Prediction with Finitely many Errors Almost Surely

Changlong Wu
University of Hawaii at Manoa

Narayana Santhanam
University of Hawaii at Manoa

Abstract

Using only samples from a probabilistic model, we predict properties of the model and of future observations. The prediction game continues in an online fashion as the sample size grows with new observations. After each prediction, the predictor incurs a binary (0-1) loss. The probability model underlying a sample is otherwise unknown except that it belongs to a known class of models. The goal is to make finitely many errors (i.e. loss of 1) with probability 1 under the generating model, no matter what it may be in the known model class.

Model classes admitting predictors that make only finitely many errors are *eventually almost surely (eas) predictable*. When the losses incurred are observable (the supervised case), we completely characterize eas predictable classes. We provide analogous results in the unsupervised case. Our results have a natural interpretation in terms of regularization. In eas-predictable classes, we study if it is possible to have a universal stopping rule that identifies (to any given confidence) when no more errors will be made. Classes admitting such a stopping rule are *eas learnable*. When samples are generated *i.i.d.*, we provide a complete characterization of eas learnability. We also study cases when samples are not generated *i.i.d.*, but a full characterization remains open at this point.

1 Introduction

This paper considers a learning framework where the learner has little prior knowledge of the environment, but is able to make as many observations (measurements) as needed. Predictions are required on either an unknown property of the environment or future observations, and can potentially use past observations. These predictions incur a loss. In this setup, we focus on when a learner can ensure that its loss is finite almost surely. This framework forms the general background for multiple lines of research, two of which we outline here.

The first is the learning of *recursive functions*, see (Zeugmann and Zilles, 2008) for a survey. Perhaps the earliest work of this flavor is the language identification problem of Gold (1967). Here one tries to identify a language from a set of languages by observing *information presentations* of the language, and asks when an identification would be correct after finite observations.

A typical setup for learning recursive functions starts with a set $\mathcal{H}$ of functions that maps $\mathbb{N} \to \{0, 1\}$. Nature fixes a function h from $\mathcal{H}$ at the beginning. At each time step n, the learner makes attempts to predict $h(n)$ using the history $h(1), \cdots, h(n-1)$ thus far. Nature then reveals the true value $h(n)$ after the learner has made the prediction. The goal is a *computable* learner that makes only *finitely* many errors no matter what $h \in \mathcal{H}$ is.

The second line of research involves randomized observations, and was initialized by Cover (1973). Here, the learner's goal is to predict the irrationality of the mean of a random variable over $[0, 1]$ using *i.i.d.* observations of it. The prediction can be updated after every observation, but the learner is allowed only finitely many errors with probability 1—and perhaps surprisingly, this is possible in a variety of non-trivial setups, underscoring the distinction between prediction and estimation. Cover's setup was generalized in (Dembo and Peres, 1994) to identify general properties of distributions over $\mathbb{R}^d$, and in continued in (Kulkarni and Tse, 1994; Koplowitz et al., 1995; Naaman, 2016). In (Wu and Santhanam, 2019), the authors predict upper bounds on the next observation of *i.i.d.* sampling from distributions over $\mathbb{N}$, such that the next observation violates the bound only finitely often with probability 1, and

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

in (Santhanam and Anantharam, 2015), the authors obtain a stopping rule that additionally indicates when such a prediction has made the last mistake.

This paper subsumes the problem setups mentioned above into a unified framework. Observations are modeled as a general discrete time random process $X_1, X_2, \cdots$ whose underlying probability measure (not necessarily *i.i.d.*) p belongs to a known collection $\mathcal{P}$. The learning process is a game between Nature and a learner, where the learner attempts to predict a property of p or of future observations. Nature fixes a random process $p \in \mathcal{P}$. At each time step n, the learner makes a prediction Y_n using $X_1, \cdots, X_{n-1}$. The prediction, the next realization X_n, and potentially the underlying process p are associated with a binary 0-1 loss ℓ, where 1 indicates an error/unsatisfactory prediction. The collection $\mathcal{P}$ together with loss ℓ is *eventually almost surely* predictable (e.a.s.-predictable), if there is a strategy such that the learner makes only *finitely* errors with probability 1 no matter what the underlying $p \in \mathcal{P}$ is.

Our contributions in this paper establish a comprehensive theoretical framework for this problem setup.

First, we provide a general characterization of e.a.s.-predictability. In particular, this characterization subsumes prior known results into (e.g. (Cover, 1973; Koplowitz et al., 1995; Dembo and Peres, 1994; Wu and Santhanam, 2019)), and resolves conjectures posed therein.

Second, for e.a.s.-predictable classes, we characterize whether for arbitrary confidence, a stopping rule can exist indicating that the final error is in the past. In particular, this subsumes results in (Santhanam and Anantharam, 2015).

Finally, we demonstrate our characterizations in the context of estimating the singularity, rank and eigenvalues of random matrices.

As we obtain the results, we note that the results provide a natural understanding of regularization, where the model class is restricted to accommodate the amount of data at hand. In particular, we provide striking examples of such regularization at work in the context of Cover's problem described above in Example 2.

2 Problem Setup

Let $\mathcal{X}$ be a set and $\mathcal{P}$ be a collection of probability measures over a fixed cylinder σ-algebras over $\mathcal{X}^\infty$. We consider a discrete time random process $\mathbf{X} = \{X_n\}_{n \in \mathbb{N}^+}$ generated by sampling from a probability law $p \in \mathcal{P}$.

Prediction is modeled as a function $\Phi : \mathcal{X}^* \to \mathcal{Y}$, where $\mathcal{X}^*$ denotes the set of all finite strings of sequences from $\mathcal{X}$, and $\mathcal{Y}$ is the set of all predictions. The *loss* function is a measurable function $\ell : \mathcal{P} \times \mathcal{X}^* \times \mathcal{Y} \to \{0, 1\}$. We consider the property we are estimating to be defined implicitly by the subset of $\mathcal{P} \times \mathcal{X}^* \times \mathcal{Y}$ where $\ell = 0$, and therefore, in a slight abuse of notation sometimes refer to ℓ as a *property* as well.

We consider the following game that proceeds in time indexed by $\mathbb{N}^+$. The game has two parties: the learner and nature. Nature chooses some model $p \in \mathcal{P}$ to begin the game. At