r\}| - n_u \geq \frac{\epsilon\delta}{2^8}t_{k_l} - \frac{\epsilon\delta}{2^9}t_{k_l} = \frac{\epsilon\delta}{2^9}t_{k_l}.$$

Defining $T' := \{k \in T,\ r - \frac{r}{2^{u+3}} \leq \rho_{\mathcal{X}}(x_{\phi(t_k)}, \bar{x}) < r\}$, the same arguments as in the original proof yield

$$|\{i,\ P_i \cap \boldsymbol{x}_{\leq t_{k_l}} \neq \emptyset\}| \geq |T'| \geq |\mathcal{U}_u| - |\{i,\ P_i(\tau_u) \cap \boldsymbol{x}_{\mathcal{U}_u} \neq \emptyset\}| \geq \frac{\epsilon\delta}{2^9} t_{k_l} - \frac{\epsilon\delta}{2^{10}} t_{k_l} = \frac{\epsilon\delta}{2^{10}} t_{k_l}.$$

B.1.4. *Step 4.* In conclusion, in all cases, we obtain

$$|\{Q \in \mathcal{Q},\ Q \cap \boldsymbol{x}_{\leq t_{k_l}} \neq \emptyset\}| \geq \max(|\{i,\ A_i \cap \boldsymbol{x}_{\leq t_{k_l}} \neq \emptyset\}|, |\{i,\ P_i \cap \boldsymbol{x}_{\leq t_{k_l}} \neq \emptyset\}|) \geq \frac{\epsilon\delta}{2^{10}} t_{k_l}.$$

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

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

This shows that $\mathbb{X} \notin$ SMV which is absurd. Therefore $(1+\delta)$C1NN is consistent on f^*. This ends the proof of the proposition. □

Using the fact that in the $(1+\delta)$C1NN learning rule, no time t can have more than 2 children, as the 2C1NN rule, we obtain with the same proof as in [1] the following proposition.

PROPOSITION B.2. *Fix $0 < \delta \leq 1$. Let $(\mathcal{X}, \mathcal{B})$ be a separable Borel space. For the binary classification setting, the learning rule $(1+\delta)$C1NN is universally consistent for all processes $\mathbb{X} \in$ SMV.*

Finally, we use a result from [2] which gives a reduction from any near-metric bounded value space to binary classification.

THEOREM B.3 ([2]). *If $(1+\delta)$C1NN is universally consistent under a process $\mathbb{X}$ for binary classification, it is also universally consistent under $\mathbb{X}$ for any separable near-metric setting $(\mathcal{Y}, \ell)$ with bounded loss.*

Together with Proposition B.2, Theorem B.3 ends the proof of Theorem 4.1.

B.2. Proof of Theorem 4.3. Let $0 < \epsilon \leq 1$. We first analyze the prediction of the learning rule $f^\epsilon_\cdot$. In the rest of the proof, we denote $\bar{\ell}(\hat{Y}_t(\epsilon), Y_t) := \sum_{y\in\mathcal{Y}_\epsilon} \mathbb{P}(\hat{Y}_t(\epsilon) = y)\ell(y, Y_t)$ the immediate expected loss at each iteration. The learning rule was constructed so that we perform exactly the classical Hedge / exponentially weighted average forecaster on each cluster of times $\mathcal{C}(t) = \{u \leq t : u \overset{\phi}{\sim} t\}$. As a result [3] (Theorem 2.2), we have that for any $t \geq 1$,

$$\begin{aligned}
\frac{1}{\bar{\ell}} \sum_{u\in\mathcal{C}(t)} \bar{\ell}(\hat{Y}_u(\epsilon), Y_u) &\leq \frac{1}{\bar{\ell}} \min_{y\in\mathcal{Y}_\epsilon} \sum_{u\in\mathcal{C}(t)} \ell(y, Y_u) + \frac{\ln|\mathcal{Y}_\epsilon|}{\bar{\ell}\eta_\epsilon} + \frac{|\mathcal{C}(t)|\bar{\ell}\eta_\epsilon}{8} \\
&\leq \frac{1}{\bar{\ell}} \min_{y\in\mathcal{Y}_\epsilon} \sum_{u\in\mathcal{C}(t)} \ell(y, Y_u) + \sqrt{\frac{\ln|\mathcal{Y}_\epsilon|}{8T_\epsilon}}(T_\epsilon + |\mathcal{C}(t)|) \\
&\leq \frac{1}{\bar{\ell}} \min_{y\in\mathcal{Y}_\epsilon} \sum_{u\in\mathcal{C}(t)} \ell(y, Y_u) + \frac{\epsilon}{\bar{\ell}} \max(T_\epsilon, |\mathcal{C}(t)|)
\end{aligned}$$

Now consider a horizon $T \geq 1$, and enumerate all the clusters $\mathcal{C}_1(T), \ldots, \mathcal{C}_{p(T)}(T)$ at horizon T, i.e. the classes of equivalence of ϕ among the times $\{t \leq T\}$. Note that if a cluster $i \leq p$ has $|\mathcal{C}_i(T)| < T_\epsilon$, then either it must contain a time $t \in \mathcal{N}$ which is a leaf of the tree formed by ϕ until time T, or it is a cluster of duplicates of an instance X_u which has already had $\frac{T_\epsilon}{\epsilon}$ occurrences. As a result, the times falling into such clusters of duplicates with less than T_ϵ members form at most a proportion ϵ of the total T times. Denote by $\mathcal{A}_i := \{t \leq T : t \in \mathcal{N}, |\{u \leq T : \phi(u) = t\}| = i\}$ times which have excactly i children for $i \in \{0, 1, 2\}$. Note that no time can have more than 2 children. In particular $\mathcal{A}_0$ is the set of leaves. Then, by summing the above equations we obtain

$$\begin{aligned}\sum_{t=1}^{T} \bar{\ell}(\hat{Y}_t(\epsilon), Y_t) &\leq \sum_{i=1}^{p(T)} \left(\min_{y \in \mathcal{Y}_\epsilon} \sum_{u \in \mathcal{C}_i(T)} \ell(y, Y_u) + \epsilon \max(T_\epsilon, |\mathcal{C}_i(T)|) \right) \\ &\leq \sum_{i=1}^{p(T)} \min_{y \in \mathcal{Y}_\epsilon} \sum_{u \in \mathcal{C}_i(T)} \ell(y, Y_u) + \epsilon T + T_\epsilon |\{1 \leq i \leq p : |\mathcal{C}_i(T)| < T_\epsilon\}| \\ &\leq \sum_{i=1}^{p(T)} \min_{y \in \mathcal{Y}_\epsilon} \sum_{u \in \mathcal{C}_i(T)} \ell(y, Y_u) + \epsilon T + T_\epsilon |\mathcal{A}_0| + \epsilon T_\epsilon,\end{aligned}$$

where in the last inequality we used the fact that all clusters with $|\mathcal{C}_i(T)| < T_\epsilon$ contain a leaf from $\mathcal{A}_0$, which is therefore distinct for each such cluster. Now note that by counting the number of edges of the tree structure we obtain $\frac{1}{2}(3|\mathcal{A}_2| + 2|\mathcal{A}_1| + |\mathcal{A}_0| - 1) = T - 1 = |\mathcal{A}_0| + |\mathcal{A}_1| + |\mathcal{A}_2| - 1$, where the -1 on the left-hand side accounts for the root of this tree which does not have a parent. Hence we obtain $|\mathcal{A}_0| = |\mathcal{A}_2| + 1$. Further, $|\mathcal{A}_2| \leq |\{t \leq T : U_t = 1\}|$ which follows a binomial distribution $\mathcal{B}(T, \delta_\epsilon)$. Therefore, using the Chernoff bound, with probability $1 - e^{-T\delta_\epsilon/3}$ we have

$$\begin{aligned}\sum_{t=1}^{T} \bar{\ell}(\hat{Y}_t(\epsilon), Y_t) &\leq \sum_{i=1}^{p(T)} \min_{y \in \mathcal{Y}_\epsilon} \sum_{u \in \mathcal{C}_i(T)} \ell(y, Y_u) + 2\epsilon T + T_\epsilon(1 + 2T\delta_\epsilon) \\ &\leq \sum_{i=1}^{p(T)} \min_{y \in \mathcal{Y}_\epsilon} \sum_{u \in \mathcal{C}_i(T)} \ell(y, Y_u) + T_\epsilon + 3\epsilon T.\end{aligned}$$

We now observe that the sequence $\{\ell(\hat{Y}_t(\epsilon), Y_t) - \bar{\ell}(\hat{Y}_t(\epsilon), Y_t)\}_{T \geq 1}$ is a sequence of martingale differences bounded by $\bar{\ell}$ in absolute value. Hence, the Hoeffding-Azuma inequality yields that for any $T \geq 1$, with probability $1 - \frac{1}{T^2} - e^{-T\delta_\epsilon/3}$,

$$\sum_{t=1}^{T} \ell(\hat{Y}_t(\epsilon), Y_t) \leq \sum_{i=1}^{p(T)} \min_{y \in \mathcal{Y}_\epsilon} \sum_{u \in \mathcal{C}_i(T)} \ell(y, Y_u) + T_\epsilon + 3\epsilon T + 2\bar{\ell}\sqrt{T \ln T}.$$

Because $\sum_{T \geq 1} \frac{1}{T^2} + e^{-T\delta_\epsilon/3} < \infty$ the Borel-Cantelli lemma implies that with probability one, there exists a time $\hat{T}$ such that

$$\forall T \geq \hat{T}, \quad \sum_{t=1}^{T} \ell(\hat{Y}_t(\epsilon), Y_t) \leq \sum_{i=1}^{p(T)} \min_{y \in \mathcal{Y}_\epsilon} \sum_{u \in \mathcal{C}_i(T)} \ell(y, Y_u) + T_\epsilon + 2\bar{\ell}\sqrt{T \ln T} + 3\epsilon T.$$

We denote by $\mathcal{E}_\epsilon$ this event. We are now ready to analyze the risk of the learning rule $f^\epsilon_\cdot$. Let $f : \mathcal{X} \to \mathcal{Y}$ a measurable function to which we compare the prediction of $f^\epsilon_\cdot$. By Theorem 4.1, the rule $(1+\delta_\epsilon)$C1NN is optimistically universal in the noiseless setting. Therefore, because $\mathbb{X}\in$ SOUL we have in particular

$$\frac{1}{T}\sum_{t=1}^T \ell((1+\delta_\epsilon)C1NN_t(\mathbb{X}_{\leq t-1}, f(\mathbb{X}_{\leq t-1}), X_t), f(X_t)) \to 0 \quad (a.s.),$$

i.e., almost surely, $\frac{1}{T}\sum_{t\leq T, t\in\mathcal{N}} \ell(f(X_{\phi(t)}), f(X_t)) \to 0$ — the times corresponding to duplicate instances incur a 0 loss by memorization. We denote by $\mathcal{F}_\epsilon$ this event of probability one. Using Lemma A.1, we write for any $u=1,\ldots,T_\epsilon-1$,

$$\begin{aligned}
&\sum_{t\leq T, t\in\mathcal{N}} \ell(f(X_{\phi^u(t)}), f(X_t)) \\
&\leq 2^{\alpha-1}\sum_{t\leq T, t\in\mathcal{N}} \ell(f(X_{\phi^{u-1}(t)}), f(X_t)) + 2^{\alpha-1}\sum_{t\leq T, t\in\mathcal{N}} \ell(f(X_{\phi^l(t)}), f(X_{\phi^{u-1}(t)})) \\
&\leq 2^{\alpha-1}\sum_{t\leq T, t\in\mathcal{N}} \ell(f(X_{\phi^{u-1}(t)}), f(X_t)) \\
&\qquad + 2^{\alpha-1}\sum_{t\leq T, t\in\mathcal{N}} \ell(f(X_{\phi(t)}), f(X_t))\cdot |\{l\leq T : \phi^{u-1}(l)=t\}| \\
&\leq 2^{\alpha-1}\sum_{t\leq T, t\in\mathcal{N}} \ell(f(X_{\phi^{u-1}(t)}), f(X_t)) + 2^{\alpha+u-2}\sum_{t\leq T, t\in\mathcal{N}} \ell(f(X_{\phi(t)}), f(X_t))
\end{aligned}$$

where we used the fact that times have at most 2 children. Therefore, iterating the above equations, we obtain that on $\mathcal{F}_\epsilon$, for any $u=1,\ldots,T_\epsilon-1$

$$\begin{aligned}
\frac{1}{T}\sum_{t\leq T, t\in\mathcal{N}} \ell(f(X_{\phi^u(t)}), f(X_t)) &\leq \left(\sum_{k=1}^u 2^{\alpha+k-2+(\alpha-1)(u-k)}\right)\frac{1}{T}\sum_{t\leq T, t\in\mathcal{N}} \ell(f(X_{\phi(t)}), f(X_t)) \\
&\leq \frac{2^{u\alpha}}{T}\sum_{t\leq T, t\in\mathcal{N}} \ell(f(X_{\phi(t)}), f(X_t)) \to 0.
\end{aligned}$$

In the rest of the proof, for any $y\in\mathcal{Y}$, we will denote by y^ϵ a value in the $\epsilon-$net $\mathcal{Y}_\epsilon$ such that $\ell(y,y^\epsilon)\leq\epsilon$. We now pose $\mu_\epsilon = \min\{0<\mu\leq 1 : c_\mu^\alpha \leq \frac{1}{\sqrt{\epsilon}}\}$ if the corresponding set is non-empty and $\mu_\epsilon = 1$ otherwise. Note that because c_μ^α is non-increasing in μ, we have $\mu_\epsilon \longrightarrow_{\epsilon\to 0} 0$. Now let $0<\mu\leq 1$. $\mu := \epsilon^{\frac{1}{\alpha+1}}$. Finally, for any cluster $\mathcal{C}_i(T)$, let $t_i = \min\{u\in\mathcal{C}_i(T)\}$. Putting everything together, on the event $\mathcal{E}_\epsilon\cap\mathcal{F}_\epsilon$, for any $T\geq\hat{T}$, we have

$$\begin{aligned}
\sum_{t=1}^T \ell(\hat{Y}_t(\epsilon), Y_t) &\leq \sum_{i=1}^{p(T)} \min_{y\in\mathcal{Y}_\epsilon}\sum_{u\in\mathcal{C}_i(T)} \ell(y, Y_u) + T_\epsilon + 2\bar{\ell}\sqrt{T\ln T} + 3\epsilon T \\
&\leq \sum_{i=1}^{p(T)}\sum_{u\in\mathcal{C}_i(T)} \ell(f(X_{t_i})^\epsilon, Y_u) + T_\epsilon\bar{\ell} + 2\bar{\ell}\sqrt{T\ln T} + 3\epsilon T \\
&\leq \sum_{i=1}^{p(T)}\sum_{u\in\mathcal{C}_i(T)} [c_{\mu_\epsilon}^\alpha \ell(f(X_{t_i})^\epsilon, f(X_{t_i})) + (c_{\mu_\epsilon}^\alpha)^2\ell(f(X_{t_i}), f(X_u)) \\
&\qquad + (1+\mu_\epsilon)^2\ell(f(X_u), Y_u)] + T_\epsilon\bar{\ell} + 2\bar{\ell}\sqrt{T\ln T} + 3\epsilon T
\end{aligned}$$

$$
\begin{aligned}
&\leq (1+\mu_\epsilon)^2 \sum_{t=1}^T \ell(f(X_t),Y_t) + (c^\alpha_{\mu_\epsilon})^2 \frac{T_\epsilon}{\epsilon} \sum_{u=1}^{T_\epsilon - 1} \sum_{t\leq T, t\in\mathcal{N}} \ell(f(X_t), f(X_{\phi^u(t)}))\\
&\qquad + T_\epsilon \bar\ell + 2\bar\ell\sqrt{T\ln T} + (3+c^\alpha_{\mu_\epsilon})\epsilon T\\
&\leq \sum_{t=1}^T \ell(f(X_t),Y_t) + \frac{(c^\alpha_{\mu_\epsilon})^2 T_\epsilon}{\epsilon} \sum_{u=1}^{T_\epsilon - 1} \sum_{t\leq T, t\in\mathcal{N}} \ell(f(X_t), f(X_{\phi^u(t)}))\\
&\qquad + T_\epsilon \bar\ell + 2\bar\ell\sqrt{T\ln T} + (3\epsilon + \epsilon c^\alpha_{\mu_\epsilon} + 3\mu_\epsilon) T,
\end{aligned}
$$

where in the third inequality we used Lemma A.1 twice, and in the fourth inequality we used the fact that clusters containing distinct instances have at most $\frac{T_\epsilon}{\epsilon}$ duplicates of each instance. Hence, for any $\epsilon < (c^\alpha_1)^{-2}$, on the event $\mathcal{E}_\epsilon \cap \mathcal{F}_\epsilon$, we obtain

$$
\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^T \ell(\hat Y_t(\epsilon), Y_t) - \ell(f(X_t),Y_t) \leq 3\epsilon + \epsilon c^\alpha_{\mu_\epsilon} + 3\mu_\epsilon \leq 3\epsilon + \sqrt{\epsilon} + 3\mu_\epsilon,
$$

where $\mu_\epsilon \longrightarrow_{\epsilon\to 0} 0$. We now denote $\delta_\epsilon := 2\epsilon + \sqrt{\epsilon} + 3\mu_\epsilon$ and $i_0 = \lceil \frac{2\ln c^\alpha_1}{\ln 2}\rceil$. We now turn to the final learning rule and show that by using the predictions of the rules $f^{\epsilon_i}_\cdot$ for $i\geq 0$, it achieves zero risk. First, by the union bound, on the event $\bigcap_{i\geq 0} \mathcal{E}_{\epsilon_i} \cap \mathcal{F}_{\epsilon_i}$ of probability one,

$$
\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^T \ell(\hat Y_t(\epsilon_i), Y_t) - \ell(f(X_t),Y_t) \leq \delta_{\epsilon_i}, \quad \forall i \geq i_0.
$$

Now define $\mathcal{H}$ the event probability one according to Lemma 4.2 such that there exists $\hat t$ for which

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

In the rest of the proof we will suppose that the event $\mathcal{H}\cap\bigcap_{i\geq 0} \mathcal{E}_{\epsilon_i} \cap \mathcal{F}_{\epsilon_i}$ is met. Let $i\geq i_0$. For any $T\geq \max(\hat t, t_i)$, we have

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^T \ell(\hat Y_t, Y_t) - \ell(f(X_t),Y_t) &\leq \frac{t_i}{T}\bar\ell + \frac{1}{T}\sum_{t=t_i}^T \ell(\hat Y_t, Y_t) - \ell(f(X_t),Y_t)\\
&\leq \frac{t_i}{T}\bar\ell + \frac{1}{T}\sum_{t=t_i}^T \ell(\hat Y_t(\epsilon_i), Y_t) - \ell(f(X_t),Y_t) + (2+\bar\ell+\bar\ell^2)\sqrt{\frac{\ln T}{T}}\\
&\leq \frac{1}{T}\sum_{t=1}^T \ell(\hat Y_t(\epsilon_i), Y_t) - \ell(f(X_t),Y_t) + \frac{2t_i}{T}\bar\ell + (2+\bar\ell+\bar\ell^2)\sqrt{\frac{\ln T}{T}}.
\end{aligned}
$$

Therefore we obtain $\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^T \ell(\hat Y_t, Y_t) - \ell(f(X_t),Y_t) \leq \delta_{\epsilon_i}$. Because this holds for any $i\geq i_0$ on the event $\mathcal{H}\cap\bigcap_{i\geq 0} \mathcal{E}_{\epsilon_i} \cap \mathcal{F}_{\epsilon_i}$ of probability one, and $\delta_{\epsilon_i}\to 0$ for $i\to\infty$, we have

$$
\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^T \ell(\hat Y_t, Y_t) - \ell(f(X_t),Y_t) \leq 0.
$$

This ends the proof of the theorem.

B.3. Proof of Lemma 4.2. We first introduce the following helper lemma which can be found in [3].

LEMMA B.4 ([3]). *For all $N \geq 2$, for all $\beta \geq \alpha \geq 0$ and for all $d_1, \ldots, d_N \geq 0$ such that $\sum_{i=1}^N e^{-\alpha d_i} \geq 1$,*

$$\ln \frac{\sum_{i=1}^N e^{-\alpha d_i}}{\sum_{i=1}^N e^{-\beta d_i}} \leq \frac{\beta - \alpha}{\alpha} \ln N.$$

We are now ready to compare the predictions of the learning rule $f_\cdot$ to the predictions of the rules $f_\cdot^\epsilon$.

For any $t \geq 0$, we define the instantaneous regret $r_{t,i} = \hat{\ell}_t - \ell(\hat{Y}_t(\epsilon_i), Y_t)$. We first note that $|r_{t,i}| \leq \bar{\ell}$. We now define $w'_{t-1,i} := e^{\eta_{t-1}(\hat{L}_{t-1,i} - L_{t-1,i})}$. We also introduce $W_{t-1} = \sum_{i \in I_t} w_{t-1,i}$ and $W'_{t-1} = \sum_{i \in I_{t-1}} w'_{t-1,i}$. We denote the index $k_t \in I_t$ such that $\hat{L}_{t,k_t} - L_{t,k_t} = \max_{i \in I_t} \hat{L}_{t,i} - L_{t,i}$. Then we write

$$\frac{1}{\eta_t} \ln \frac{w_{t-1,k_{t-1}}}{W_{t-1}} - \frac{1}{\eta_{t+1}} \ln \frac{w_{t,k_t}}{W_t} = \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \ln \frac{W_t}{w_{t,k_t}} + \frac{1}{\eta_t} \ln \frac{W_t / w_{t,k_t}}{W'_t / w'_{t,k_t}} \\ + \frac{1}{\eta_t} \ln \frac{w_{t-1,k_{t-1}}}{w'_{t,k_t}} + \frac{1}{\eta_t} \ln \frac{W'_t}{W_{t-1}}.$$

By construction, we have $\ln \frac{W_t}{w_{t,k_t}} \leq \ln |I_t| \leq \ln(1 + \ln t)$. Further, we have that

$$\begin{aligned} \frac{1}{\eta_t} \ln \frac{W_t / w_{t,k_t}}{W'_t / w'_{t,k_t}} &= \frac{1}{\eta_t} \ln \frac{\sum_{i \in I_{t+1}} e^{\eta_{t+1}(\hat{L}_{t,i} - L_{t,i} - \hat{L}_{t,k_t} + L_{t,k_t})}}{\sum_{i \in I_t} e^{\eta_t(\hat{L}_{t,i} - L_{t,i} - \hat{L}_{t,k_t} + L_{t,k_t})}} \\ &= \frac{1}{\eta_t} \ln \frac{\sum_{i \in I_{t+1}} w_{t,i}}{\sum_{i \in I_t} w_{t,i}} + \frac{1}{\eta_t} \ln \frac{\sum_{i \in I_{t+1}} e^{\eta_{t+1}(\hat{L}_{t,i} - L_{t,i} - \hat{L}_{t,k_t} + L_{t,k_t})}}{\sum_{i \in I_{t+1}} e^{\eta_t(\hat{L}_{t,i} - L_{t,i} - \hat{L}_{t,k_t} + L_{t,k_t})}} \\ &\leq \frac{1}{\eta_t} \ln \frac{\sum_{i \in I_{t+1}} w_{t,i}}{\sum_{i \in I_t} w_{t,i}} + \frac{1}{\eta_t} \left(\frac{\eta_t - \eta_{t+1}}{\eta_{t+1}} \right) \ln |I_{t+1}| \\ &\leq \frac{|I_{t+1}| - |I_t|}{\eta_t \sum_{i \in I_t} w_{t,i}} + \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \ln(1 + \ln(t+1)), \end{aligned}$$

where in the first inequality we applied Lemma B.4. We also have

$$\frac{1}{\eta_t} \ln \frac{w_{t-1,k_{t-1}}}{w'_{t,k_t}} = (\hat{L}_{t-1,k_{t-1}} - L_{t-1,k_{t-1}}) - (\hat{L}_{t,k_t}, L_{t,k_t}).$$

Last, because $|r_{t,i}| \leq \bar{\ell}$ for all $i \in I_t$, we can use Hoeffding's lemma to obtain

$$\frac{1}{\eta_t} \ln \frac{W'_t}{W_{t-1}} = \frac{1}{\eta_t} \ln \sum_{i \in I_t} \frac{w_{t-1,i}}{W_{t-1}} e^{\eta_t r_{t,i}} \leq \frac{1}{\eta_t} \left(\eta_t \sum_{i \in I_t} r_{t,i} \frac{w_{t-1,i}}{W_{t-1}} + \frac{\eta_t^2 (2\bar{\ell})^2}{8} \right) = \frac{1}{2} \eta_t \bar{\ell}^2.$$

Putting everything together gives

$$(1) \quad \frac{1}{\eta_t} \ln \frac{w_{t-1,k_{t-1}}}{W_{t-1}} - \frac{1}{\eta_{t+1}} \ln \frac{w_{t,k_t}}{W_t} \leq 2 \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \ln(1 + \ln(t+1)) + \frac{|I_{t+1}| - |I_t|}{\eta_t \sum_{i \in I_t} w_{t,i}} \\ + (\hat{L}_{t-1,k_{t-1}} - L_{t-1,k_{t-1}}) - (\hat{L}_{t,k_t} - L_{t,k_t}) + \frac{1}{2} \eta_t \bar{\ell}^2.$$

First suppose that we have $\sum_{i\in I_t} w_{t,i} \leq 1$. Then either $k_t \in I_{t+1}\setminus I_t$ in which case $\hat L_{t,k_t} - L_{t,k_t} = 0$, or we have directly

$$\hat L_{t,k_t} - L_{t,k_t} \leq \frac{1}{\eta_{t+1}} \ln\left[\sum_{i\in I_t} w_{t,i}\right] \leq 0.$$

Otherwise, let $t' = \min\{1\leq s\leq t : \forall s\leq s'\leq t, \sum_{i\in I_{s'}} w_{s',i}\geq 1\}$. We sum equation (1) for $s = t',\ldots,t$ which gives

$$\begin{aligned}\frac{1}{\eta_1}\ln\frac{w_{t'-1,k_{t'-1}}}{W_{t'-1}} - \frac{1}{\eta_{t+1}}\ln\frac{w_{t,k_t}}{W_t} &\leq \frac{2}{\eta_{t+1}}\ln(1+\ln(t+1)) + \frac{|I_{t+1}|}{\eta_t}\\ &+ (\hat L_{t'-1,k_{t'-1}} - L_{t'-1,k_{t'-1}}) - (\hat L_{t,k_t} - L_{t,k_t}) + \frac{\bar\ell^2}{2}\sum_{s=t'}^t \eta_s.\end{aligned}$$

Note that we have $\frac{w_{t,k_t}}{W_t}\leq 1$ and $\frac{w_{t'-1,k_{t'-1}}}{W_{t'-1}} \geq \frac{1}{|I_{t'-1}|}\geq \frac{1}{1+\ln t}$. Also, assuming $t'\geq 2$, since $\sum_{i\in I_{t'-1}} w_{t'-1,i} < 1$, we have for any $i\in I_{t'-1}$ that $\hat L_{t'-1,i} - L_{t'-1,i}\leq 0$, hence $\hat L_{t'-1,k_{t'-1}} - L_{t'-1,k_{t'-1}} \leq 0$. If $t'=1$ we have directly $\hat L_{0,k_0} - L_{0,k_0} = 0$. Finally, using the fact that $\sum_{s=1}^t \frac{1}{\sqrt s}\leq 2\sqrt t$, we obtain

$$\begin{aligned}\hat L_{t,k_t} - L_{t,k_t} &\leq \ln(1+\ln(t+1))\left(1+2\sqrt{\frac{t+1}{\ln(t+1)}}\right) + (1+\ln(t+1))\sqrt{\frac{t}{\ln t}} + \bar\ell^2\sqrt{t\ln t}\\ &\leq (3/2+\bar\ell^2)\sqrt{t\ln t},\end{aligned}$$

for all $t\geq t_0$ where t_0 is a fixed constant. This in turn implies that for all $t\geq t_0$ and $i\in I_t$, we have $\hat L_{t,i} - L_{t,i}\leq (3/2+\bar\ell^2)\sqrt{t\ln t}$. Now note that $|\ell(\hat Y_t, Y_t) - \hat\ell_t|\leq \bar\ell$. Hence, we can use Hoeffding-Azuma inequality for the variables $\ell(\hat Y_t,Y_t)-\hat\ell_t$ that form a sequence of martingale differences to obtain $\mathbb{P}\left[\sum_{s=t_i}^t \ell(\hat Y_s,Y_s) > \hat L_{t,i} + u\right]\leq e^{-\frac{2u^2}{t\bar\ell^2}}$. Hence, for $t\geq t_0$ and $i\in I_t$, with probability $1-\delta$, we have

$$\sum_{s=t_i}^t \ell(\hat Y_s,Y_s) \leq \hat L_{t,i} + \bar\ell\sqrt{\frac{t}{2}\ln\frac{1}{\delta}} \leq L_{t,i} + (3/2+\bar\ell^2)\sqrt{t\ln t} + \bar\ell\sqrt{\frac{t}{2}\ln\frac{1}{\delta}}.$$

Therefore, since $|I_t|\leq 1+\ln t$, by union bound with probability $1-\frac{1}{t^2}$ we obtain that for all $i\in I_t$,

$$\sum_{s=t_i}^t \ell(\hat Y_s,Y_s)\leq L_{t,i} + (3/2+\bar\ell^2)\sqrt{t\ln t} + \bar\ell\sqrt{\frac{t}{2}\ln(1+\ln t)} + \bar\ell\sqrt{t\ln t} \leq (2+\bar\ell+\bar\ell^2)\sqrt{t\ln t},$$

for all $t\geq t_1$ where $t_1\geq t_0$ is a fixed constant. The Borel-Cantelli lemma implies that almost surely, there exists $\hat t\geq 0$ such that

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

This ends the proof of the lemma.

APPENDIX C: PROOFS OF SECTION 5

C.1. Proof of Theorem 5.1. We start by checking that with the defined loss $(\mathbb{N},\ell)$ is indeed a metric space $(\mathbb{N},\ell)$. We only have to check that the triangular inequality is satisfied, the other properties of a metric being directly satisfied. By construction, the loss has values in $\{0,\frac{1}{2},1\}$. Now let $i,j,k\in\mathbb{N}$. The triangular inequality $\ell(i,j)\leq \ell(i,k)+\ell(k,j)$ is directly satisfied if two of these indices are equal. Therefore, we can suppose that they are all distinct and as a result $\ell(i,j),\ell(i,k),\ell(k,j)\in\{\frac{1}{2},1\}$. Therefore

$$\ell(i,j)\leq 1\leq \ell(i,k)+\ell(k,j),$$

which ends the proof that ℓ is a metric.

Now let $(\mathcal{X},\mathcal{B})$ be a separable metrizable Borel space. Let $\mathbb{X}\notin$ CS. We aim to show that universal online learning under adversarial responses is not achievable under $\mathbb{X}$. Because $\mathbb{X}\notin$ CS, there exists a sequence of decreasing measurable sets $\{A_i\}_{i\geq 1}$ with $A_i\downarrow\emptyset$ such that $\mathbb{E}[\hat\mu_{\mathbb{X}}(A_i)]$ does not converge to 0 for $i\to\infty$. In particular, there exist $\epsilon>0$ and an increasing subsequence $(i_l)_{l\geq 1}$ such that $\mathbb{E}[\hat\mu_{\mathbb{X}}(A_{i_l})]\geq\epsilon$ for all $l\geq 1$. We now denote $B_l:=A_{i_l}\setminus A_{i_{l+1}}$ for any $l\geq 1$. Then $\{B_l\}_{l\geq 1}$ forms a sequence of disjoint measurable sets such that

$$\mathbb{E}\left[\hat\mu_{\mathbb{X}}\left(\bigcup_{l'\geq l}B_{l'}\right)\right]\geq\epsilon,\quad l\geq 1.$$

Therefore, for any $l\geq 1$ because $\mathbb{E}\left[\hat\mu_{\mathbb{X}}\left(\bigcup_{l'\geq l}B_{l'}\right)\right]\leq\mathbb{P}\left[\hat\mu_{\mathbb{X}}\left(\bigcup_{l'\geq l}B_{l'}\right)\geq\frac{\epsilon}{2}\right]+\frac{\epsilon}{2}$ we obtain

$$\mathbb{P}\left[\hat\mu_{\mathbb{X}}\left(\bigcup_{l'\geq l}B_{l'}\right)\geq\frac{\epsilon}{2}\right]\geq\frac{\epsilon}{2}.$$

Now because $\hat\mu$ is increasing we obtain

$$\begin{aligned}\mathbb{P}\left[\hat\mu_{\mathbb{X}}\left(\bigcup_{l'\geq l}B_{l'}\right)\geq\frac{\epsilon}{2},\forall l\geq 1\right]&=\lim_{L\to\infty}\mathbb{P}\left[\hat\mu_{\mathbb{X}}\left(\bigcup_{l'\geq l}B_{l'}\right)\geq\frac{\epsilon}{2},1\leq l\leq L\right]\\&=\lim_{L\to\infty}\mathbb{P}\left[\hat\mu_{\mathbb{X}}\left(\bigcup_{l'\geq L}B_{l'}\right)\geq\frac{\epsilon}{2}\right]\geq\frac{\epsilon}{2}.\end{aligned}$$

We will denote by $\mathcal{A}$ this event in which for all $l\geq 1$, we have $\hat\mu_{\mathbb{X}}\left(\bigcup_{l'\geq l}B_{l'}\right)\geq\frac{\epsilon}{2}$. Under the event $\mathcal{A}$, for any $l,t^0\geq 1$, there always exists $t^1>t^0$ such that $\frac{1}{t^1}\sum_{t=1}^{t^1}\mathbb{1}_{\bigcup_{l'\geq l}B_{l'}}(X_t)\geq\frac{3\epsilon}{8}$. We construct a sequence of times $(t_p)_{p\geq 1}$ and indices $(l_p)_{p\geq 1}$, $(u_p)_{p\geq 1}$ by induction as follows. We first pose $u_0=t_0=0$. Now assume that for $p\geq 1$, the time t_{p-1} and index u_{p-1} are defined. We first construct an index $l_p>u_{p-1}$ such that

$$\mathbb{P}\left[\mathbb{X}_{\leq t_{p-1}}\cap\left(\bigcup_{l\geq l_p}B_l\right)\neq\emptyset\right]\leq\frac{\epsilon}{2^{p+3}}.$$

We will denote by $\mathcal{E}_p$ the complementary of this event. Note that finding such index l_p is possible because the considered events $\{\mathbb{X}_{\leq t_{p-1}}\cap\left(\bigcup_{l'\geq l}B_{l'}\right)\neq\emptyset\}$ are decreasing as $l>u_{p-1}$ increases and we have $\bigcap_{l>u_{p-1}}\left\{\mathbb{X}_{\leq t_{p-1}}\cap\left(\bigcup_{l'\geq l}B_{l'}\right)\neq\emptyset\right\}=$

$\left\{\mathbb{X}_{\leq t_{p-1}} \cap \left(\bigcap_{l>u_{p-1}} \bigcup_{l'\geq l} B_{l'}\right) \neq \emptyset\right\} = \emptyset$. We then construct $t_p > t_{p-1}$ such that

$$\mathbb{P}\left[\mathcal{A}^c \cup \bigcup_{t_{p-1}<t\leq t_p} \left\{\frac{1}{t}\sum_{u=1}^{t} \mathbb{1}_{\bigcup_{l\geq l_p} B_l}(X_u) \geq \frac{3\epsilon}{8}\right\}\right] \geq 1 - \frac{\epsilon}{2^{p+4}}.$$

This is also possible because $\mathcal{A} \subset \bigcup_{t>\frac{8}{\epsilon}t_{p-1}} \left\{\frac{1}{t}\sum_{u=1}^{t} \mathbb{1}_{\bigcup_{l\geq l_p} B_l}(X_u) \geq \frac{3\epsilon}{8}\right\}$. Last, we can now construct $u_p \geq l_p$ such that

$$\mathbb{P}\left[\mathcal{A}^c \cup \bigcup_{t_{p-1}<t\leq t_p} \left\{\frac{1}{t}\sum_{u=1}^{t} \mathbb{1}_{\bigcup_{l_p\leq l\leq u_p} B_l}(X_u) \geq \frac{\epsilon}{4}\right\}\right] \geq 1 - \frac{\epsilon}{2^{p+3}},$$

which is possible using similar arguments as above. We denote $\mathcal{F}_p$ this event. This ends the recursive construction of times t_p and indices l_p for all $p \geq 1$. Note that by construction, $\mathbb{P}[\mathcal{E}_p^c], \mathbb{P}[\mathcal{F}_p^c] \leq \frac{\epsilon}{2^{p+3}}$. Hence, by union bound, the event $\mathcal{A} \cap \bigcap_{p\geq 1}(\mathcal{E}_p \cap \mathcal{F}_p)$ has probability $\mathbb{P}[\mathcal{A} \cap \bigcap_{p\geq 1}(\mathcal{E}_p \cap \mathcal{F}_p)] \geq \mathbb{P}[\mathcal{A}] - \frac{\epsilon}{4} \geq \frac{\epsilon}{4}$. To simplify the rest of the proof, we denote $\tilde{B}_p = \bigcup_{l_p\leq l\leq u_p} B_l$ for any $p \geq 1$. Also, for any $t_1 \leq t_2$, we denote by

$$N_p(t_1, t_2) = \sum_{t=t_1}^{t_2} \mathbb{1}_{\tilde{B}_p}(X_t)$$

the number of times that set $\tilde{B}_p$ has been visited between times t_1 and t_2.

We now fix a learning rule $f.$ and construct a process $\mathbb{Y}$ for which consistency will not be achieved on the event $\mathcal{A} \cap \bigcap_{p\geq 1}(\mathcal{E}_p \cap \mathcal{F}_p)$. Precisely, we first construct a family of processes $\mathbb{Y}^b$ indexed by a sequence of binary digits $b = (b_i)_{i\geq 1}$. The process $\mathbb{Y}^b$ is defined such that for any $p \geq 1$, and for all $t_{p-1} < t \leq t_p$,

$$Y_t^b := \begin{cases} n_{t_p} + 4u_p(t) + 2b_{i(p,u_p(t))} + b_{i(p,u_p(t))+1} & \text{if } X_t \in \tilde{B}_p, \\ n_{t_{p'}} + 4t_{p'} + \{b_{i(p',t_{p'}-1)} \dots b_{i(p',1)} b_{i(p',0)}\}_2 & \text{if } X_t \in \tilde{B}_{p'}, p' < p, \\ 0 & \text{otherwise,} \end{cases}$$

where we denoted $u_p(t) = N_p(t_{p-1}+1, t-1)$ and posed for any $p \geq 1$ and $u \geq 1$:

$$i(p,u) = 2\sum_{p'<p} t_{p'} + 2u.$$

Note in particular that conditionally on $\mathbb{X}$, $\mathbb{Y}^b$ is deterministic: it does not depends on the random predictions of the learning rule. Because we always have $N_p(t_{p-1}+1, t-1) \leq t_p$ for any $t \leq t_p$, the process is designed so that we have $Y_t^b \in I_{t_p}$ if $X_t \in \tilde{B}_p$ and $t_{p-1} < t \leq t_p$. Further, for $t_{p-1} < t \leq t_p$, if $X_t \in \bigcup_{p'<p} \tilde{B}_{p'}$ then $Y_t^b \in J_{t_{p'}}$. We now consider an i.i.d. Bernoulli $\mathcal{B}(\frac{1}{2})$ sequence of random bits $\boldsymbol{b}$ independent from the process $\mathbb{X}$—and any learning rule predictions. We analyze the responses of the learning rule for responses $\mathbb{Y}^{\boldsymbol{b}}$. We first fix a realization $\boldsymbol{x}$ of the process $\mathbb{X}$, which falls in the event $\mathcal{A} \cap \bigcap_{p\geq 1}(\mathcal{E}_p \cap \mathcal{F}_p)$. For any $p \geq 1$ we define $\mathcal{T}_p := \{t_{p-1} < t \leq t_p : \ x_t \in \tilde{B}_p\}$. For simplicity of notation, for any $t \in \mathcal{T}_p$ we denote $i(t) = i(p, u_p(t))$. We will also denote $\hat{Y}_t := f_t(\boldsymbol{x}_{<t}, \mathbb{Y}^{\boldsymbol{b}}_{<t}, x_t)$. Last, denote by r_t the possible randomness used by the learning rule f_t at time t. For any $t \in \mathcal{T}_p$, we have

$$\begin{aligned} \mathbb{E}_{\boldsymbol{b},\boldsymbol{r}} \ell(\hat{Y}_t, Y_t^{\boldsymbol{b}}) &= \mathbb{E}_{\{b_{i(p',u')}, b_{i(p',u')+1},\, p'\leq p, u'\leq t_{p'}\} \cup \{r_{t'}, t'\leq t\}} \ell(\hat{Y}_t, Y_t^{\boldsymbol{b}}) \\ &= \mathbb{E}\left[\mathbb{E}_{b_{i(t)}, b_{i(t)+1}}\ \ell(\hat{Y}_t, Y_t^{\boldsymbol{b}}) \middle| b_{i(t')}, b_{i(t')+1}, t' < t, t' \in \mathcal{T}_p;\ b_i, i < i(p,0);\ r_{t'}, t' \leq t\right] \end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left[\mathbb{E}_{b_{i(t)},b_{i(t)+1}}\ \ell(\hat{Y}_t, Y_t^{\boldsymbol{b}})\middle| \hat{Y}_t\right]\\
&= \mathbb{E}_{\hat{Y}_t}\left[\frac{1}{4}\sum_{m=0}^{3}\ell(\hat{Y}_t, n_{t_p}+4u_p(t)+m)\right]\\
&= \mathbb{E}_{\hat{Y}_t}\left[\mathbb{1}_{\hat{Y}_t\notin\{n_{t_p}+4u_p(t)+m, 0\leq m\leq 3\}\cup J_{t_p}} + \frac{3}{4}\mathbb{1}_{\hat{Y}_t\in\{n_{t_p}+4u_p(t)+m,0\leq m\leq 3\}} + \frac{3}{4}\mathbb{1}_{\hat{Y}_t\in J_{t_p}}\right]\\
&\geq \frac{3}{4}.
\end{aligned}
$$

where in the last equality, we used the fact that if $j\in J_{k(t)}$ then by construction $\ell(j, n_{t_p}+4u_p(t)) = \ell(j, n_{t_p}+4u_p(t)+1)$, $\ell(j,n_{t_p}+4u_p(t)+2) = \ell(j,n_{t_p}+4u_p(t)+3)$, and $\{\ell(j,n_{t_p}+4u_p(t)), \ell(j,n_{t_p}+4u_p(t)+2)\} = \{\frac{1}{2},1\}$. Summing all equations, we obtain for any $t_{p-1}<T\leq t_p$,

$$
\mathbb{E}_{\boldsymbol{b},\boldsymbol{r}}\left[\sum_{t=1}^{T}\ell(f_t(\boldsymbol{x}_{<t},\mathbb{Y}^{\boldsymbol{b}}_{<t},x_t),Y_t^{\boldsymbol{b}})\right] \geq \frac{3}{4}\sum_{p'<p}|\mathcal{T}_{p'}| + \frac{3}{4}|\mathcal{T}_p\cap\{t\leq T\}|.
$$

This holds for all $p\geq 1$. Let us now compare this loss to the best prediction of a fixed measurable function. Specifically, for any binary sequence b, we consider the following function $f^b:\mathcal{X}\to\mathbb{N}$:

$$
f^b(x) = \begin{cases} n_{t_p}+4t_p+\{b_{i(p,t_p-1)}\dots b_{i(p,1)}b_{i(p,0)}\}_2 & \text{if } x\in\tilde{B}_p\\ 0 & \text{if } x\notin\bigcup_{p\geq 1}\tilde{B}_p.\end{cases}
$$

Now let $t_{p-1}<t\leq t_p$ and $p\geq 1$. If $x_t\in\bigcup_{p'<p}\tilde{B}_{p'}$ we have $f^{\boldsymbol{b}}(x_t) = Y_t^{\boldsymbol{b}}$, hence $\ell(f^{\boldsymbol{b}}(x_t),Y_t^{\boldsymbol{b}}) = 0$. Now if $X_t\in\tilde{B}_p$ by construction we have $\ell(f^{\boldsymbol{b}}(x_t),Y_t^{\boldsymbol{b}}) = \frac{1}{2}$. Finally, observe that because the event $\mathcal{E}_{p+1}$ is satisfied by $\boldsymbol{x}$ there does not exist $t_{p-1}<t\leq t_p$ such that $t\in\bigcup_{p'>p}\tilde{B}_{p'}\subset\bigcup_{l\geq l_{p+1}}B_l$. As a result, we have $\ell(f^{\boldsymbol{b}}(x_t),Y_t^{\boldsymbol{b}}) = \frac{1}{2}\mathbb{1}_{t\in\mathcal{T}_p}$ for any $t_{p-1}<t\leq t_p$. Thus, we obtain for any $t_{p-1}<T\leq t_p$,

$$
\mathbb{E}_{\boldsymbol{b},\boldsymbol{r}}\left[\sum_{t=1}^{T}\ell(\hat{Y}_t,Y_t^{\boldsymbol{b}}) - \ell(f^{\boldsymbol{b}}(X_t),Y_t^{\boldsymbol{b}})\right] \geq \frac{1}{4}\sum_{p'\leq p}|\mathcal{T}_{p'}| + \frac{1}{4}|\mathcal{T}_p\cap\{t\leq T\}| \geq \frac{1}{4}|\mathcal{T}_p\cap\{t\leq T\}|.
$$

Recall that the event $\mathcal{F}_p$ is satisfied by $\boldsymbol{x}$ for any $p\geq 1$. Therefore, there exists a time $t_{p-1}<T_p\leq t_p$ such that $\sum_{t=1}^{T_p}\mathbb{1}_{\tilde{B}_p}(x_t)\geq\frac{\epsilon T_p}{4}$. Then, note that because the event $\mathcal{E}_p$ is satisfied, we have $\sum_{t=1}^{t_{p-1}}\mathbb{1}_{\tilde{B}_p}(x_t) = 0$. Therefore, we obtain $|\mathcal{T}_p\cap\{t\leq T_p\}|\geq\frac{\epsilon T_p}{4}$, and as a result,

$$
\mathbb{E}_{\boldsymbol{b},\boldsymbol{r}}\left[\frac{1}{T_p}\sum_{t=1}^{T_p}\ell(\hat{Y}_t,Y_t^{\boldsymbol{b}}) - \ell(f^{\boldsymbol{b}}(X_t),Y_t^{\boldsymbol{b}})\right]\geq\frac{\epsilon}{16}.
$$

Because this holds for any $p\geq 1$ and as $p\to\infty$ we have $T_p\to\infty$, we can now use Fatou lemma which yields

$$
\mathbb{E}_{\boldsymbol{b},\boldsymbol{r}}\left[\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\ell(\hat{Y}_t,Y_t^{\boldsymbol{b}}) - \ell(f^{\boldsymbol{b}}(X_t),Y_t^{\boldsymbol{b}})\right]\geq\frac{\epsilon}{16}.
$$

This holds for any realization in $\mathcal{A}\cap\bigcap_{p\geq 1}(\mathcal{E}_p\cap\mathcal{F}_p)$ which we recall has probability at least $\frac{\epsilon}{4}$. Therefore we finally obtain

$$
\mathbb{E}_{\boldsymbol{b},\boldsymbol{r},\mathbb{X}}\left[\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\ell(\hat{Y}_t,Y_t^{\boldsymbol{b}}) - \ell(f^{\boldsymbol{b}}(X_t),Y_t^{\boldsymbol{b}})\right]\geq\frac{\epsilon^2}{2^6}.
$$

As a result, there exists a specific realization of $\boldsymbol{b}$ which we denote b such that

$$\mathbb{E}_{\boldsymbol{r},\mathbb{X}}\left[\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\ell(\hat{Y}_t,Y_t^b)-\ell(f^b(X_t),Y_t^b)\right]\geq\frac{\epsilon^2}{2^6},$$

which shows that with nonzero probability $\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\ell(\hat{Y}_t,Y_t^b)-\ell(f^b(X_t),Y_t^b)>0$. This ends the proof of the theorem. As a remark, one can note that the construction of our bad example $\mathbb{Y}^b$ is a deterministic function of $\mathbb{X}$: it is independent from the realizations of the randomness used by the learning rule.

C.2. Proof of Lemma 5.3. We first construct our online learning algorithm, which is a simple variant of the classical exponential forecaster. We first define a step $\eta:=\sqrt{2\ln t_0/t_0}$. At time $t=1$ we always predict 0. For time step $t\geq 2$, we define the set $S_{t-1}:=\{y\in\mathbb{N},\sum_{u=1}^{t-1}\mathbb{1}_{y=y_u}>0\}$ the set of values which have been visited. Then, we construct weights for all $y\in\mathbb{N}$ such that

$$w_{y,t-1}=\begin{cases}e^{\eta\sum_{u=1}^{t-1}\mathbb{1}_{y=y_u}}, & y\in S_{t-1}\\ 0 & \text{otherwise},\end{cases}$$

and output a randomized prediction independent of the past history such that

$$\mathbb{P}(\hat{y}_t=y)=\frac{w_{y,t-1}}{\sum_{y'\in\mathbb{N}}w_{y',t-1}}.$$

This defines a proper online learning rule. Note that the denominator is well defined since $w_{y,t-1}$ is non-zero only for values in S_{t-1}, which contains at most $t-1$ elements. We now define the expected success at time $1\leq t\leq T$ as $\hat{s}_t:=\frac{w_{y_t,t-1}}{\sum_{y\in\mathbb{N}}w_{y,t-1}}\mathbb{1}_{y_t\in S_t}$. Note that $\hat{s}_t=\mathbb{E}[\mathbb{1}_{f_t(\boldsymbol{y}_{\leq t-1})=y_t}]$. We first show that we have

$$\sum_{t=1}^{T}\hat{s}_t\geq\max_{y\in\mathbb{N}}\sum_{t=1}^{T}\mathbb{1}_{y=y_t}-\sqrt{T}\ln T.$$

To do so, we analyze the quantity $W_t:=\frac{1}{\eta}\ln\left(\sum_{y\in S_t}e^{\eta\sum_{u=1}^{t}(\mathbb{1}_{y=y_u}-\hat{s}_u)}\right)$. Let $2\leq t\leq T$. Supposing that $y_t\in S_{t-1}$, i.e., $S_t=S_{t-1}$, we define the operator $\Phi:\boldsymbol{x}\in\mathbb{R}^{|S_{t-1}|}\mapsto\frac{1}{\eta}\ln\left(\sum_{y\in S_{t-1}}e^{\eta x_y}\right)$ and use the Taylor expansion of Φ to obtain

$$\begin{aligned}
W_t&=\frac{1}{\eta}\ln\left(\sum_{y\in S_{t-1}}e^{\eta\sum_{u=1}^{t-1}(\mathbb{1}_{y=y_u}-\hat{s}_u)+\eta(\mathbb{1}_{y=y_t}-\hat{s}_t)}\right)\\
&=W_{t-1}+\sum_{y\in S_{t-1}}(\mathbb{1}_{y=y_t}-\hat{s}_t)\frac{e^{\eta\sum_{u=1}^{t-1}\mathbb{1}_{y=y_u}}}{\sum_{y'\in S_{t-1}}e^{\eta\sum_{u=1}^{t-1}\mathbb{1}_{y'=y_u}}}\\
&\qquad+\frac{1}{2}\sum_{y_1,y_2\in S_{t-1}}\left.\frac{\partial^2\Phi}{\partial x_{y_1}\partial x_{y_2}}\right|_{\xi}(\mathbb{1}_{y_1=y_u}-\hat{s}_u)(\mathbb{1}_{y_2=y_u}-\hat{s}_u)\\
&=W_{t-1}+\frac{1}{2}\sum_{y_1,y_2\in S_{t-1}}\left.\frac{\partial^2\Phi}{\partial x_{y_1}\partial x_{y_2}}\right|_{\xi}(\mathbb{1}_{y_1=y_t}-\hat{s}_u)(\mathbb{1}_{y_2=y_t}-\hat{s}_u)\\
&\leq W_{t-1}+\frac{1}{2}\sum_{y\in S_{t-1}}\frac{\eta e^{\eta\xi_y}}{\sum_{y'\in S_{t-1}}e^{\eta\xi_{y'}}}(\mathbb{1}_{y=y_t}-\hat{s}_u)^2
\end{aligned}$$

$$\leq W_{t-1} + \frac{\eta}{2},$$

for some vector $\xi \in \mathbb{R}^{|S_{t-1}|}$, where in the last inequality we used the fact $|\mathbb{1}_{y=y_t} - \hat{s}_u| \leq 1$. We now suppose that $y_t \notin S_{t-1}$ and $W_{t-1} \geq 1 + \frac{\ln 2 + 2\ln\frac{1}{\eta}}{\eta}$. In that case, $e^{\eta W_t} = e^{\eta W_{t-1}} + e^{\eta(1-\sum_{u=1}^{t-1}\hat{s}_u)}$. Hence, we obtain

$$W_t = W_{t-1} + \frac{\ln\left(1 + e^{\eta(1-W_{t-1}-\sum_{u=1}^{t-1}\hat{s}_u)}\right)}{\eta} \leq W_{t-1} + \frac{e^{\eta(1-W_{t-1})}}{\eta} \leq W_{t-1} + \frac{\eta}{2}.$$

Now let $l = \max\{1\} \cup \left\{1 \leq t \leq T : W_t < 1 + \frac{\ln 2 + 2\ln\frac{1}{\eta}}{\eta}\right\}$. Note that for any $l < t \leq T$ the above arguments yield $W_t \leq W_{t-1} + \frac{\eta}{2}$. As a result, noting that $W_1 \leq 1$, we finally obtain

$$W_T \leq W_l + \eta\frac{T-l}{2} \leq 1 + \frac{\ln 2 + 2\ln\frac{1}{\eta}}{\eta} + \eta\frac{T}{2} \leq 1 + \ln 2\sqrt{\frac{t_0}{2\ln t_0}} + \sqrt{\frac{\ln t_0}{2t_0}}(t_0+T).$$

Therefore, for any $y \in S_T$, we have

$$\sum_{t=1}^{T}(\mathbb{1}_{y=y_t} - \hat{s}_t) \leq W_T \leq 1 + \ln 2\sqrt{\frac{t_0}{2\ln t_0}} + \sqrt{\frac{\ln t_0}{2t_0}}(t_0+T).$$

In particular, this shows that

$$\sum_{t=1}^{T}\hat{s}_t \geq \max_{y\in S_T}\sum_{t=1}^{T}\mathbb{1}_{y=y_t} - 1 - \ln 2\sqrt{\frac{t_0}{2\ln t_0}} - \sqrt{\frac{\ln t_0}{2t_0}}(t_0+T).$$

Now note that if $y \notin S_T$, then $\sum_{t=1}^{T}\mathbb{1}_{y=y_t} = 0$, which yields $\max_{y\in S_T}\sum_{t=1}^{T}\mathbb{1}_{y=y_t} = \max_{y\in\mathbb{N}}\sum_{t=1}^{T}\mathbb{1}_{y=y_t}$. For the sake of conciseness, we will now denote by $\hat{y}_t$ the prediction of the online learning rule at time t. We observe that the variables $\mathbb{1}_{\hat{y}_t=y_t} - \hat{s}_t$ for $1 \leq t \leq T$ form a sequence of martingale differences. Further, $|\mathbb{1}_{\hat{y}_t=y_t} - \hat{s}_t| \leq 1$. Therefore, the Hoeffding-Azuma inequality shows that with probability $1-\delta$,

$$\sum_{t=1}^{T}(\mathbb{1}_{\hat{y}_t=y_t} - \hat{s}_t) \geq -\sqrt{2T\ln\frac{1}{\delta}}.$$

Putting everything together yields that with probability $1-\delta$,

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{1}_{\hat{y}_t=y_t} &\geq \sum_{t=1}^{T}\hat{s}_t - \sqrt{2T\ln\frac{1}{\delta}}\\ &\geq \max_{y\in\mathbb{N}}\sum_{t=1}^{T}\mathbb{1}_{y=y_t} - 1 - \ln 2\sqrt{\frac{t_0}{2\ln t_0}} - \sqrt{\frac{\ln t_0}{2t_0}}(t_0+T) - \sqrt{2T\ln\frac{1}{\delta}}.\end{aligned}$$

This ends the proof of the lemma.

C.3. Proof of Theorem 5.4. We use a similar learning rule to the one constructed in Section 4 for totally-bounded spaces. We only make a slight modification of the learning rules $f^\epsilon_\cdot$. Precisely, we pose for $0 < \epsilon \leq 1$,

$$T_\epsilon := \left\lceil \frac{2^4 \cdot 3^2(1+\ln\frac{1}{\epsilon})}{\epsilon^2} \right\rceil \quad \text{and} \quad \delta_\epsilon := \frac{\epsilon}{2T_\epsilon}.$$

Then, let ϕ be the representative function from the $(1+\delta_\epsilon)$C1NN learning rule. Similarly as for the $\epsilon-$approximation learning rule from Section 4, we consider the same equivalence relation $\overset{\phi}{\sim}$ on times to define clusters. The learning rule then performs its prediction based on the values observed on the corresponding cluster using the learning rule from Lemma 5.3 using $t_0 = T_\epsilon$. Precisely, let $\eta_\epsilon := \sqrt{2\ln T_\epsilon / T_\epsilon}$ and define the weights $w_{y,t} = e^{\eta_\epsilon \sum_{u<t: u \overset{\phi}{\sim} t} \mathbb{1}(Y_u = y)}$ for all $y \in \tilde{S} := \{y' \in \mathbb{N} : \sum_{u<t: u\overset{\phi}{\sim} t} \mathbb{1}(Y_u = y') > 0\}$ and $w_{y,t} = 0$ otherwise. The learning rule $f_t^\epsilon(\mathbb{X}_{\leq t-1}, \mathbb{Y}_{\leq t-1}, X_t)$ outputs a random value in $\mathbb{N}$ independent of the past history such that

$$\mathbb{P}(\hat{Y}_t = y) = \frac{w_{y,t}}{\sum_{y'\in\mathbb{N}} w_{y',t}}, \quad y \in \mathbb{N}.$$

The final learning rule $f.$ is then defined similarly as before from the learning rules $f_\cdot^\epsilon$ for $\epsilon > 0$. Therefore, Lemma 4.2 still holds. Also, using the same notations as in the proof of Theorem 4.3, Lemma 5.3 implies that for any $t \geq 1$, we can write for any $t \geq 1$ on the cluster $\mathcal{C}(t) = \{u < t : u \overset{\phi}{\sim} t\}$,

$$\begin{aligned}
\sum_{u\in\mathcal{C}(t)} \bar{\ell}_{01}(\hat{Y}_u(\epsilon), Y_u) &\leq \min_{y\in\mathbb{N}} \sum_{u\in\mathcal{C}(t)} \ell_{01}(y, Y_u) + 1 + \ln 2\sqrt{\frac{T_\epsilon}{2\ln T_\epsilon}} + \sqrt{\frac{\ln T_\epsilon}{2T_\epsilon}}(T_\epsilon + |\mathcal{C}(t)|) \\
&\leq \min_{y\in\mathbb{N}} \sum_{u\in\mathcal{C}(t)} \ell_{01}(y, Y_u) + \left(\frac{1}{T_\epsilon} + \frac{\ln 2}{\sqrt{2T_\epsilon \ln T_\epsilon}} + \sqrt{\frac{2\ln T_\epsilon}{T_\epsilon}}\right) \max(T_\epsilon, |\mathcal{C}(t)|) \\
&\leq \min_{y\in\mathbb{N}} \sum_{u\in\mathcal{C}(t)} \ell_{01}(y, Y_u) + \left(\frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3}\right) \max(T_\epsilon, |\mathcal{C}(t)|) \\
&= \min_{y\in\mathbb{N}} \sum_{u\in\mathcal{C}(t)} \ell_{01}(y, Y_u) + \epsilon \max(T_\epsilon, |\mathcal{C}(t)|)
\end{aligned}$$

Therefore, the same proof of Theorem 4.3 holds by replacing all $\epsilon-$nets $\mathcal{Y}_\epsilon$ directly by $\mathbb{N}$. The martingale argument still holds since the learning rule used is indeed online. This ends the proof of this theorem.

C.4. Proof of Theorem 5.5. We first need the following simple result which intuitively shows that we can assume that the predictions of the learning rule for mean estimation $g^\epsilon_{\leq t_\epsilon}$ are unrelated for $t = 1, \ldots, t_\epsilon$.

LEMMA C.1. *Let $(\mathcal{Y}, \ell)$ satisfying 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*

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

and such that the random variables $g_t(\boldsymbol{y}_{\leq t-1})$ are independent.

PROOF. Fix $\eta > 0$, $T_\eta \geq 1$ and $g_{\leq T_\eta}$ such that this online learning rule satisfies the condition from F-TiME for $\eta > 0$. We consider the following learning rule $\tilde{g}.$. For any $t \geq 1$ and $\boldsymbol{y} \in \mathcal{Y}^{t-1}$,

$$\tilde{g}_t(\boldsymbol{y}_{\leq t-1}) = g_t^t(\boldsymbol{y}_{\leq t-1}),$$

where $(g^t_\cdot)$ are i.i.d. samples of the learning rule $g_\cdot$. By construction, we still have that for any sequence $\boldsymbol{y}_{T_\eta} \in \mathcal{Y}^{T_\eta}$,

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

This ends the proof of the lemma. □

From now on, by Lemma C.1, we will suppose without loss of generality that the learning rule g^ϵ has predictions that are independent at each step (conditionally on the observed values). For simplicity, we refer to the prediction of the defined learning rule $f_\cdot$ (resp. $f^\epsilon_\cdot$) at time t as $\hat{Y}_t$ (resp. $\hat{Y}_t(\epsilon)$). We now show that is optimistically universal for arbitrary responses. By construction of the learning rule $f_\cdot$, Lemma 4.2 still holds. Therefore, we only have to focus on the learning rules $f^\epsilon_\cdot$ and prove that we obtain similar results as before. Let $T \geq 1$ and denote by $\mathcal{A}_i := \{t \leq T : |\{u \leq T : \phi(u) = t\}| = i\}$ the set of times which have exactly i children within horizon T for $i = 0, 1, 2$. Then, we define

$$\mathcal{B}_T = \{t \leq T : L_t = 0 \text{ and } |\{t < u \leq T : u \overset{\phi}{\sim} t\}| \geq t_\epsilon\},$$

i.e., times that start a new learning block and such that there are at least t_ϵ future times falling in their cluster within horizon T. Note that the function ψ defines a parent-relation (similarly to ϕ, but defined for all times $t \geq 1$). To simplify notations, for any $t \in \mathcal{B}_T$, we denote t^u the $\psi-$children of t at generation $u-1$ for $1 \leq u \leq t_\epsilon$, i.e., we have $\psi^{u-1}(t^u) = t$ for all $1 \leq u \leq t_\epsilon$. In particular $t = t^1$. By construction, blocks have length at most t_ϵ. More precisely, the block started at any $t \in \mathcal{B}_T$ has had time to finish completely, hence has length exactly t_ϵ. By construction of the indices L_t, the blocks $\{t^u, 1 \leq u \leq t_\epsilon\}$, for $t \in \mathcal{B}_T$, are all disjoint. This implies in particular $|\mathcal{B}_T| t_\epsilon \leq T$. We first analyze the predictions along these blocks and for any $t \in \mathcal{B}_T$ and $y \in \mathcal{Y}$, we pose $\delta_t(y) := \frac{1}{t_\epsilon}\sum_{u=1}^{t_\epsilon}\left(\ell(\hat{Y}_{t^u}, Y_{t^u}) - \ell(y, Y_{t^u}) - \epsilon\right)$. Now by construction of the learning rule $f^\epsilon_\cdot$, we have

$$t_\epsilon \delta_t(y^t) = \sum_{u=1}^{t_\epsilon}\left(\ell(g_u^{\epsilon,t}(\{Y_{t^l}\}_{l=1}^{u-1}), Y_{t^u}) - \ell(y^t, Y_{t^u})\right) - \epsilon t_\epsilon.$$

Next, for any $t \leq t_\epsilon$ and sequence $\boldsymbol{y}_{\leq t-1}$ and value $y \in \mathcal{Y}$, we write $\bar{\ell}(g^\epsilon_t(\boldsymbol{y}_{\leq t-1}), y) := \mathbb{E}\left[\ell(g^\epsilon_t(\boldsymbol{y}_{\leq t-1}), y)\right]$. Now by hypothesis on the learning rule $g^\epsilon_{\leq t_\epsilon}$,

$$\frac{1}{t_\epsilon}\sum_{u=1}^{t_\epsilon} \bar{\ell}(\hat{Y}_{t^u}, Y_{t^u}) - \ell(y^t, Y_{t^u}) \leq \epsilon. \tag{2}$$

Now consider the following sequence $(\ell(\hat{Y}_{t^u}, Y_{t^u}) - \bar{\ell}(\hat{Y}_{t^u}, Y_{t^u}))_{t \in \mathcal{B}_T, 1 \leq u \leq s(t)}$. Because of the definition of the learning rule, which uses i.i.d. copies of the learning rule $g^\epsilon_\cdot$, if we order the former sequence by increasing order of t^u, we obtain a sequence of martingale differences. We can continue this sequence by zeros to ensure that it has length exactly T. As a result, we obtain a sequence of T martingale differences, which are all bounded by $\bar{\ell}$ in absolute value. Now, the Azuma-Hoeffding inequality implies that for $\delta > 0$, with probability $1 - \delta$, we have

$$\sum_{t \in \mathcal{B}_T}\sum_{u=1}^{t_\epsilon} \ell(\hat{Y}_{t^u}, Y_{t^u}) \leq \sum_{t \in \mathcal{B}_T}\sum_{u=1}^{t_\epsilon} \bar{\ell}(\hat{Y}_{t^u}, Y_{t^u}) + \bar{\ell}\sqrt{2T \ln \frac{1}{\delta}}.$$

Thus, using Eq (2), with probability at least $1-\delta$,

$$\sum_{t\in\mathcal{B}_T} t_\epsilon \delta_t(y^t) \leq \bar\ell\sqrt{2T\ln\frac{1}{\delta}}. \tag{3}$$

We also denote $\mathcal{T}=\bigcup_{t\in\mathcal{B}_T}\{t^u, 1\leq u\leq t_\epsilon\}$ the union of all blocks within horizon T. This set contains all times $t\leq T$ except *bad* times close to the last times of their corresponding cluster $\{u\leq T: u\overset{\phi}{\sim} t\}$. Precisely, these are times t such that $|\{t<u\leq T: u\overset{\phi}{\sim} t\}|<t_\epsilon - L_t$. As a result, there are at most t_ϵ such times for each cluster. Using the same arguments as in the proof of Theorem 4.3, if we consider only clusters of duplicates (i.e., the cluster started for a specific instance which has high number of duplicates), the corresponding *bad* times contribute to a proportion $\leq \frac{t_\epsilon}{T_\epsilon/\epsilon}\leq\epsilon^2$ of times. Now consider clusters that have at least T_ϵ times. Their *bad* times contribute to a proportion $\leq\frac{t_\epsilon}{T_\epsilon}\leq\epsilon$ of times. Last, we need to account for clusters of size $<T_\epsilon$ which necessarily contain leaves of the tree ϕ: there are at most $|\mathcal{A}_0|$ such clusters. By the Chernoff bound, with probability at least $1-e^{-T\delta_\epsilon/3}$ we have

$$T-|\mathcal{T}|\leq(\epsilon^2+\epsilon)T+|\mathcal{A}_0|t_\epsilon\leq t_\epsilon+(\epsilon^2+\epsilon+2\delta_\epsilon t_\epsilon)T\leq t_\epsilon+3\epsilon T.$$

By the Borel-Cantelli lemma, because $\sum_{T\geq1}e^{-T\delta_\epsilon/3}<\infty$, almost surely there exists a time $\hat T$ such that for $T\geq\hat T$ we have $T-|\mathcal{T}|\leq t_\epsilon+3\epsilon T$. We denote by $\mathcal{E}_\epsilon$ this event. Then, on the event $\mathcal{E}_\epsilon$, for any $T\geq\hat T$ and for any sequence of values $(y^t)_{t\geq1}$ we have

$$\begin{aligned}\sum_{t=1}^T\ell(\hat Y_t(\epsilon),Y_t)&\leq\sum_{t\in\mathcal{B}_T}\sum_{u=1}^{t_\epsilon}\ell(\hat Y_{t^u},Y_{t^u})+(T-|\mathcal{T}|)\bar\ell\\&\leq\sum_{t\in\mathcal{B}_T}\sum_{u=1}^{t_\epsilon}\ell(y^t,Y_{t^u})+\sum_{t\in\mathcal{B}_T}t_\epsilon\delta_t(y^t)+\epsilon|\mathcal{B}_T|t_\epsilon+t_\epsilon\bar\ell+3\epsilon T\\&\leq\sum_{t\in\mathcal{B}_T}\sum_{u=1}^{t_\epsilon}\ell(y^t,Y_{t^u})+\sum_{t\in\mathcal{B}_T}t_\epsilon\delta_t(y^t)+t_\epsilon\bar\ell+4\epsilon T.\end{aligned}$$

Now let $f:\mathcal{X}\to\mathcal{Y}$ be a measurable function to which we compare $f^\epsilon_\cdot$. By Theorem 4.1, because $(1+\delta_\epsilon)$C1NN is optimistically universal without noise and $\mathbb{X}\in$ SOUL, almost surely $\frac{1}{T}\sum_{t=1}^T\ell(f(X_{\phi(t)}),f(X_t))\to0$. We denote by $\mathcal{F}_\epsilon$ this event of probability one. The proof of Theorem 4.3 shows that on $\mathcal{F}_\epsilon$, for any $0\leq u\leq T_\epsilon-1$ we have

$$\frac{1}{T}\sum_{t=1}^T\ell(f(X_{\phi^u(t)}),f(X_t))\to0.$$

We let $y^t=f(X_t)$ for all $t\geq1$. Then, recalling that for any $t\in\mathcal{B}_T$, we have $t=\phi^{u-1}(t^u)$, on the event $\mathcal{E}_\epsilon$, for any $T\geq\hat T$ we have

$$\begin{aligned}&\sum_{t=1}^T\ell(\hat Y_t(\epsilon),Y_t)\\&\leq\sum_{t\in\mathcal{B}_T}\sum_{u=1}^{t_\epsilon}((1+\epsilon)\ell(f(X_{t^u}),Y_{t^u})+c_\epsilon^\alpha\ell(f(X_t),f(X_{t^u})))+\sum_{t\in\mathcal{B}_T}t_\epsilon\delta_t(y^t)+t_\epsilon\bar\ell+4\epsilon T\\&\leq\sum_{t=1}^T\ell(f(X_t),Y_t)+c_\epsilon^\alpha\frac{T_\epsilon}{\epsilon}\sum_{u=0}^{T_\epsilon-1}\sum_{t=1}^T\ell(f(X_{\phi^u(t)}),f(X_t))+\sum_{t\in\mathcal{B}_T}t_\epsilon\delta_{\varphi(t)}(y^t)+t_\epsilon\bar\ell+5\epsilon T,\end{aligned}$$

where in the first inequality we used Lemma A.1, and in the second inequality we used the fact that cluster with distinct instance values have at most $\frac{T_\epsilon}{\epsilon}$ duplicates of each instance. Next, using Eq (3), with probability $1-\frac{1}{T^2}$, we have

$$\sum_{t\in\mathcal{B}_T} t_\epsilon \delta_t(y^t) \leq 2\bar{\ell}\sqrt{T\ln T}.$$

Because $\sum_{T\geq 1}\frac{1}{T^2}<0$, the Borel-Cantelli lemma implies that on an event $\mathcal{G}_\epsilon$ of probability one, there exists $\hat{T}_2$ such that for all $T\geq \hat{T}_2$ the above inequality holds. As a result, on the event $\mathcal{E}_\epsilon\cap\mathcal{F}_\epsilon\cap\mathcal{G}_\epsilon$ we obtain for any $T\geq\max(\hat{T},\hat{T}_2)$ that

$$\sum_{t=1}^T \ell(\hat{Y}_t(\epsilon),Y_t) \leq \sum_{t=1}^T \ell(f(X_t),Y_t) + \frac{c_\epsilon^\alpha T_\epsilon}{\epsilon}\sum_{u=0}^{T_\epsilon-1}\sum_{t=1}^T \ell(f(X_{\phi^u(t)}),f(X_t)) \\ + 2\bar{\ell}\sqrt{T\ln T} + t_\epsilon\bar{\ell} + 5\epsilon T.$$

where $\frac{1}{T}\sum_{u=0}^{T_\epsilon-1}\sum_{t=1}^T \ell(f(X_{\phi^u(t)}),f(X_t))\to 0$ because the event $\mathcal{F}_\epsilon$ is met. Therefore, we obtain that on the event $\mathcal{E}_\epsilon\cap\mathcal{F}_\epsilon\cap\mathcal{G}_\epsilon$ of probability one,

$$\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^T\left[\ell(\hat{Y}_t(\epsilon),Y_t)-\ell(f(X_t),Y_t)\right]\leq 5\epsilon,$$

i.e., almost surely, the learning rule $f^\epsilon_\cdot$ achieves risk at most 5ϵ compared to the fixed function f. By union bound, on the event $\bigcap_{i\geq 0}(\mathcal{E}_{\epsilon_i}\cap\mathcal{F}_{\epsilon_i}\cap\mathcal{G}_{\epsilon_i})$ of probability one we have that

$$\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^T\left[\ell(\hat{Y}_t(\epsilon_i),Y_t)-\ell(f(X_t),Y_t)\right]\leq 5\epsilon_i,\quad \forall i\geq 0.$$

The rest of the proof uses similar arguments as in the proof of Theorem 4.3. Precisely, let $\mathcal{H}$ be the almost sure event of Lemma 4.2 such that there exists $\hat{t}$ for which

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

In the rest of the proof we will suppose that the event $\mathcal{H}\cap\bigcap_{i\geq 0}(\mathcal{E}_{\epsilon_i}\cap\mathcal{F}_{\epsilon_i}\cap\mathcal{G}_{\epsilon_i})$ of probability one is met. Let $i\geq 0$. For all $t\geq\max(\hat{t},t_i)$ we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^T\ell(\hat{Y}_t,Y_t)-\ell(f(X_t),Y_t) &\leq \frac{t_i}{T}\bar{\ell}+\frac{1}{T}\sum_{t=t_i}^T\ell(\hat{Y}_t,Y_t)-\ell(f(X_t),Y_t)\\ &\leq \frac{t_i}{T}\bar{\ell}+\frac{1}{T}\sum_{t=t_i}^T\ell(\hat{Y}_t(\epsilon_i),Y_t)-\ell(f(X_t),Y_t)+(2+\bar{\ell}+\bar{\ell}^2)\sqrt{\frac{\ln T}{T}}\\ &\leq \frac{1}{T}\sum_{t=1}^T\ell(\hat{Y}_t(\epsilon_i),Y_t)-\ell(f(X_t),Y_t)+\frac{2t_i}{T}\bar{\ell}+(2+\bar{\ell}+\bar{\ell}^2)\sqrt{\frac{\ln T}{T}}.\end{aligned}$$

Therefore we obtain $\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^T\ell(\hat{Y}_t,Y_t)-\ell(f(X_t),Y_t)\leq 5\epsilon_i$. Because this holds for any $i\geq 0$ we finally obtain

$$\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^T\ell(\hat{Y}_t,Y_t)-\ell(f(X_t),Y_t)\leq 0.$$

As a result, $f_\cdot$ is universally consistent for adversarial responses under all SOUL processes. Hence, SOLAR = SOUL and $f_\cdot$ is in fact optimistically universal. This ends the proof of the theorem.

C.5. Proof of Lemma 5.7. We first note that with the same horizon time T_η, we have that F-TiME implies Property 2. We now show that Property 2 implies F-TiME. Let $(\mathcal{Y}, \ell)$ satisfying Property 2. We now fix $\eta > 0$ and let $T, g_{\leq\tau}$ such that for any $\boldsymbol{y} := (y_t)_{t=1}^T$ of values in $\mathcal{Y}$ and any value $y \in \mathcal{Y}$, we have

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

We now construct a random time $1 \leq \tilde\tau \leq T$ such that $\mathbb{P}[\tilde\tau = t] = \frac{\mathbb{P}[\tau=t]}{t\mathbb{E}[1/\tau]}$ for all $1 \leq t \leq T$. This indeed defines a proper random variable because $\sum_{t=1}^T \frac{\mathbb{P}[\tau=t]}{t\mathbb{E}[1/\tau]} = 1$. Let $Supp(\tau) := \{1 \leq t \leq T : \mathbb{P}[\tau = t] > 0\}$ be the support of τ. For any $t \in Supp(\tau)$, we denote by $g^t_{\leq t}$ the learning rule obtained by conditioning $g_{\leq\tau}$ on the event $\{\tau = t\}$, i.e., $g^t_{\leq t} = g_{\leq \tau} | \tau = t$. More precisely, recall that τ only uses the randomness of g_t. It is not an online random time. Hence, a practical way to simulate $g^t_{\leq t}$ for all $t \in Supp(\tau)$ is to first draw an i.i.d. sequence of learning rules $(g_{i,\leq\tau_i})_{i\geq 1}$. Then, for each $t \in Supp(\tau)$ we select the randomness which first satisfies $\tau = t$. Specifically, we define the time $i_t = \min\{i : \tau_i = t\}$ for all $t \in Supp(\tau)$. With probability one, these times are finite for all $t \in Supp(\tau)$. Denote this event $\mathcal{E}$. Then, letting $\bar y \in \mathcal{Y}$ be an arbitrary fixed value, for all $1 \leq t \leq T$ we pose

$$g^t_{\leq t} = \begin{cases} g_{i_t,\leq t} & \text{if } \mathcal{E} \text{ is met,} \\ \bar y_{\leq t} & \text{otherwise,}\end{cases} \quad t \in Supp(\tau) \quad \text{and} \quad g^t_{\leq t} = \bar y_{\leq t}, \quad t \notin Supp(\tau).$$

where $\bar y_{\leq t}$ denotes the learning rules which always outputs value $\bar y$ for all steps $u \leq t$. Intuitively, $g^t_{\leq t}$ has the same distribution as $g_{\leq\tau}$ conditioned on the event $\{\tau = t\}$. We are now ready to define a new learning rule $\tilde g_{\leq\tilde\tau}$, by $\tilde g_{\leq\tilde\tau} := g^{\tilde\tau}_{\leq\tilde\tau}$. Noting that for any $t \notin Supp(\tau)$ we have $\mathbb{P}[\tilde\tau = t] = 0$, we can write

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{\tau}\left(\ell(\tilde g_t(\boldsymbol{y}_{\leq t-1}), y_t) - \ell(y, y_t)\right) - \eta\tau\right] \\
&= \sum_{t=1}^{T}\mathbb{P}[\tilde\tau = t]\mathbb{E}\left[\sum_{u=1}^{t}\left(\ell(\tilde g_u(\boldsymbol{y}_{\leq u-1}), y_u) - \ell(y, y_u)\right) - \eta t \,\middle|\, \tilde\tau = t\right] \\
&= \sum_{t\in Supp(\tau)}\mathbb{P}[\tilde\tau = t]\mathbb{E}\left[\sum_{u=1}^{t}\left(\ell(\tilde g_u(\boldsymbol{y}_{\leq u-1}), y_u) - \ell(y, y_u)\right) - \eta t \,\middle|\, \tilde\tau = t, \mathcal{E}\right] \\
&= \frac{1}{\mathbb{E}[1/\tau]}\sum_{t\in Supp(\tau)}\mathbb{P}[\tau = t]\mathbb{E}\left[\frac{1}{t}\sum_{u=1}^{t}\left(\ell(g_{i_t,u}(\boldsymbol{y}_{\leq u-1}), y_u) - \ell(y, y_u)\right) - \eta \,\middle|\, \tilde\tau = t, \mathcal{E}\right] \\
&= \frac{1}{\mathbb{E}[1/\tau]}\sum_{t\in Supp(\tau)}\mathbb{P}[\tau = t]\mathbb{E}\left[\frac{1}{t}\sum_{u=1}^{t}\left(\ell(g_{i_t,u}(\boldsymbol{y}_{\leq u-1}), y_u) - \ell(y, y_u)\right) - \eta\right] \\
&= \frac{1}{\mathbb{E}[1/\tau]}\sum_{t\in Supp(\tau)}\mathbb{P}[\tau = t]\mathbb{E}\left[\frac{1}{t}\sum_{u=1}^{t}\left(\ell(g_u(\boldsymbol{y}_{\leq u-1}), y_u) - \ell(y, y_u)\right) - \eta \,\middle|\, \tau = t\right]
\end{aligned}$$

$$= \frac{1}{\mathbb{E}[1/\tau]}\mathbb{E}\left[\frac{1}{\tau}\sum_{t=1}^{\tau}\big(\ell(g_t(\boldsymbol{y}_{\leq t-1}), y_t) - \ell(y, y_t)\big) - \eta\right] \leq 0.$$

where in the second and fourth equality we used the fact that $\mathbb{P}[\mathcal{E}] = 1$. As a result, there exists a learning rule $\tilde{g}_{\leq \tilde{\tau}}$ such that $1 \leq \tilde{\tau} \leq T_\eta$, and for any $\boldsymbol{y}_{\leq T_\eta} \in \mathcal{Y}^{T_\eta}$ and $y \in \mathcal{Y}$ one has

$$\mathbb{E}\left[\sum_{t=1}^{\tilde{\tau}}\big(\ell(\tilde{g}_t(\boldsymbol{y}_{\leq t-1}), y_t) - \ell(y, y_t)\big) - \eta\tilde{\tau}\right] \leq 0.$$

We now pose $T'_\eta = \lceil T_\eta/\eta \rceil$ and draw an i.i.d. sequence of learning rules $(\tilde{g}^i_{\leq \tilde{\tau}_i})_{i\geq 1}$. Denote $\theta_i = \sum_{j<i}\tilde{\tau}_i$ with the convention $\theta_1 = 0$. We are now ready to define a learning rule $h_{\leq T'_\eta}$ as follows. For any $1 \leq t \leq T'_\eta$ and $\boldsymbol{y}_{\leq t} \in \mathcal{Y}^t$,

$$h_t(\boldsymbol{y}_{\leq t-1}) = \tilde{g}^i_{\leq t-\theta_i}((y_{t'})_{\theta_i < t' \leq t-1}), \qquad \theta_i < t \leq \theta_{i+1}, i \geq 1.$$

In other words, the learning rule performs independent learning rules $\tilde{g}_{\leq\tilde{\tau}}$ and when the time horizon $\tilde{\tau}$ is reached, we re-initialize the learning rule with a new randomness. Now let $\boldsymbol{y}_{\leq T'_\eta} \in \mathcal{Y}^{T'_\eta}$ and $y \in \mathcal{Y}$. We denote by $\hat{i} = \max\{i \geq 1, \theta_i \leq t\}$, the index of the last learning rule which had time to finish completely. Then, because $\tilde{\tau}_{\hat{i}} \leq T_\eta$,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T'_\eta}(\ell(h_t(\boldsymbol{y}_{\leq t-1}), y_t) - \ell(y, y_t)) - 2\eta T'_\eta\right] \\
&\leq \mathbb{E}\left[\sum_{i\leq \hat{i}}\sum_{t=1}^{\tilde{\tau}_i}(\ell(\tilde{g}^i_{t-\theta_i}(\boldsymbol{y}_{\theta_i < \cdot \leq t-1}), y_t) - \ell(y, y_t)) - \eta T'_\eta\right] - \eta T'_\eta + T_\eta \\
&\leq \mathbb{E}\left[\sum_{i\leq \hat{i}}\left(\sum_{t=1}^{\tilde{\tau}_i}(\ell(\tilde{g}^i_{t-\theta_i}(\boldsymbol{y}_{\theta_i < \cdot \leq t-1}), y_t) - \ell(y, y_t)) - \eta\tilde{\tau}_i\right)\right].
\end{aligned}$$

We now analyze the last term. First, note that by construction, the sequence

$$\left\{S_j := \sum_{j\leq i}\left(\sum_{t=1}^{\tilde{\tau}_j}(\ell(\tilde{g}^j_{t-\theta_j}(\boldsymbol{y}_{\theta_j < \cdot \leq t-1}), y_t) - \ell(y, y_t)) - \eta\tilde{\tau}_j\right)\right\}_{j\geq 1}$$

is a super-martingale. Now, note that $\hat{i} \leq 1 + T'_\eta$ since for all i, $\theta_i = \sum_{j<i}\tau_i \geq i-1$. As a result, $\hat{i}$ is bounded, is a stopping time for the considered filtration (after finishing period $\hat{i}$ we stop if and only we exceed time T'_η) and we can apply Doob's optimal sampling theorem to obtain $\mathbb{E}[S_{\hat{i}}] \leq 0$. Thus, combining the above equations gives

$$\frac{1}{T'_\eta}\mathbb{E}\left[\sum_{t=1}^{T'_\eta}(\ell(h_t(\boldsymbol{y}_{\leq t-1}), y_t) - \ell(y, y_t))\right] \leq 2\eta.$$

Because this holds for all $\eta > 0$, F-TiME is satisfied. This ends the proof of the lemma.

C.6. Proof of Theorem 5.8. We first prove that adversarial regression for processes outside of CS is not achievable. Precisely, we show that for any $\mathbb{X} \notin$ CS, for any online learning rule $f_\cdot$, there exists a process $\mathbb{Y}$ on $\mathcal{Y}$, a measurable function $f^* : \mathcal{X} \to \mathcal{Y}$ and $\delta > 0$ such that with non-zero probability $\mathcal{L}_{(\mathbb{X},\mathbb{Y})}(f_\cdot, f^*) > \delta$.

Because F-TiME is not satisfied by $(\mathcal{Y},\ell)$, by Lemma 5.7, Property 2 is not satisfied either. Hence, we can fix $\eta>0$ such that for any horizon $T\geq 1$ and any online learning rule $g_{\leq\tau}$ with $1\leq\tau\leq T$, there exist a sequence $\boldsymbol{y}:=(y_t)_{t=1}^T$ of values in $\mathcal{Y}$ and a value y such that

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

as in the assumption of the space $(\mathcal{Y},\ell)$. Let $\mathbb{X}\notin$ CS. The proof of Theorem 5.1 shows that there exist $0<\epsilon<1$, a sequence of disjoint measurable sets $\{B_p\}_{p\geq 1}$ and a sequence of times $(t_p)_{p\geq 0}$ with $t_0=0$ and such that with $\mu:=\max(1,\frac{8\bar{\ell}}{\epsilon\eta})$, for any $p\geq 1$, $t_p>\mu t_{p-1}$, and defining the events

$$\mathcal{E}_p=\left\{\mathbb{X}_{\leq t_{p-1}}\cap\left(\bigcup_{p'\geq p}B_p\right)=\emptyset\right\}\text{ and }\mathcal{F}_p:=\bigcup_{\mu t_{p-1}<t\leq t_p}\left\{\frac{1}{t}\sum_{u=1}^{t}\mathbb{1}_{B_p}(X_u)\geq\frac{\epsilon}{4}\right\},$$

we have $\mathbb{P}[\bigcap_{p\geq 1}(\mathcal{E}_p\cap\mathcal{F}_p)]\geq\frac{\epsilon}{4}$. We now fix a learning rule $f_\cdot$ and construct a “bad” process $\mathbb{Y}$ recursively. Fix $\bar{y}\in\mathcal{Y}$ an arbitrary value. We start by defining the random variables $N_p(t)=\sum_{u=t_{p-1}+1}^{t}\mathbb{1}_{B_p}(X_u)$ for any $p\geq 1$. We now construct (deterministic) values y_p and sequences $(y_p^u)_{u=1}^{t_p}$ for all $p\geq 1$, of values in $\mathcal{Y}$. Suppose we have already constructed the values y_q as well as the sequences $(y_q^u)_{u=1}^{t_q}$ for all $q<p$. We will now construct y_p and $(y_p^u)_{u=1}^{t_p}$. Assuming that the event $\mathcal{E}_p\cap\mathcal{F}_p$ is met, there exists $\mu t_{p-1}<t\leq t_p$ such that

$$N_p(t)=\sum_{u=t_{p-1}+1}^{t}\mathbb{1}_{B_p}(X_u)=\sum_{u=1}^{t}\mathbb{1}_{B_p}(X_u)\geq\frac{\epsilon}{4}t,$$

where in the first equality we used the fact that on $\mathcal{E}_p$, the process $\mathbb{X}_{\leq t_{p-1}}$ does not visit B_p. In the rest of the construction, we will denote

$$T_p=\begin{cases}\min\{\mu t_{p-1}<t\leq t_p:N_p(t)\geq\frac{\epsilon}{4}t\} & \text{if } \mathcal{E}_p\cap\mathcal{F}_p \text{ is met.}\\ t_p & \text{otherwise.}\end{cases}$$

Now consider the process $\mathbb{Y}_{t\leq t_{p-1}}(\mathbb{X})$ defined as follows. For any $1\leq q<p$ we pose

$$Y_t(\mathbb{X})=\begin{cases}y_q^{N_q(t)} & \text{if } t\leq T_q \text{ and } X_t\in B_q,\\ y_q & \text{if } t>T_q \text{ and } X_t\in B_q,\\ y_{q'} & \text{if } X_t\in B_{q'},\ q'<q,\\ \bar{y} & \text{otherwise,}\end{cases}\qquad t_{q-1}<t\leq t_q.$$

Similarly, for $M\geq 1$ and given any sequence $\{\tilde{y}_i\}_{i=1}^M$, we define the following process $\mathbb{Y}_{t_{p-1}<u\leq t_p}\left(\mathbb{X},\{\tilde{y}_i\}_{i=1}^M\right)$ by

$$Y_u\left(\mathbb{X},\{\tilde{y}_i\}_{i=q1}^M\right)=\begin{cases}\tilde{y}_{\min(N_p(u),M)} & \text{if } X_t\in B_p,\\ y_q & \text{if } X_t\in B_q,\ q<p,\\ \bar{y} & \text{otherwise.}\end{cases}$$

We now construct a learning rule $g^p_\cdot$. First, we define the event $\mathcal{B}:=\bigcap_{p\geq 1}(\mathcal{E}_p\cap\mathcal{F}_p)$. We will denote by $\tilde{\mathbb{X}}=\mathbb{X}|\mathcal{B}$ a sampling of the process $\mathbb{X}$ on the event $\mathcal{B}$ which has probability at least $\frac{\epsilon}{4}$. For instance we draw i.i.d. samplings following the same distribution as $\mathbb{X}$ then select the process which first falls into $\mathcal{B}$. We are now ready to define a learning rule $(g_u^p)_{u\leq\tau}$ where τ is a random time. To do so, we first draw a sample $\tilde{\mathbb{X}}$ which is now fixed for the learning rule

$g^p_\cdot$. We define the stopping time as $\tau = N_p(T_p)$. Finally, for all $1 \leq u \leq \tau$, and any sequence of values $\tilde{\boldsymbol{y}}_{\leq u-1}$, we pose

$$g^p_u(\tilde{\boldsymbol{y}}_{\leq u-1}) = f_{T_p(u)}\left(\tilde{\mathbb{X}}_{\leq T_p(u)-1}, \left\{\mathbb{Y}_{\leq t_{p-1}}(\tilde{\mathbb{X}}), \mathbb{Y}_{t_{p-1}<u\leq T_p(u)-1}\left(\tilde{\mathbb{X}}, \{\tilde{y}_i\}_{i=1}^{u-1}\right)\right\}, \tilde{X}_{T_p(u)}\right),$$

where we used the notation $T_p(u) := \min\{t_{p-1} < t' \leq t_p : N_p(t) = u\}$ for the time of the $u-$th visit of B_p, which exists because $u \leq \tau = N_p(T_p) \leq N_p(t_p)$ since the event $\mathcal{B}$ is satisfied by $\tilde{\mathbb{X}}$. Note that the prediction of the rule $g^p_\cdot$ is random because of the dependence on $\tilde{\mathbb{X}}$. Also, observe that the random time τ is bounded by $1 \leq \tau \leq T_p \leq t_p$. Therefore, by hypothesis on the value space $(\mathcal{Y}, \ell)$, there exists a sequence $\{y_p^u\}_{u=1}^{t_p}$ and a value $y_p \in \mathcal{Y}$ such that

$$\mathbb{E}\left[\frac{1}{\tau}\sum_{u=1}^{\tau}\left(\ell(g^p_u(\boldsymbol{y_p}^{\leq u-1}), y_p^u) - \ell(y_p, y_p^u)\right)\right] \geq \eta.$$

This ends the recursive construction of the values y_p and the sequences $(y_p^u)_{u=1}^{t_p}$ for all $p \geq 1$. We are now ready to define the process $\mathbb{Y}(\mathbb{X})$, using a similar construction as before. For any $p \geq 1$ we define

$$Y_t(\mathbb{X}) = \begin{cases} y_p^{N_p(t)} & \text{if } t \leq T_p \text{ and } X_t \in B_p, \\ y_p & \text{if } t > T_p \text{ and } X_t \in B_p, \\ y_q & \text{if } X_t \in B_q,\ q < p, \\ \bar{y} & \text{otherwise,} \end{cases} \qquad t_{p-1} < t \leq t_p.$$

We also define a function $f^* : \mathcal{X} \to \mathcal{Y}$ by

$$f^*(x) = \begin{cases} y_p & \text{if } x \in B_p, \\ \bar{y} & \text{otherwise.} \end{cases}$$

This function is simple hence measurable. From now, we will suppose that the event $\mathcal{B}$ is met. For simplicity, we will denote by $\hat{Y}_t := f_t(\mathbb{X}_{\leq t-1}, \mathbb{Y}_{\leq t-1}, X_t)$ the prediction of the learning rule at time t. For any $p \geq 1$, because $\mathcal{E}_p \cap \mathcal{F}_p$ is met, for all $1 \leq u \leq N_p(T_p)$, we have $t_{p-1} < T_p(u) \leq T_p$, and $X_{T_p(u)} \in B_p$. Hence, by construction, we have $\hat{Y}_{T_q(u)} = y_q^u$ and we can write

$$\begin{aligned}
\sum_{t=1}^{T_p} \ell(\hat{Y}_t, Y_t) &\geq \sum_{t=t_{p-1}+1}^{T_p} \ell(\hat{Y}_t, Y_t) \\
&\geq \sum_{u=1}^{N_p(T_p)} \ell(\hat{Y}_{T_p(u)}, Y_{T_p(u)}) \\
&= \sum_{u=1}^{\tau} \ell(f_{T_p(u)}\left(\mathbb{X}_{\leq T_p(u)-1}, \mathbb{Y}_{\leq T_p(u)-1}, X_{T_p(u)}\right), y_p^u).
\end{aligned}$$

Now note that because the construction was similar to the construction of $g^p_\cdot$, we have $\mathbb{Y}_{\leq T_p(u)-1} = \left\{\mathbb{Y}_{\leq t_{p-1}}(\mathbb{X}), \mathbb{Y}_{t_{p-1}<t\leq T_p(u)-1}\left(\mathbb{X}, \{y_p^i\}_{i=1}^{u-1}\right)\right\}$, i.e., $\hat{Y}_{T_p(u)}$ coincides with the prediction $g^p_u(\{y_p^i\}_{i=1}^{u-1})$ provided that g^p_u precisely used the realization $\mathbb{X}$. Hence, conditioned on $\mathcal{B}$ for all $u \leq \tau_p$, $\hat{Y}_{T_p(u)}$ has the same distribution as $g^p_u(\boldsymbol{y_p}^{\leq u-1})$. Therefore we obtain

$$\mathbb{E}\left[\frac{1}{\tau}\sum_{t=1}^{T_p}\ell(\hat{Y}_t, Y_t) - \frac{1}{\tau}\sum_{u=1}^{\tau}\ell(y_p, y_p^u)\middle| \mathcal{B}\right] \geq \mathbb{E}\left[\frac{1}{\tau}\sum_{u=1}^{\tau}\left(\ell(g^p_u(\hat{Y}_{T_p(u)}, y_p^u) - \ell(y_p, y_p^u)\right)\middle| \mathcal{B}\right]$$

$$= \mathbb{E}\left[\frac{1}{\tau}\sum_{u=1}^{\tau}\left(\ell(g_u^p(\boldsymbol{y_p}^{\leq u-1}), y_p^u) - \ell(y_p, y_p^u)\right)\right]$$
$$\geq \eta.$$

We now turn to the loss obtained by the simple function f^*. By construction, assuming that the event $\mathcal{B}$ is met, we have

$$\sum_{t=1}^{T_p} \ell(f^*(X_t), Y_t) \leq \bar{\ell} t_{p-1} + \sum_{u=1}^{N_p(T_p)} \ell(f^*(X_{T_p(u)}), y_p^u) = \bar{\ell} t_{p-1} + \sum_{u=1}^{\tau} \ell(y_p, y_p^u).$$

Recalling that $T_p > \mu t_{p-1} \geq \frac{8\bar{\ell}}{\epsilon\eta} t_{p-1}$ and noting that $\tau = N_p(T_p) \geq \frac{\epsilon}{4} T_p$, we obtain

$$\mathbb{E}\left[\sup_{t_{p-1} < T \leq t_p} \frac{1}{T}\sum_{t=1}^{T}(\ell(\hat{Y}_t, Y_t) - \ell(f(X_t), Y_t)) \middle| \mathcal{B}\right]$$
$$\geq \mathbb{E}\left[\frac{\tau}{T_p}\frac{1}{\tau}\left(\sum_{t=1}^{T}\ell(\hat{Y}_t, Y_t) - \sum_{u=1}^{\tau}\ell(y_p, y_p^u)\right) - \bar{\ell}\frac{t_{p-1}}{T_p} \middle| \mathcal{B}\right]$$
$$\geq \frac{\epsilon}{4}\mathbb{E}\left[\frac{1}{\tau}\sum_{t=1}^{T_p}\ell(\hat{Y}_t, Y_t) - \frac{1}{\tau}\sum_{u=1}^{\tau}\ell(y_p, y_p^u) \middle| \mathcal{B}\right] - \frac{\epsilon\eta}{8}$$
$$\geq \frac{\epsilon\eta}{8}.$$

Because this holds for any $p \geq 1$, Fatou lemma yields

$$\mathbb{E}\left[\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}\ell(\hat{Y}_t, Y_t) - \ell(f(X_t), Y_t)\right]$$
$$\geq \mathbb{E}\left[\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T}(\ell(\hat{Y}_t, Y_t) - \ell(f(X_t), Y_t)) \middle| \mathcal{B}\right]\mathbb{P}[\mathcal{B}]$$
$$\geq \frac{\epsilon^2\eta}{32}.$$

Hence, we do note have almost surely $\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} \ell(\hat{Y}_t, Y_t) - \ell(f(X_t), Y_t) \leq 0$. This shows that $\mathbb{X} \notin$ SOLAR, which in turn implies SOLAR $\subset$ CS. This ends the proof that SOLAR $\subset$ CS. The proof that CS $\subset$ SOLAR and the construction of an optimistically universal learning rule for adversarial regression is deferred to Section 7 where we give a stronger result which also holds for unbounded losses. Note that generalizing Theorem 5.2 to adversarial responses already shows that CS $\subset$ SOLAR and provides an optimistically universal learning rule when the loss ℓ is a metric $\alpha = 1$.

APPENDIX D: PROOFS OF SECTION 6

D.1. Proof of Theorem 3.6. We first show that there exists $t_1 \geq 1$ such that for any $t \geq t_1$, with high probability, for all $i \in I_t$,

$$\sum_{s=t_i}^{t} \ell(\hat{Y}_s, Y_s) \leq L_{t,i} + 3\ln^2 t\sqrt{t}.$$

For any $t \geq 0$, note that we have $\hat{\ell}_t = \mathbb{E}[\ell(\hat{Y}_t, Y_t) \mid \mathbb{Y}_{\leq t}]$. We define the instantaneous regret $r_{t,i} = \hat{\ell}_t - \ell(y^i, Y_t)$. We now define $w'_{t-1,i} := e^{\eta_{t-1}(\hat{L}_{t-1,i} - L_{t-1,i})}$ and pose $W_{t-1} = \sum_{i \in I_t} w_{t-1,i}$ and $W'_{t-1} = \sum_{i \in I_{t-1}} w'_{t-1,i}$, i.e., which induces the most regret. We also denote the index $k_t \in I_t$ such that $\hat{L}_{t,k_t} - L_{t,k_t} = \max_{i \in I_t} \hat{L}_{t,i} - L_{t,i}$. We first note that for any $i, j \in I_t$, we have $\ell(y^i, Y_t) - \ell(y^j, Y_t) \leq \ell(y^i, y^0) + \ell(y^0, y^j) \leq 2 \ln t$. Therefore, we also have $|r_{t,i}| \leq 2 \ln t$. Hence, we can apply Hoeffding's lemma to obtain

$$\frac{1}{\eta_t} \ln \frac{W'_t}{W_{t-1}} = \frac{1}{\eta_t} \ln \sum_{i \in I_t} \frac{w_{t-1,i}}{W_{t-1}} e^{\eta_t r_{t,i}} \leq \frac{1}{\eta_t} \left(\eta_t \sum_{i \in I_t} r_{t,i} \frac{w_{t-1,i}}{W_{t-1}} + \frac{\eta_t^2 (4 \ln t)^2}{8} \right) = 2\eta_t \ln^2 t.$$

The same computations as in the proof of Lemma 4.2 then show that

$$\begin{aligned} (4) \quad \frac{1}{\eta_t} \ln \frac{w_{t-1,k_{t-1}}}{W_{t-1}} - \frac{1}{\eta_{t+1}} \ln \frac{w_{t,k_t}}{W_t} \leq 2 \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \ln(1 + \ln(t+1)) + \frac{|I_{t+1}| - |I_t|}{\eta_t \sum_{i \in I_t} w_{t,i}} \\ + (\hat{L}_{t-1,k_{t-1}} - L_{t-1,k_{t-1}}) - (\hat{L}_{t,k_t} - L_{t,k_t}) + 2\eta_t \ln^2 t. \end{aligned}$$

First suppose that we have $\sum_{i \in I_t} w_{t,i} \leq 1$. Similarly to Lemma 4.2, we obtain $\hat{L}_{t,k_t} - L_{t,k_t} \leq 0$. Otherwise, let $t' = \min\{1 \leq s \leq t : \forall s \leq s' \leq t, \sum_{i \in I_{s'}} w_{s',i} \geq 1\}$. We sum equation (4) for $s = t', \ldots, t$ which gives

$$\begin{aligned} \frac{1}{\eta_1} \ln \frac{w_{t'-1,k_{t'-1}}}{W_{t'-1}} - \frac{1}{\eta_{t+1}} \ln \frac{w_{t,k_t}}{W_t} \leq \frac{2}{\eta_{t+1}} \ln(1 + \ln(t+1)) + \frac{|I_{t+1}|}{\eta_t} \\ + (\hat{L}_{t'-1,k_{t'-1}} - L_{t'-1,k_{t'-1}}) - (\hat{L}_{t,k_t} - L_{t,k_t}) + 2 \sum_{s=t'}^{t} \eta_s \ln^2 s. \end{aligned}$$

Similarly as in Lemma 4.2, we have $\frac{w_{t,k_t}}{W_t} \leq 1$, $\frac{w_{t'-1,k_{t'-1}}}{W_{t'-1}} \geq \frac{1}{1+\ln t}$ and $\hat{L}_{t'-1,k_{t'-1}} - L_{t'-1,k_{t'-1}} \leq 0$. Finally, using the fact that $\sum_{s=1}^{t} \frac{1}{\sqrt{s}} \leq 2\sqrt{t}$, we obtain

$$\hat{L}_{t,k_t} - L_{t,k_t} \leq \ln(1 + \ln(t+1))(4 + 8\sqrt{t+1}) + 4(1 + \ln(t+1))\sqrt{t} + \ln^2 t \sqrt{t} \leq 2 \ln^2 t \sqrt{t},$$

for all $t \geq t_0$ where t_0 is a fixed constant, and as a result, for all $t \geq t_0$ and $i \in I_t$, we have $\hat{L}_{t,i} - L_{t,i} \leq 2 \ln^2 t \sqrt{t}$.

Now note that $|\ell(\hat{Y}_t, Y_t) - \mathbb{E}[\ell(\hat{Y}_t, Y_t) \mid \mathbb{Y}_{\leq t}]| \leq 2 \ln t$ because for all $i \in I_t$, we have $\ell(y^i, y^0) \leq \ln t$. Hence, we can apply Hoeffding-Azuma inequality to the variables $\ell(\hat{Y}_t, Y_t) - \hat{\ell}_t$ that form a sequence of differences of a martingale, which yields

$$\mathbb{P}\left[\sum_{s=t_i}^{t} \ell(\hat{Y}_s, Y_s) > \hat{L}_{t,i} + u \right] \leq e^{-\frac{u^2}{8t \ln^2 t}}.$$

Hence, for $t \geq t_0$ and $i \in I_t$, with probability $1 - \delta$, we have

$$\sum_{s=t_i}^{t} \ell(\hat{Y}_s, Y_s) \leq \hat{L}_{t,i} + \ln t \sqrt{2t \ln \frac{1}{\delta}} \leq L_{t,i} + 2 \ln^2 t \sqrt{t} + \ln t \sqrt{2t \ln \frac{1}{\delta}}.$$

Therefore, since $|I_t| \leq 1 + \ln t$, by union bound with probability $1 - \frac{1}{t^2}$ we obtain that for all $i \in I_t$,

$$\sum_{s=t_i}^{t} \ell(\hat{Y}_s, Y_s) \leq L_{t,i} + 2 \ln^2 t \sqrt{t} + \ln t \sqrt{2t \ln(1 + \ln t)} + \ln t \sqrt{4t \ln t} \leq 3 \ln^2 t \sqrt{t}$$

for all $t \geq t_1$ where $t_1 \geq t_0$ is a fixed constant. Now because $\sum_{t\geq 1} \frac{1}{t^2} < \infty$, the Borel-Cantelli lemma implies that almost surely, there exists $\hat{t} \geq 0$ such that

$$\forall t \geq \hat{t}, \forall i \in I_t, \quad \sum_{s=t_i}^{t} \ell(\hat{Y}_s, Y_s) \leq L_{t,i} + 3\ln^2 t\sqrt{t}.$$

We denote by $\mathcal{A}$ this event. Now let $y \in \mathcal{Y}$, $\epsilon > 0$ and consider $i \geq 0$ such that $\ell(y^i, y) < \epsilon$. On the event $\mathcal{A}$, we have for all $t \geq \max(\hat{t}, t_i)$,

$$\sum_{s=t_i}^{t} \ell(\hat{Y}_s, Y_s) \leq \sum_{s=t_i}^{t} \ell(y^i, Y_s) + 3\ln^2 t\sqrt{t} \leq \sum_{s=t_i}^{t} \ell(y, Y_s) + \epsilon t + 3\ln^2 t\sqrt{t}.$$

Therefore, $\limsup_{t\to\infty} \frac{1}{t}\sum_{s=1}^{t} \left(\ell(\hat{Y}_s, Y_s) - \ell(y, Y_s)\right) \leq \epsilon$ on $\mathcal{A}$. Because this holds for any $\epsilon > 0$ we finally obtain $\limsup_{t\to\infty} \frac{1}{t}\sum_{s=1}^{t} \left(\ell(\hat{Y}_s, Y_s) - \ell(y, Y_s)\right) \leq 0$ on the event $\mathcal{A}$ of probability one, which holds for all $y \in \mathcal{Y}$ simultaneously. This ends the proof of the theorem.

D.2. Proof of Corollary 6.2. We denote by $g_\cdot$ the learning rule on values $\mathcal{Y}$ for mean estimation described in Theorem 3.6. Because processes in $\mathbb{X} \in \text{FS}$ visit only finite number of different instance points in $\mathcal{X}$ almost surely, we can simply perform the learning rule $g_\cdot$ on each sub-process $\mathbb{Y}_{\{t: X_t = x\}}$ separately for any $x \in \mathcal{X}$. Note that the learning rule $g_\cdot$ does not explicitely re-use past randomness for its prediction. Hence, we will not specify that the randomness used for all learning rules—for each x visited by $\mathbb{X}$—should be independent. Let us formally describe our learning rule. Consider a sequence $\boldsymbol{x}_{\leq t-1}$ of instances in $\mathcal{X}$ and $\boldsymbol{y}_{\leq t-1}$ of values in $\mathcal{Y}$. We denote by $S_{t-1} = \{x : \boldsymbol{x}_{\leq t-1} \cap \{x\} \neq \emptyset\}$ the support of $\boldsymbol{x}_{\leq t-1}$. Further, for any $x \in S_{t-1}$, we denote $N_{t-1}(x) = \sum_{u \leq t-1} \mathbb{1}_{x_u = x}$ the number of times that the specific instance x was visited by the sequence $\boldsymbol{x}_{\leq t-1}$. Last, for any $x \in S_{t-1}$, we denote $\boldsymbol{y}^x_{\leq N(x)}$ the values $\boldsymbol{y}_{\{u \leq t: X_u = x\}}$ obtained when the instance was precisely x in the sequence $\boldsymbol{x}_{\leq t-1}$, ordered by increasing time u. We are now ready to define our learning rule f_t at time t. Given a new instance point x_t, we pose

$$f_t(\boldsymbol{x}_{\leq t-1}, \boldsymbol{y}_{\leq t-1}, x_t) = \begin{cases} g_{N_{t-1}(x)+1}(\boldsymbol{y}^x_{\leq N_{t-1}(x)}) & \text{if } x_t \in S_{t-1}, \\ g_1(\emptyset) & \text{otherwise.} \end{cases}$$

Recall that for any $u \geq 1$, g_u uses some randomness. The only subtlety is that at each iteration $t \geq 1$ of the learning rule $f_\cdot$, the randomness used by the subroutine call to $g_\cdot$ should be independent from the past history. We now show that $f_\cdot$ is universally consistent for adversarial regression under all processes $\mathbb{X} \in \text{FS}$.

Let $\mathbb{X} \in \text{FS}$. For simplicity, we will denote by $\hat{Y}_t$ the prediction of the learning rule $f_\cdot$ at time t. We denote $S = \{x : \{x\} \cap \mathbb{X} \neq \emptyset\}$ the random support of $\mathbb{X}$. By hypothesis, we have $|S| < \infty$ with probability one. Denote by $\mathcal{E}$ this event. We now consider a specific realization $\boldsymbol{x}$ of $\mathbb{X}$ falling in the event $\mathcal{E}$. Then, S is a fixed set. We also denote $\tilde{S} := \{x \in S : \lim_{t\to\infty} N_t(x) = \infty\}$ the instances which are visited an infinite number of times by the sequence $\boldsymbol{x}$. Now, we can write for any function $f : \mathcal{X} \to \mathcal{Y}$,

$$\sum_{t=1}^{T} \left(\ell(\hat{Y}_t, Y_t) - \ell(f(x_t), Y_t)\right) = \sum_{x\in S} \sum_{u=1}^{N_t(x)} \left(\ell(g_u(\mathbb{Y}^x_{\leq u-1}), Y^x_u) - \ell(f(x), Y_u)\right)$$

$$\leq \sum_{s\in S\setminus\tilde{S}} \bar{\ell}|\{t \geq 1 : x_t = x\}| + \sum_{s\in\tilde{S}} \sum_{u=1}^{N_t(x)} \left(\ell(g_u(\mathbb{Y}^x_{\leq u-1}), Y^x_u) - \ell(f(x), Y_u)\right).$$

Now, because the randomness in $g.$ was taken independently from the past at each iteration, we can apply directly Theorem 3.6. For all $x \in \tilde{S}$, with probability one, for all $y^x \in \mathcal{Y}$,

$$\limsup_{t'\to\infty} \frac{1}{t'} \sum_{u=1}^{t'} \left(\ell(g_u(\mathbb{Y}^x_{\leq u-1}), Y^x_u) - \ell(y^x, Y_u)\right) \leq 0.$$

We denote by $\mathcal{E}_x$ this event. Then, on the event $\bigcap_{x\in\tilde{S}} \mathcal{E}_x$ of probability one, we have for any measurable function $f : \mathcal{X} \to \mathcal{Y}$,

$$\begin{aligned}
&\limsup_{T\to\infty} \frac{1}{T} \left(\ell(\hat{Y}_t, Y_t) - \ell(f(x_t), Y_t)\right) \\
&\qquad\leq \sum_{s\in\tilde{S}} \limsup_{T\to\infty} \frac{1}{T} \sum_{u=1}^{N_t(x)} \left(\ell(g_u(\mathbb{Y}^x_{\leq u-1}), Y^x_u) - \ell(f(x), Y_u)\right) \\
&\qquad\leq \sum_{s\in\tilde{S}} \limsup_{T\to\infty} \frac{1}{N_t(x)} \sum_{u=1}^{N_t(x)} \left(\ell(g_u(\mathbb{Y}^x_{\leq u-1}), Y^x_u) - \ell(f(x), Y_u)\right) \leq 0.
\end{aligned}$$

As a result, averaging on realisations of $\mathbb{X}$, we obtain that with probability one, we have that $\mathcal{L}_{(\mathbb{X},\mathbb{Y})}(f., f) \leq 0$ for all measurable functions $f : \mathcal{X} \to \mathcal{Y}$. Note that this is stronger than the notion of universal consistency which we defined in Section 2, where we ask that for all measurable function $f : \mathcal{X} \to \mathcal{Y}$, we have almost surely $\mathcal{L}_{(\mathbb{X},\mathbb{Y})}(f., f) \leq 0$. In particular, this shows that FS $\subset$ SOLAR-U. As result SOLAR-U $=$ FS and $f.$ is optimistically universal. This ends the proof of the result.

D.3. Proof of Theorem 6.3. We first show that mean-estimation is not achievable. To do so, let $f.$ be a learning rule. For simplicity, we will denote by $\hat{Y}_t$ its prediction at step t. We aim to construct a process $\mathbb{Y}$ on $\mathbb{R}$ and a value $y^* \in \mathbb{R}$ such that with non-zero probability we have

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

We now pose $\beta := \frac{2\alpha}{\alpha-1} > 2$. For any sequence $\boldsymbol{b} := (b_t)_{t\geq 1}$ in $\{-1, 1\}$, we consider the following process $\mathbb{Y}^{\boldsymbol{b}}$ such that for any $t \geq 1$ we have $Y^{\boldsymbol{b}}_t = 2^{\beta^t} b_t$. Let $\boldsymbol{B} := (B_t)_{t\geq 1}$ be an i.i.d. sequence of Rademacher random variables, i.e., such that $B_1 = 1$ (resp. $B_1 = -1$) with probability $\frac{1}{2}$. We consider the random variables $e_t := \mathbb{1}_{\hat{Y}_t \cdot Y_t \leq 0}$ which intuitively correspond to flags for large mistakes of the learning rule $f.$ at time t. Because $f.$ is an online learning rule, we have

$$\mathbb{E}[e_t \mid \mathbb{Y}_{\leq t-1}] = \mathbb{E}_{\hat{Y}_t}\left[\mathbb{E}_{B_t}[\mathbb{1}_{\hat{Y}_t \cdot Y_t \leq 0} \mid \hat{Y}_t]\right] = \mathbb{E}_{\hat{Y}_t}\left[\mathbb{1}_{\hat{Y}_t = 0} + \frac{1}{2}\mathbb{1}_{\hat{Y}_t \neq 0}\right] \geq \frac{1}{2}.$$

where the expectation $\mathbb{E}_{\hat{Y}_t}$ refers to the expectation on the randomness of the rule f_t. As a result, the random variables $e_t - \frac{1}{2}$ form a sequence of differences of a sub-martingale bounded by $\frac{1}{2}$ in absolute value. By the Azuma-Hoeffding inequality, we obtain $\mathbb{P}\left[\sum_{t=1}^T e_t \leq \frac{T}{4}\right] \leq e^{-T/8}$. Because $\sum_{t\geq 1} e^{-t/8} < \infty$, the Borel-Cantelli lemma implies that on an event $\mathcal{E}$ of probability one, we have $\limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^T e_t \geq \frac{1}{4}$. As a result, there exists a specific realization $\boldsymbol{b}$ of $\boldsymbol{B}$ such that on an event $\tilde{\mathcal{E}}$ of probability one, we have $\limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^T e_t \geq$

$\frac{1}{4}$. Note that the sequence $\mathbb{Y}^{\boldsymbol{b}}$ is now deterministic. Then, writing $e_t = e_t \mathbb{1}_{Y_t>0} + e_t \mathbb{1}_{Y_t<0}$, we obtain

$$\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^T e_t \mathbb{1}_{Y_t>0} + \limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^T e_t \mathbb{1}_{Y_t<0} \geq \frac{1}{4}.$$

Without loss of generality, we can suppose that $\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^T \mathbb{1}_{\hat{Y}_t\cdot Y_t\leq 0}\mathbb{1}_{Y_t>0} \geq \frac{1}{8}$. We now pose $y^*=1$. In the other case, we pose $y^*=-1$. We now compute for any $T\geq 1$ such that $\hat{Y}_t\cdot Y_t\leq 0$ and $Y_t>0$,

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^T (\ell(f_t(\mathbb{Y}_{\leq t-1}),Y_t) - \ell(y^*,Y_t)) &\geq \frac{\ell(0,2^{\beta^T}) - \ell(1,2^{\beta^T})}{T} - \frac{1}{T}\sum_{t=1}^{T-1}\ell(1,-2^{\beta^t}).\\ &= \frac{\alpha}{T}2^{(\alpha-1)\beta^T} + O\left(\frac{1}{T}2^{(\alpha-2)\beta^T}\right) - 2^{\alpha(1+\beta^{T-1})}\\ &= \frac{\alpha}{T}2^{2\alpha\beta^{T-1}}(1+o(1)).\end{aligned}$$

Because, by construction $\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^T \mathbb{1}_{\hat{Y}_t\cdot Y_t\leq 0}\mathbb{1}_{Y_t>0} \geq \frac{1}{8}$, we obtain

$$\limsup \frac{1}{T}\sum_{t=1}^T (\ell(f_t(\mathbb{Y}_{\leq t-1}),Y_t) - \ell(y^*,Y_t)) = \infty,$$

on the event $\tilde{E}$ of probability one. This end the proof that mean-estimation is not achievable. Because mean-estimation is the easiest regression setting, this directly implies SOLAR-U $=\emptyset$. Formally, let $\mathbb{X}$ a process on $\mathcal{X}$. and $f.$ a learning rule for regression. We consider the same processes $\mathbb{Y}^{\boldsymbol{B}}$ where $\boldsymbol{B}$ is i.i.d. Rademacher and independent from $\mathbb{X}$. The same proof shows that there exists a realization $\boldsymbol{b}$ for which we have almost surely $\mathcal{L}_{(\mathbb{X},\mathbb{Y})}(f., f^*:=y^*) = \infty$, where $f^*=y^*$ denotes the constant function equal to y^* where $y^*\in\mathbb{R}$ is the value constructed as above. Hence, $\mathbb{X}\notin$ SOLAR-U, and as a result, SOLAR-U $=\emptyset$.

D.4. Proof of Proposition 6.4. Suppose that there exists an online learning rule $g.$ for mean-estimation. In the proof of Corollary 6.2, instead of using the learning rule for mean-estimation for metric losses introduced in Theorem 3.6, we can use the learning rule $g.$ to construct the learning rule $f.$ for adversarial regression on FS instance processes, which simply performs $f.$ separately on each subprocess $\mathbb{Y}_{t:X_t=x}$ with the same instance $x\in\mathcal{X}$ for all visited $x\in\mathcal{X}$ in the process $\mathbb{X}$. The same proof shows that because almost surely $\mathbb{X}$ visits a finite number of different instances, $f.$ is universally consistent under any process $\mathbb{X}\in$ FS. Hence, FS $\subset$ SOLAR-U. Because SOLAR-U $\subset$ SOUL $=$ FS, we obtain directly SOLAR-U $=$ FS and $f.$ is optimistically universal.

On the other hand, if mean-estimation with adversarial responses is not achievable, we can use similar arguments as for the proof of Theorem 6.3. Let $f.$ a learning rule for regression, and consider the following learning rule $g.$ for mean-estimation. We first draw a process $\tilde{\mathbb{X}}$ with same distribution as $\mathbb{X}$. Then, we pose

$$g_t(\boldsymbol{y}_{\leq t-1}) := f_t(\tilde{\mathbb{X}}_{\leq t-1}, \boldsymbol{y}_{\leq t-1}, \tilde{X}_t).$$

Then, because mean-estimation is not achievable, there exists an adversarial process $\mathbb{Y}$ on $(\mathcal{Y},\ell)$ such that with non-zero probability,

$$\limsup \frac{1}{T}\sum_{t=1}^T (\ell(g_t(\mathbb{Y}_{\leq t-1}),Y_t) - \ell(y^*,Y_t)) > 0.$$

Then, we obtain that with non-zero probability, $\mathcal{L}_{(\tilde{\mathbb{X}},\mathbb{Y})} > 0$. Hence, $f.$ is not universally consistent. Note that the “bad” process $\mathbb{Y}$ is not correlated with $\tilde{\mathbb{X}}$ in this construction.

APPENDIX E: PROOFS OF SECTION 7

E.1. Proof of Theorem 7.1. Let $(x^k)_{k\geq 0}$ a sequence of distinct points of $\mathcal{X}$. Now fix a value $y_0 \in \mathcal{Y}$ and construct a sequence of values y_k^1, y_k^2 for $k \geq 1$ such that $\ell(y_k^1, y_k^2) \geq c_\ell 2^{k+1}$. Because $\ell(y_k^1, y_k^2) \leq c_\ell \ell(y_0, y_k^1) + c_\ell \ell(y_0, y_k^2)$, there exists $i_k \in \{1,2\}$ such that $\ell(y_0, y_k^{i_k}) \geq 2^k$. For simplicity, we will now write $y_k := y_k^{i_k}$ for all $k \geq 1$. We define

$$t_k = \left\lfloor \sum_{l=1}^{k} \ell(y_0, y_l) \right\rfloor .$$

This forms an increasing sequence of times because $t_{k+1} - t_k \geq \ell(y_0, y_{k+1}) \geq 1$. Consider the deterministic process $\mathbb{X}$ that visits x^k at time t_k and x^0 otherwise, i.e., such that

$$X_t = \begin{cases} x^k & \text{if } t = t_k, \\ x^0 & \text{otherwise.} \end{cases}$$

The process $\mathbb{X}$ visits $\mathcal{X} \setminus \{x^0\}$ a sublinear number of times. Hence we have for any measurable set A:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}_A(X_t) = \begin{cases} 1 & \text{if } x^0 \in A \\ 0 & \text{otherwise.} \end{cases}$$

As a result, $\mathbb{X} \in$ CRF. We will now show that universal learning under $\mathbb{X}$ with the first moment condition on the responses is not achievable. For any sequence $b := (b_k)_{k\geq 1}$ of binary variables $b_k \in \{0,1\}$, we define the function $f_b^* : \mathcal{X} \to \mathcal{Y}$ such that

$$f_b^*(x^k) = \begin{cases} y_0 & \text{if } b_k = 0, \\ y_k & \text{otherwise,} \end{cases} \quad k \geq 0 \quad \text{and} \quad f_b^*(x) = y_0 \text{ if } x \notin \{x_k, k \geq 0\}.$$

These functions are simple, hence measurable. We will first show that for any binary sequence b, the function f_b^* satisfies the moment condition on the target functions. Indeed, we note that for any $T \geq t_1$, with $k := \max\{l \geq 1 : t_l \leq T\}$, we have

$$\frac{1}{T} \sum_{t=1}^{T} \ell(y_0, f_b^*(X_t)) \leq \frac{1}{T} \sum_{l=1}^{k} \ell(y_0, y_k) \leq \frac{t_k + 1}{T} \leq \frac{T+1}{T}.$$

Therefore, $\limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \ell(y_0, f_b^*(X_t)) \leq 1$. We now consider any online learning rule $f_\cdot$. Let $B = (B_k)_{k\geq 1}$ be an i.i.d. sequence of Bernouilli variables independent from the learning rule randomness. For any $k \geq 1$, denoting by $\hat{Y}_{t_k} := f_{t_k}(\mathbb{X}_{\leq t_k - 1}, f_B^*(\mathbb{X}_{\leq t_k - 1}), X_{t_k})$ we have

$$\mathbb{E}_{B_k} \ell(\hat{Y}_{t_k}, f_B^*(X_{t_k})) = \frac{\ell(\hat{Y}_{t_k}, y_0) + \ell(\hat{Y}_{t_k}, y_k)}{2} \geq \frac{1}{2c_\ell} \ell(y_0, y_k).$$

In particular, taking the expectation over both B and the learning rule, we obtain

$$\mathbb{E}\left[\frac{1}{t_k} \sum_{t=1}^{t_k} \ell(f_t(\mathbb{X}_{\leq t-1}, f_B^*(\mathbb{X}_{\leq t-1}), X_t), f_B^*(X_t))\right] \geq \frac{1}{2c_\ell t_k} \sum_{l=1}^{k} \ell(y_0, y_k) \geq \frac{1}{2c_\ell}.$$

As a result, using Fatou's lemma we obtain

$$\mathbb{E}\left[\limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \ell(f_t(\mathbb{X}_{\leq t-1}, f_B^*(\mathbb{X}_{\leq t-1}), X_t), f_B^*(X_t))\right]$$

$$
\begin{aligned}
&\geq \limsup_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \ell(f_t(\mathbb{X}_{\leq t-1}, f_B^*(\mathbb{X}_{\leq t-1}), X_t), f_B^*(X_t))\right] \\
&\geq \frac{1}{2c_\ell}.
\end{aligned}
$$

Therefore, the learning rule $f_\cdot$ is not consistent under $\mathbb{X}$ for all target functions of the form f_b^* for some sequence of binary variables b. Indeed, otherwise for all binary sequence $b = (b_k)_{k\geq 1}$, we would have $\mathbb{E}_{\mathbb{X}}\left[\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^T \ell(f_t(\mathbb{X}_{\leq t-1}, f_b^*(\mathbb{X}_{\leq t-1}), X_t), f_b^*(X_t))\right] = 0$ and as a result

$$
\mathbb{E}_B\mathbb{E}_{\mathbb{X}}\left[\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^T \ell(f_t(\mathbb{X}_{\leq t-1}, f_B^*(\mathbb{X}_{\leq t-1}), X_t), f_B^*(X_t))\right] = 0.
$$

This ends the proof of the theorem.

E.2. Proof of Lemma 7.3. It suffices to prove that empirical integrability implies the latter property. We pose $\epsilon_i = 2^{-i}$ for any $i \geq 0$. By definition, there exists an event $\mathcal{E}_i$ of probability one such that on $\mathcal{E}_i$ we have

$$
\exists M_i \geq 0, \quad \limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^T \ell(y_0, Y_t)\mathbb{1}_{\ell(y_0,Y_t)\geq M_i} \leq \epsilon_i.
$$

As a result, on $\bigcap_{i\geq 0}\mathcal{E}_i$ of probability one, we obtain

$$
\forall \epsilon > 0, \exists M := M_{\lceil \log_2 \frac{1}{\epsilon}\rceil} \geq 0, \quad \limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^T \ell(y_0, Y_t)\mathbb{1}_{\ell(y_0,Y_t)\geq M} \leq \epsilon.
$$

This ends the proof of the lemma.

E.3. Proof of Theorem 3.1. Let $\mathbb{X} \in$ SOUL and $f^* : \mathcal{X} \to \mathcal{Y}$ such that $f^*(\mathbb{X})$ is empirically integrable. By Lemma 7.3, there exists some value $y_0 \in \mathcal{Y}$ such that on an event $\mathcal{A}$ of probability one, for all $\epsilon > 0$ there exists $M_\epsilon \geq 0$ such that

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

For any $M \geq 1$ we define the function f_M^* by

$$
f_M^*(x) = \begin{cases} f^*(x) & \text{if } \ell(y_0, f^*(x)) \leq M, \\ y_0 & \text{otherwise.} \end{cases}
$$

We know that 2C1NN is optimistically universal in the noiseless setting for bounded losses. Therefore, restricting the study to the output space $(B_\ell(y_0, M), \ell)$ we obtain that 2C1NN is consistent for f_M^* under $\mathbb{X}$, i.e.

$$
\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^T \ell(2C1NN_t(\mathbb{X}_{t-1}, f_M^*(\mathbb{X}_{\leq t-1}), X_t), f_M^*(X_t)) = 0 \quad (a.s.).
$$

For any $t \geq 1$, we denote $\phi(t)$ the representative used by the 2C1NN learning rule. We denote $\mathcal{E}_M$ the above event such that $\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^T \ell(f_M^*(X_{\phi(t)}), f_M^*(X_t)) = 0$. We now write for any $T \geq 1$ and $M \geq 1$,

$$\frac{1}{T}\sum_{t=1}^{T}\ell(f^*(X_{\phi(t)}), f^*(X_t)) \leq \frac{c_\ell^2}{T}\sum_{t=1}^{T}\ell(f^*_M(X_{\phi(t)}), f^*_M(X_t)) + \frac{c_\ell^2}{T}\sum_{t=1}^{T}\ell(f^*(X_t), f^*_M(X_t)) + \frac{c_\ell}{T}\sum_{t=1}^{T}\ell(f^*(X_{\phi(t)}), f^*_M(X_{\phi(t)})).$$

We now note that by construction of the 2C1NN learning rule,

$$\frac{1}{T}\sum_{t=1}^{T}\ell(f^*(X_{\phi(t)}), f^*_M(X_{\phi(t)})) = \frac{1}{T}\sum_{u=1}^{T}\ell(f^*(X_u), f^*_M(X_u))|\{u < t \leq T : \phi(t) = u\}| \leq \frac{2}{T}\sum_{t=1}^{T}\ell(f^*(X_t), f^*_M(X_t)).$$

Hence, we obtain

$$\frac{1}{T}\sum_{t=1}^{T}\ell(f^*(X_{\phi(t)}), f^*(X_t)) \leq \frac{c_\ell^2}{T}\sum_{t=1}^{T}\ell(f^*_M(X_{\phi(t)}), f^*_M(X_t)) + \frac{c_\ell(2+c_\ell)}{T}\sum_{t=1}^{T}\ell(y_0, f^*(X_t))\mathbb{1}_{\ell(y_0, f^*(X_t))>M}.$$

As a result, on the event $\mathcal{A} \cap \bigcap_{M\geq 1}\mathcal{E}_M$ of probability one, for any $M \geq 1$, we obtain

$$\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\ell(f^*(X_{\phi(t)}), f^*(X_t)) \leq c_\ell(2+c_\ell)\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\ell(y_0, f^*(X_t))\mathbb{1}_{\ell(y_0, f^*(X_t))\geq M}.$$

In particular, if $\epsilon > 0$ we can apply this result with $M := \lceil M_\epsilon \rceil$, which shows that $\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\ell(f^*(X_{\phi(t)}), f^*(X_t)) \leq c_\ell(2+c_\ell)\epsilon$. Because this holds for any $\epsilon > 0$ we finally obtain that on the event $\mathcal{A} \cap \bigcap_{M\geq 1}\mathcal{E}_M$ we have

$$\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\ell(f^*(X_\phi(t)), f^*(X_t)) = 0.$$

This ends the proof of the theorem.

E.4. Proof of Theorem 3.3. We first define the learning rule. Using Lemma 23 of [5], let $\mathcal{T} \subset \mathcal{B}$ a countable set such that for all $\mathbb{X} \in \text{CS}, A \subset \mathcal{B}$ we have

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

Now let $(y^i)_{i\geq 0}$ be a dense sequence in $\mathcal{Y}$. For any $k \geq 0$, any indices $l_1, \ldots, l_k \in \mathbb{N}$ and any sets $A_1, \ldots, A_k \in \mathcal{T}$, we define the function $f_{\{l_1,\ldots,l_k\},\{A_1,\ldots,A_k\}} : \mathcal{X} \to \mathcal{Y}$ as

$$f_{\{l_1,\ldots,l_k\},\{A_1,\ldots,A_k\}}(x) = y^{\max\{0\leq j\leq k:\, x\in A_j\}}$$

where $A_0 = \mathcal{X}$. These functions are simple hence measurable. Because the set of such functions is countable, we enumerate these functions as $f^0, f^1 \ldots$ Without loss of generality, we suppose that $f^0 = y^0$. For any $i \geq 0$, we denote $k^i \geq 0$, $\{l_1^i, \ldots, l_{k^i}^i\}$ and $\{A_1^i, \ldots, A_{k^i}^i\}$ such that f^i was defined as $f^i := f_{\{l_1^i,\ldots,l_k^i\},\{A_1^i,\ldots,A_k^i\}}$. We now define a sequence of sets $(I_t)_{t\geq 1}$ of indices and a sequence of sets $(\mathcal{F}_t)_{t\geq 1}$ of measurable functions by

$$I_t := \{i \leq \ln t : \ell(y^{l_p^i}, y^0) \leq 2^{-\alpha+1}\ln t,\ \forall 1 \leq p \leq k^i\} \quad \text{and} \quad \mathcal{F}_t := \{f^i : i \in I_t\}.$$

Then, clearly I_t is finite and $\bigcup_{t\geq 1} I_t = \mathbb{N}$. For any $i \geq 0$, we define $t_i = \min\{t : i \in I_t\}$. We are now ready to construct our learning rule. Let $\eta_t = \frac{1}{\ln t\sqrt{t}}$. Fix any sequences $(x_t)_{t\geq 1}$ in $\mathcal{X}$ and $(y_t)_{t\geq 1}$ in $\mathcal{Y}$. At step $t \geq 1$, after observing the values x_i for $1 \leq i \leq t$ and y_i for $1 \leq i \leq t-1$, we define for any $i \in I_t$ the loss $L_{t-1,i} := \sum_{s=t_i}^{t-1} \ell(f^i(x_s), y_s)$. For any $M \geq 1$ we define the function $\phi_M : \mathcal{Y} \to \mathcal{Y}$ such that

$$\phi_M(y) = \begin{cases} y & \text{if } \ell(y, y^0) < M, \\ y^0 & \text{otherwise.} \end{cases}$$

We now construct construct some weights $w_{t,i}$ for $t \geq 1$ and $i \in I_t$ recursively in the following way. Note that $I_1 = \{0\}$. Therefore, we pose $w_{0,0} = 1$. Now let $t \geq 2$ and suppose that $w_{s-1,i}$ have been constructed for all $1 \leq s \leq t-1$. We define

$$\hat{\ell}_s := \frac{\sum_{j\in I_s} w_{s-1,j}\ell(f^j(x_s), \phi_{2^{-\alpha+1}\ln s}(y_s))}{\sum_{j\in I_s} w_{s-1,j}}$$

and for any $i \in I_t$ we note $\hat{L}_{t-1,i} := \sum_{s=t_i}^{t-1} \hat{\ell}_s$. In particular, if $t_i = t$ we have $\hat{L}_{t-1,i} = L_{t-1,i} = 0$. The weights at time t are constructed as $w_{t-1,i} := e^{\eta_t(\hat{L}_{t-1,i} - L_{t-1,i})}$ for any $i \in I_t$. Last, let $\{\hat{i}_t\}_{t\geq 1}$ a sequence of independent random $\mathbb{N}-$valued variables such that

$$\mathbb{P}(\hat{i}_t = i) = \frac{w_{t-1,i}}{\sum_{j\in I_t} w_{t-1,j}}, \quad i \in I_t.$$

Finally, the prediction is defined as $\hat{y}_t := f^{\hat{i}_t}(x_t)$. The learning rule is summarized in Algorithm 1.

For simplicity, we will refer to the predictions of the learning rule as $(\hat{Y}_t)_{t\geq 1}$. Now consider a process $(\mathbb{X}, \mathbb{Y})$ with $\mathbb{X} \in \text{CS}$ and such that $\mathbb{Y}$ is empirically integrable. By Lemma 7.3, there exists $y_0 \in \mathcal{Y}$ such that on an event $\mathcal{A}$ of probability one, for any $\epsilon > 0$, there exists $M_\epsilon \geq 0$ with $\limsup_{T\to\infty} \frac{1}{T}\sum_{t=1}^T \ell(y_0, Y_t)\mathbb{1}_{\ell(y_0,Y_t)\geq M_\epsilon} \leq \epsilon$. We will now denote $\tilde{\mathbb{Y}}$ the process defined by $\tilde{Y}_t = \phi_{2^{-\alpha+1}\ln t}(Y_t)$ for all $t \geq 1$. Then, for any $i \in I_t$, note that using Lemma A.1 we have

$$0 \leq \ell(f^i(x_t), \tilde{Y}_t) \leq 2^{\alpha-1}\left(\ell(f^i(x_t), y^0) + \ell(y^0, \tilde{Y}_t)\right) \leq 2\ln t,$$

by construction of the set I_t. As a result, for any $i, j \in I_t$, we obtain $|\ell(f^i(x_t), \tilde{Y}_t^M) - \ell(f^j(x_t) - \tilde{Y}_t^M)| \leq 2\ln t$. Hence, we can use the same proof as for Theorem 3.6 and show that almost surely, there exists $\hat{t} \geq 1$ such that

$$\forall t \geq \hat{t}, \forall i \in I_t, \quad \sum_{s=t_i}^{t} \ell(\hat{Y}_s, \tilde{Y}_s^M) \leq L_{t,i} + 3\ln^2 t\sqrt{t}.$$

We denote by $\mathcal{B}$ this event. Now let $f : \mathcal{X} \to \mathcal{Y}$ to which we compare the predictions of our learning rule. For any $M \geq 1$, the function $\phi_M \circ f$ is measurable and has values in the

Input: Historical samples $(X_t, Y_t)_{t<T}$ and new input point X_T
Output: Predictions $\hat{Y}_t$ for $t \leq T$
Construct the sequence of measurable functions $\{f^i, i\geq 0\}$ with $f^i = f_{\{l_1^i,\ldots,l_k^i\},\{A_1^i,\ldots,A_k^i\}}$
$I_t := \{i \leq \ln t, \ell(y^{l_p^i}, y^0) \leq 2^{-\alpha+1}\ln t, \forall 1\leq p\leq k^i\}, \mathcal{F}_t := \{f^i, i\in I_t\}, \eta_t := \frac{1}{\ln t\sqrt{t}}, t\geq 1$
$t_i = \min\{t : i \in I_t\}, i\geq 0$
$w_{0,0} := 1, \quad \hat{Y}_1 = y^0 (= f^0(X_0))$ `// Initialisation`
for $t = 2,\ldots,T$ **do**
$L_{t-1,i} = \sum_{s=t_i}^{t-1} \ell(f^i(X_s), \phi_{2^{-\alpha+1}\ln t}(Y_s)), \quad \hat{L}_{t-1,i} = \sum_{s=t_i}^{t-1}\hat{\ell}_s, \quad i\in I_t$
$w_{t-1,i} := \exp(\eta_t(\hat{L}_{t-1,i} - L_{t-1,i})), \quad i \in I_t$
$p_t(i) = \frac{w_{t-1,i}}{\sum_{j\in I_t} w_{t-1,j}}, \quad i\in I_t$
$\hat{i}_t \sim p_t(\cdot)$ `// Function selection`
$\hat{Y}_t = f^{\hat{i}_t}(X_t)$
$\hat{\ell}_t := \frac{\sum_{j\in I_t} w_{t-1,j}\ell(f^j(X_s), \phi_{2^{-\alpha+1}\ln t}(Y_t))}{\sum_{j\in I_t} w_{t-1,j}}$
end

Algorithm 1: A learning rule for adversarial empirically integrable responses under CS processes.

ball $B_\ell(y_0, M)$ where the loss is bounded by $2^\alpha M$. Hence, by Lemma 24 from [5] because $\mathbb{X}\in\mathcal{C}_1$ we have

$$\inf_{i\geq 0} \mathbb{E}\left[\hat{\mu}_{\mathbb{X}}(\ell(\phi_M\circ f(\cdot), f^i(\cdot)))\right] = 0.$$

Now for any $k\geq 0$, let $i_k\geq 0$ such that $\mathbb{E}\left[\hat{\mu}_{\mathbb{X}}(\ell(\phi_M\circ f(\cdot), f^{i_k}(\cdot)))\right] < 2^{-2k}$. By Markov inequality, we have

$$\mathbb{P}\left[\hat{\mu}_{\mathbb{X}}(\ell(\phi_M\circ f(\cdot), f^i(\cdot))) < 2^{-k}\right] \geq 1 - 2^{-k}.$$

Because $\sum_k 2^{-k} < \infty$, the Borel-Cantelli lemma implies that almost surely there exists $\hat{k}$ such that for any $k\geq\hat{k}$, the above inequality is met. We denote $\mathcal{E}_M$ this event. On the event $\mathcal{B}\cap\mathcal{E}_M$ of probability one, for $k\geq\hat{k}$ and any $T\geq\max(t_{i_k},\hat{t})$ we have for any $\epsilon>0$,

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^T\left(\ell(\hat{Y}_t,\tilde{Y}_t) - \ell(\phi_M\circ f(X_t),\tilde{Y}_t)\right)\\
&= \frac{1}{T}\sum_{t=1}^T \ell(\hat{Y}_t,\tilde{Y}_t) - \ell(f^{i_k}(X_t),\tilde{Y}_t) + \frac{1}{T}\sum_{t=1}^T \ell(f^{i_k}(X_t),\tilde{Y}_t) - \ell(\phi_M\circ f(X_t),\tilde{Y}_t)\\
&\leq \frac{1}{T}\sum_{t=1}^{t_{i_k}-1}\ell(\hat{Y}_t,\tilde{Y}_t) + \frac{1}{T}\left(\sum_{t=t_{i_k}}^T \ell(\hat{Y}_t,\tilde{Y}_t) - L_{T,i_k}\right) + \frac{\epsilon}{T}\sum_{t=1}^T \ell(\phi_M\circ f(X_t),\tilde{Y}_t)\\
&\qquad + \frac{c_\epsilon^\alpha}{T}\sum_{t=1}^T \ell(f^{i_k}(X_t),\phi_M\circ f(X_t))\\
&\leq \frac{2\ln t_{i_k}}{T} + \frac{3\ln^2 T}{\sqrt{T}} + \epsilon 2^{\alpha-1}M + \epsilon 2^{\alpha-1}\frac{1}{T}\sum_{t=1}^T\ell(y^0,\tilde{Y}_t) + \frac{c_\epsilon^\alpha}{T}\sum_{t=1}^T \ell(f^{i_k}(X_t),\phi_M\circ f(X_t))
\end{aligned}$$

$$\leq \frac{2\ln t_{i_k}}{T} + \frac{3\ln^2 T}{\sqrt{T}} + \epsilon 2^{\alpha-1}M + \epsilon 2^{\alpha-1}\frac{1}{T}\sum_{t=1}^T \ell(y^0, Y_t) + \frac{c_\epsilon^\alpha}{T}\sum_{t=1}^T \ell(f^{i_k}(X_t), \phi_M\circ f(X_t)),$$

where in the last inequality we used the inequality $\ell(y^0,\tilde Y_t)\leq \ell(y^0,Y_t)$ by construction of $\tilde Y_t = \phi_{2^{-\alpha+1}\ln t}(Y_t)$. Now on the event $\mathcal{A}$, we have

$$\begin{aligned}
Z_1 := \limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^T \ell(y^0,Y_t) &\leq 2^{\alpha-1}\ell(y_0,y^0) + 2^{\alpha-1}\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^T \ell(y_0,Y_t)\\
&\leq 2^{\alpha-1}\ell(y_0,y^0) + 2^{\alpha-1}\left(M_1 + \limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^T \ell(y_0,Y_t)\mathbb{1}_{\ell(y_0,Y_t)\geq M_1}\right)\\
&\leq 2^{\alpha-1}\ell(y_0,y^0) + 2^{\alpha-1}(M_1+1) < \infty.
\end{aligned}$$

Thus, on the event $\mathcal{A}\cap\mathcal{B}\cap\mathcal{E}_M$, for any $k\geq \hat k$ we have for any $\epsilon>0$,

$$\limsup_T \frac{1}{T}\sum_{t=1}^T \ell(\hat Y_t,\tilde Y_t) - \ell(\phi_M\circ f(X_t),\tilde Y_t)) \leq \epsilon 2^{\alpha-1}M + \epsilon 2^{\alpha-1}Z_1 + \frac{c_\epsilon^\alpha}{2^k}.$$

Let $\delta>0$. Now taking $\epsilon = \frac{1}{2^\alpha(M+Z_1)}$, we obtain that on the event $\mathcal{A}\cap\mathcal{B}\cap\mathcal{E}_M$, for any $k\geq\hat k$, we have $\limsup_T \frac{1}{T}\sum_{t=1}^T \ell(\hat Y_t,\tilde Y_t) - \ell(\phi_M\circ f(X_t),\tilde Y_t)) \leq \delta + \frac{c_\epsilon^\alpha}{2^k}$. This yields $\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^T \ell(\hat Y_t,\tilde Y_t) - \ell(\phi_M\circ f(X_t),\tilde Y_t)) \leq \delta$. Because this holds for any $\delta>0$ we obtain $\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^T \ell(\hat Y_t,\tilde Y_t) - \ell(\phi_M\circ f(X_t),\tilde Y_t)) \leq 0$. Finally, on the event $\mathcal{A}\cap\mathcal{B}\cap\bigcap_{M=1}^\infty \mathcal{E}_M$ of probability one, we have

$$\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^T \left(\ell(\hat Y_t,\tilde Y_t) - \ell(\phi_M\circ f(X_t),\tilde Y_t)\right) \leq 0, \quad \forall M\geq 1,$$

where M is an integer. We now observe that on the event $\mathcal{A}$, the same guarantee for y_0 also holds for y^0. Indeed, let ϵ. For $\tilde M_\epsilon := 2^{\alpha-1}(M_{2^{-\alpha}\epsilon} + \ell(y^0,y_0)) + \ell(y_0,y^0)$ we have

$$\begin{aligned}
&\frac{1}{T}\sum_{T=1}^T \ell(y^0,Y_t)\mathbb{1}_{\ell(y^0,Y_t)\geq \tilde M_\epsilon}\\
&\qquad\leq 2^{\alpha-1}\ell(y^0,y_0)\frac{1}{T}\sum_{t=1}^T \mathbb{1}_{\ell(y^0,Y_t)\geq\tilde M_\epsilon} + 2^{\alpha-1}\frac{1}{T}\sum_{T=1}^T \ell(y_0,Y_t)\mathbb{1}_{\ell(y^0,Y_t)\geq\tilde M_\epsilon}\\
&\qquad\leq 2^{\alpha-1}\ell(y^0,y_0)\frac{1}{T}\sum_{t=1}^T \mathbb{1}_{\ell(y_0,Y_t)\geq 2^{-\alpha+1}M - \ell(y_0,y^0)}\\
&\qquad\qquad + 2^{\alpha-1}\frac{1}{T}\sum_{T=1}^T \ell(y_0,Y_t)\mathbb{1}_{\ell(y_0,Y_t)\geq 2^{-\alpha+1}M-\ell(y_0,y^0)}\\
&\qquad\leq 2^\alpha\frac{1}{T}\sum_{t=1}^T \ell(y_0,Y_t)\mathbb{1}_{\ell(y_0,Y_t)\geq M_{2^{-\alpha}\epsilon}}
\end{aligned}$$

Hence, we obtain $\limsup_{T\to\infty}\frac{1}{T}\sum_{T=1}^T \ell(y^0,Y_t)\mathbb{1}_{\ell(y^0,Y_t)\geq\tilde M_\epsilon}\leq\epsilon$. We now write

$$\frac{1}{T}\sum_{t=1}^T \ell(\phi_M\circ f(X_t),\tilde Y_t) - \ell(f(X_t),Y_t)$$

$$\leq \frac{1}{T}\sum_{t=1}^{T}\left(\ell(y^0,Y_t)-\ell(f(X_t),Y_t)\right)\mathbb{1}_{\ell(f(X_t),y^0)\geq M}\mathbb{1}_{\ell(Y_t,y^0)\leq \ln t}$$

$$+\frac{1}{T}\sum_{t=1}^{T}\left(\ell(f(X_t),y^0)-\ell(f(X_t),Y_t)\right)\mathbb{1}_{\ell(f(X_t),y^0)\leq M}\mathbb{1}_{\ell(Y_t,y^0)\geq 2^{-\alpha+1}\ln t}$$

$$\leq \frac{1}{T}\sum_{t=1}^{T}\left(2\ell(y^0,Y_t)-2^{-\alpha+1}\ell(f(X_t),y^0)\right)\mathbb{1}_{\ell(f(X_t),y^0)\geq M}$$

$$+\frac{1}{T}\sum_{t=1}^{T}\left(2\ell(f(X_t),y^0)-2^{-\alpha+1}\ell(y^0,Y_t)\right)\mathbb{1}_{\ell(f(X_t),y^0)\leq M}\mathbb{1}_{\ell(Y_t,y^0)\geq 2^{-\alpha+1}\ln t}$$

$$\leq \frac{2}{T}\sum_{t=1}^{T}\ell(y^0,Y_t)\mathbb{1}_{\ell(Y_t,y^0)\geq 2^{-\alpha}M}+\frac{2Me^{2^{2\alpha-1}M}}{T}.$$

As a result, on the event $\mathcal{A}\cap\mathcal{B}\cap\bigcap_{M=1}^{\infty}\mathcal{E}_M$, for any $M\geq 1$,

$$\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\ell(\phi_M\circ f(X_t),\tilde{Y}_t)-\ell(f(X_t),Y_t)\leq 2\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\ell(y^0,Y_t)\mathbb{1}_{\ell(Y_t,y^0)\geq 2^{-\alpha}M}.$$

Last, we compute

$$\frac{1}{T}\sum_{t=1}^{T}\ell(\hat{Y}_t,Y_t)-\ell(\hat{Y}_t,\tilde{Y}_t)=\frac{1}{T}\sum_{t=1}^{T}\left(\ell(\hat{Y}_t,Y_t)-\ell(\hat{Y}_t,y^0)\right)\mathbb{1}_{\ell(Y_t,y^0)\geq 2^{-\alpha+1}\ln t}$$

$$\leq \frac{1}{T}\sum_{t=1}^{T}\left(2^{\alpha-1}\ell(\hat{Y}_t,y^0)+2^{\alpha-1}\ell(Y_t,y^0)\right)\mathbb{1}_{\ell(Y_t,y^0)\geq 2^{-\alpha+1}\ln t}$$

$$\leq \frac{1}{T}\sum_{t=1}^{T}\left(\ln t+2^{\alpha-1}\ell(Y_t,y^0)\right)\mathbb{1}_{\ell(Y_t,y^0)\geq 2^{-\alpha+1}\ln t}$$

$$\leq \frac{2^{\alpha}}{T}\sum_{t=1}^{T}\ell(Y_t,y^0)\mathbb{1}_{\ell(Y_t,y^0)\geq 2^{-\alpha+1}\ln t}.$$

Note that for any $\epsilon>0$, we have on the event $\mathcal{A}$ that for any $M\geq 1$,

$$\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\ell(Y_t,y^0)\mathbb{1}_{\ell(Y_t,y^0)\geq 2^{-\alpha+1}\ln t}\leq \limsup_{T\to\infty}\frac{1}{T}\sum_{t\geq e^{2^{\alpha-1}M}}^{T}\ell(Y_t,y^0)\mathbb{1}_{\ell(Y_t,y^0)\geq M}$$

$$=\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\ell(Y_t,y^0)\mathbb{1}_{\ell(Y_t,y^0)\geq M}.$$

Hence, because this holds for any $M\geq 1$, if $\epsilon>0$ we can apply this to the integer $M:=\lceil\tilde{M}_\epsilon\rceil$ which yields $\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\ell(Y_t,y^0)\mathbb{1}_{\ell(Y_t,y^0)\geq 2^{-\alpha+1}\ln t}\leq\epsilon$. This holds for any $\epsilon>0$. Hence we obtain on the event $\mathcal{A}$ that $\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\ell(Y_t,y^0)\mathbb{1}_{\ell(Y_t,y^0)\geq 2^{-\alpha+1}\ln t}\leq 0$, which implies that $\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\ell(\hat{Y}_t,Y_t)-\ell(\hat{Y}_t,\tilde{Y}_t)\leq 0$. Putting everything together, we obtain on $\mathcal{A}\cap\mathcal{B}\cap\bigcap_{M=1}^{\infty}\mathcal{E}_M$ that for any $M\geq 1$,

$$\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\ell(\hat{Y}_t,Y_t)-\ell(f(X_t),Y_t)\leq \limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\ell(\hat{Y}_t,Y_t)-\ell(\hat{Y}_t,\tilde{Y}_t)$$

$$
\begin{aligned}
&+\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^T \ell(\hat Y_t,\tilde Y_t)-\ell(\phi_M\circ f(X_t),\tilde Y_t)\\
&+\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^T \ell(\phi_M\circ f(X_t),\tilde Y_t)-\ell(f(X_t),Y_t)\\
\leq\; & 2\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^T \ell(y^0,Y_t)\mathbb{1}_{\ell(Y_t,y^0)\geq 2^{-\alpha}M}.
\end{aligned}
$$

Because this holds for all $M\geq 1$, we can again apply this result to $M:=\lceil \tilde M_\epsilon\rceil$ which yields $\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^T\ell(\hat Y_t,Y_t)-\ell(f(X_t),Y_t)\leq\epsilon$. This holds for any $\epsilon>0$. Therefore, we finally obtain on the event $\mathcal{A}\cap\mathcal{B}\cap\bigcap_{M=1}^\infty\mathcal{E}_M$ of probability one, one has $\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^T\ell(\hat Y_t,Y_t)-\ell(f(X_t),Y_t)\leq 0$. This ends the proof that Algorithm 1 is universally consistent under CS processes for adversarial empirically integrable responses. Now because there exists a ball $B_\ell(y,r)$ of $(\mathcal{Y},\ell)$ that does not satisfy F-TiME, from Theorem 5.8, universal learning with responses restricted on this ball cannot be achieved for processes $\mathbb{X}\notin$ CS. However, these responses are empirically integrable because they are bounded. Hence, CS is still necessary for universal learning with adversarial empirically integrable responses. Thus SOLAR $=$ CS and the provided learning rule is optimistically universal. This ends the proof of the theorem.

E.5. Proof of Theorem 3.2. Fix $(\mathcal{X},\rho_\mathcal{X})$ and a value space $(\mathcal{Y},\ell)$ such that any ball satisfies F-TiME We now construct our learning rule. Let $\bar y\in\mathcal{Y}$ be an arbitrary value. For any $M\geq 1$, because $B_\ell(\bar y,M)$ is bounded and satisfies F-TiME, there exists an optimistically universal learning rule $f_\cdot^M$ for value space $(B_\ell(y_0,M),\ell)$. For any $M\geq 1$, we define the function $\phi_M:\mathcal{Y}\to\mathcal{Y}$ defined by restricting the space to the ball $B_\ell(\bar y,M)$ as follows

$$
\phi_M(y):=\begin{cases} y & \text{if } \ell(y,\bar y)<M\\ \bar y & \text{otherwise.}\end{cases}
$$

For simplicity, we will denote by $\hat Y_t^M:=f_t^M(\mathbb{X}_{\leq t-1},\phi_M(\mathbb{Y})_{\leq t-1},X_t)$ the prediction of $f_\cdot^M$ at time t for the responses which are restricted to the ball $B_\ell(\bar y,M)$. We now combine these predictors using online learning into a final learning rule $f_\cdot$. Specifically, we define $I_t:=\{0\leq M\leq 2^{-\alpha+1}\ln t\}$ for all $t\geq 1$. We also denote $t_M=\lceil e^{2^{\alpha-1}M}\rceil$ for $M\geq 0$ and pose $\eta_t=\frac{1}{4\sqrt t}$. For any $M\in I_t$, we define

$$
L_{t-1,M}:=\sum_{s=t_M}^{t-1}\ell(\hat Y_s^M,\phi_{2^{-\alpha+1}\ln s}(Y_s)).
$$

For simplicity, we will denote by $\tilde{\mathbb{Y}}$ the process defined by $\tilde Y_t=\phi_{2^{-\alpha+1}\ln t}(Y_t)$ for all $t\geq 1$. We now construct recursive weights as $w_{0,0}=1$ and for $t\geq 2$ we pose for all $1\leq s\leq t-1$

$$
\hat l_s:=\frac{\sum_{M\in I_s}w_{s-1,M}\ell(\hat Y_s^M,\tilde Y_s)}{\sum_{M\in I_s}w_{s-1,M}}.
$$

Now for any $M\in I_t$ we note $\hat L_{t-1,M}:=\sum_{s=t_M}^{t-1}\hat\ell_s$, and pose $w_{t-1,M}:=e^{\eta_t(\hat L_{t-1,M}-L_{t-1,M})}$. We then choose a random index $\hat M_t$ independent from the past history such that

$$
\mathbb{P}(\hat M_t=M):=\frac{w_{t-1,M}}{\sum_{M'\in I_t}w_{t-1,M'}},\quad M\in I_t.
$$

Input: Historical samples $(X_t, Y_t)_{t<T}$ and new input point X_T

Optimistically universal learning rule $f_\cdot^M$ for value space $B_\ell(y_0, M), \ell)$, where $y_0 \in \mathcal{Y}$ fixed.

Output: Predictions $\hat{Y}_t$ for $t \leq T$

$I_t := \{0 \leq M \leq 2^{-\alpha+1} \ln t\}, \eta_t := \frac{1}{4\sqrt{t}}, t \geq 1$

$t_M = \lceil e^{2^{\alpha-1}M} \rceil, M \geq 0$

$w_{0,0} := 1, \quad \hat{Y}_1 = y^0 (= f^0(X_0))$ `// Initialisation`

for $t = 2, \ldots, T$ **do**

$\quad L_{t-1,M} = \sum_{s=t_M}^{t-1} \ell(f_s^M(\mathbb{X}_{\leq s-1}, \phi_M(\mathbb{Y})_{\leq s-1}, X_s), \phi_{2^{-\alpha+1} \ln s}(Y_s)), \quad \hat{L}_{t-1,M} = \sum_{s=t_M}^{t-1} \hat{\ell}_s, \quad M \in I_t$

$\quad w_{t-1,M} := \exp(\eta_t(\hat{L}_{t-1,M} - L_{t-1,M})), \quad M \in I_t$

$\quad p_t(M) = \frac{w_{t-1,M}}{\sum_{M' \in I_t} w_{t-1,M'}}, \quad M \in I_t$

$\quad \hat{M}_t \sim p_t(\cdot)$ `// Model selection`

$\quad \hat{Y}_t = f_t^{\hat{M}_t}(\mathbb{X}_{\leq t-1}, \phi_M(\mathbb{Y})_{\leq t-1}, X_t)$

$\quad \hat{\ell}_t := \frac{\sum_{j \in I_t} w_{t-1,j} \ell(f_t^M(\mathbb{X}_{\leq t-1}, \phi_M(\mathbb{Y})_{\leq t-1}, X_t), \phi_{2^{-\alpha+1} \ln t}(Y_t))}{\sum_{j \in I_t} w_{t-1,j}}$

end

Algorithm 2: A learning rule for adversarial empirically integrable responses under SMV processes for value spaces $(\mathcal{Y}, \ell)$ such that any ball satisfies F-TiME.

The output the learning rule is $f_t(\mathbb{X}_{\leq t-1}, \mathbb{Y}_{\leq t-1}, X_t) := \hat{Y}_t^{\hat{M}_t}$. For simplicity, we will denote by $\hat{Y}_t := f_t(\mathbb{X}_{\leq t-1}, \mathbb{Y}_{\leq t-1}, X_t)$ the prediction of $f_\cdot$ at time t. This ends the construction of our learning rule which is summarized in Algorithm 2.

Now let $(\mathbb{X}, \mathbb{Y})$ be such that $\mathbb{X} \in$ SOUL and $\mathbb{Y}$ empirically integrable. By Lemma 7.3, there exists some value $y_0 \in \mathcal{Y}$ such that on an event $\mathcal{A}$ of probability one, we have for any ϵ, a threshold $M_\epsilon \geq 0$ with $\limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^T \ell(y_0, Y_t) \mathbb{1}_{\ell(y_0, Y_t) \geq M_\epsilon} \leq \epsilon$. We fix a measurable function $f : \mathcal{X} \to \mathcal{Y}$. Also, for any $t \geq 1$ and $M \in I_t$ we have $0 \leq \ell(\hat{Y}_t^M, \tilde{Y}_t) \leq 2^{\alpha-1}\ell(\hat{Y}_t^M, \bar{y}) + 2^{\alpha-1}\ell(\tilde{Y}_t, \bar{y}) \leq 2 \ln t$. As a result, for any $M, M' \in I_t$ we have $|\ell(\hat{Y}_t^M, \tilde{Y}_t) - \ell(\hat{Y}_t^{M'}, \tilde{Y}_t)| \leq 2 \ln t$. Because $|I_t| \leq 1 + \ln t$ for all $t \geq 1$, the same proof as Theorem 3.6 shows that on an event $\mathcal{B}$ of probability one, there exists $\hat{t} \geq 0$ such that

$$\forall t \geq \hat{t}, \forall M \in I_t, \quad \sum_{s=t_M}^{t} \ell(\hat{Y}_t, \tilde{Y}_t) \leq \sum_{s=t_M}^{t} \ell(\hat{Y}_t^M, \tilde{Y}_t) + 3 \ln^2 t \sqrt{t}.$$

Further, we know that $f_\cdot^M$ is Bayes optimistically universal for value space $(B_\ell(\bar{y}, M), \ell)$. In particular, because $\mathbb{X} \in$ SOUL and $\phi_M \circ f : \mathcal{X} \to B_\ell(\bar{y}, M)$, we have

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \ell(\hat{Y}_t^M, \phi_M(Y_t)) - \ell(\phi_M \circ f(X_t), \phi_M(Y_t)) \leq 0 \quad (a.s.).$$

For simplicity, we introduce $\delta_T^M := \frac{1}{T} \sum_{t=1}^T \ell(\hat{Y}_t^M, \phi_M(Y_t)) - \ell(\phi_M \circ f(X_t), \phi_M(Y_t))$ and define $\mathcal{E}_M$ as the event of probability one where the above inequality is satisfied, i.e., $\limsup_{T\to\infty} \delta_T^M \leq 0$. Because we always have $\ell(\hat{Y}_t, \bar{y}) \leq 2^{-\alpha+1} \ln t$, we can write

$$\begin{aligned}
\frac{1}{T} \sum_{t=1}^{T} \ell(\hat{Y}_t, Y_t) - \ell(\hat{Y}_t, \tilde{Y}_t) &= \frac{1}{T} \sum_{t=1}^{T} \left(\ell(\hat{Y}_t, Y_t) - \ell(\hat{Y}_t, \bar{y}) \right) \mathbb{1}_{\ell(Y_t, \bar{y}) \geq 2^{-\alpha+1} \ln t} \\
&\leq \frac{1}{T} \sum_{t=1}^{T} \left(2^{\alpha-1} \ell(\hat{Y}_t, \bar{y}) + 2^{\alpha-1} \ell(Y_t, \bar{y}) \right) \mathbb{1}_{\ell(Y_t, \bar{y}) \geq 2^{-\alpha+1} \ln t}
\end{aligned}$$

$$\leq \frac{2^\alpha}{T}\sum_{t=1}^T \ell(Y_t,\bar y)\mathbb{1}_{\ell(Y_t,\bar y)\geq 2^{-\alpha+1}\ln t}.$$

The proof of Theorem 3.3 shows that on the event $\mathcal{A}$,

$$\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^T \ell(Y_t,\bar y)\mathbb{1}_{\ell(Y_t,\bar y)\geq 2^{-\alpha+1}\ln t}\leq 0,$$

which implies $\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^T \ell(\hat Y_t,Y_t)-\ell(\hat Y_t,\tilde Y_t)\leq 0$. Now let $M\geq 1$. We write

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^T \ell(\hat Y_t^M,\tilde Y_t)-\ell(\hat Y_t^M,\phi_M(Y_t))\\
&\qquad\leq \frac{1}{T}\sum_{t=1}^{t_M-1}\ell(\hat Y_t^M,\tilde Y_t)+\frac{1}{T}\sum_{t=t_M}^T\left(\ell(\hat Y_t^M,Y_t)-\ell(\hat Y_t^M,\bar y)\right)\mathbb{1}_{M\leq \ell(Y_t,\bar y)<2^{-\alpha+1}\ln t}\\
&\qquad\leq \frac{e^{2^{\alpha-1}M}2^\alpha M}{T}+\frac{1}{T}\sum_{t=1}^T\left(2^{\alpha-1}\ell(\hat Y_t^M,\bar y)+2^{\alpha-1}\ell(Y_t,\bar y)\right)\mathbb{1}_{\ell(Y_t,\bar y)\geq M}\\
&\qquad\leq \frac{e^{2^{\alpha-1}M}2^\alpha M}{T}+\frac{2^\alpha}{T}\sum_{t=1}^T \ell(Y_t,\bar y)\mathbb{1}_{\ell(Y_t,\bar y)\geq M}.
\end{aligned}$$

Hence, on the event $\mathcal{A}$, we obtain

$$\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^T \ell(\hat Y_t^M,\tilde Y_t)-\ell(\hat Y_t^M,\phi_M(Y_t))\leq 2^\alpha\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^T \ell(Y_t,\bar y)\mathbb{1}_{\ell(Y_t,\bar y)\geq M}.$$

Finally, we compute

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^T \ell(\phi_M\circ f(X_t),\phi_M(Y_t))-\ell(f(X_t),Y_t)\\
&\leq \frac{1}{T}\sum_{t=1}^T\left(\ell(\bar y,Y_t)-\ell(f(X_t),Y_t)\right)\mathbb{1}_{\ell(f(X_t),\bar y)\geq M}\mathbb{1}_{\ell(Y_t,\bar y)\leq M}\\
&\qquad+\frac{1}{T}\sum_{t=1}^T\left(\ell(f(X_t),\bar y)-\ell(f(X_t),Y_t)\right)\mathbb{1}_{\ell(f(X_t),\bar y)\leq M}\mathbb{1}_{\ell(Y_t,\bar y)\geq M}\\
&\leq \frac{1}{T}\sum_{t=1}^T \ell(\bar y,Y_t)\mathbb{1}_{\ell(Y_t,\bar y)\geq 2^{-\alpha}M}+\frac{M}{T}\sum_{t=1}^T\mathbb{1}_{\ell(Y_t,\bar y)\geq M}\\
&\qquad+\frac{1}{T}\sum_{t=1}^T\left(\ell(\bar y,Y_t)-\ell(f(X_t),Y_t)\right)\mathbb{1}_{\ell(f(X_t),\bar y)\geq M}\mathbb{1}_{\ell(Y_t,\bar y)\leq 2^{-\alpha}M}\\
&\leq \frac{1}{T}\sum_{t=1}^T \ell(\bar y,Y_t)\mathbb{1}_{\ell(Y_t,\bar y)\geq 2^{-\alpha}M}+\frac{1}{T}\sum_{t=1}^T \ell(Y_t,\bar y)\mathbb{1}_{\ell(Y_t,\bar y)\geq M}\\
&\qquad+\frac{1}{T}\sum_{t=1}^T\left(2\ell(\bar y,Y_t)-2^{-\alpha+1}\ell(f(X_t),\bar y)\right)\mathbb{1}_{\ell(f(X_t),\bar y)\geq M}\mathbb{1}_{\ell(Y_t,\bar y)\leq 2^{-\alpha}M}
\end{aligned}$$

$$\leq \frac{1}{T}\sum_{t=1}^{T} \ell(\bar{y}, Y_t)\mathbb{1}_{\ell(Y_t,\bar{y})\geq 2^{-\alpha}M} + \frac{1}{T}\sum_{t=1}^{T} \ell(Y_t, \bar{y})\mathbb{1}_{\ell(Y_t,\bar{y})\geq M}.$$

We now put all these estimates together. On the event $\mathcal{A}\cap\mathcal{B}\cap\bigcap_{M=1}^{\infty}\mathcal{E}_M$, for any $M\geq 1$ and $t\geq \max(\hat{t}, t_M)$ we can write

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} \ell(\hat{Y}_t, Y_t) - \ell(f(X_t), Y_t) &\leq \frac{1}{T}\sum_{t=1}^{T}\left(\ell(\hat{Y}_t, Y_t) - \ell(\hat{Y}_t, \tilde{Y}_t)\right)\\
&+ \frac{1}{T}\sum_{t=1}^{T}\left(\ell(\hat{Y}_t, \tilde{Y}_t) - \ell(\hat{Y}_t^M, \tilde{Y}_t)\right) + \frac{1}{T}\sum_{t=1}^{T}\left(\ell(\hat{Y}_t^M, \tilde{Y}_t) - \ell(\hat{Y}_t^M, \phi_M(Y_t))\right) + \delta_T^M\\
&+ \frac{1}{T}\sum_{t=1}^{T}\left(\ell(\phi_M\circ f(X_t), \phi_M(Y_t)) - \ell(f(X_t), Y_t)\right)\\
&\leq \frac{1}{T}\sum_{t=1}^{T}\left(\ell(\hat{Y}_t, Y_t) - \ell(\hat{Y}_t, \tilde{Y}_t)\right) + \frac{3\ln^2 T}{\sqrt{T}} + \frac{1}{T}\sum_{t=1}^{T}\left(\ell(\hat{Y}_t^M, \tilde{Y}_t) - \ell(\hat{Y}_t^M, \phi_M(Y_t))\right)\\
&\qquad + \delta_T^M + \frac{1}{T}\sum_{t=1}^{T}\left(\ell(\phi_M\circ f(X_t), \phi_M(Y_t)) - \ell(f(X_t), Y_t)\right).
\end{aligned}$$

Thus, we obtain on the event $\mathcal{A}\cap\mathcal{B}\cap\bigcap_{M=1}^{\infty}\mathcal{E}_M$, for any $M\geq 1$,

$$\begin{aligned}
\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} \ell(\hat{Y}_t, Y_t) - \ell(f(X_t), Y_t) &\leq \limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\ell(\bar{y}, Y_t)\mathbb{1}_{\ell(Y_t,\bar{y})\geq 2^{-\alpha}M}\\
&\qquad + (1+2^{\alpha})\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\ell(Y_t, \bar{y})\mathbb{1}_{\ell(Y_t,\bar{y})\geq M}
\end{aligned}$$

On the event $\mathcal{A}$, the same arguments as in the proof of Theorem 3.3 show that we have same guarantees for y_0 as for $\bar{y}$, i.e., for any $\epsilon>0$, there exists $\tilde{M}_\epsilon$ such that $\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\ell(Y_t, \bar{y})\mathbb{1}_{\ell(Y_t,\bar{y})\geq \tilde{M}_\epsilon}\leq \epsilon$. Therefore, for any $\epsilon>0$, we can apply the above equation to $M:=\lceil 2^{\alpha}M_\epsilon + M_{2^{-\alpha-1}\epsilon}\rceil$ to obtain

$$\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} \ell(\hat{Y}_t, Y_t) - \ell(f(X_t), Y_t) \leq \epsilon + \frac{1+2^{\alpha}}{2^{\alpha+1}} \leq 2\epsilon.$$

Because this holds for all $\epsilon>0$, we can in finally get

$$\limsup_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\left(\ell(\hat{Y}_t, Y_t) - \ell(f(X_t), Y_t)\right) \leq 0,$$

on the event $\mathcal{A}\cap\mathcal{E}\cap\bigcap_{M\geq 1}\mathcal{F}_M$ of probability one. This ends the proof of the theorem.

REFERENCES

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

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

[3] CESA-BIANCHI, N. and LUGOSI, G. (2006). *Prediction, learning, and games*. Cambridge university press.

[4] EVANS, S. N. and JAFFE, A. Q. (2020). Strong laws of large numbers for Fréchet means. *arXiv preprint arXiv:2012.12859*.

[5] HANNEKE, S. (2021). Learning whenever learning is possible: Universal learning under general stochastic processes. *Journal of Machine Learning Research* **22** 1–116.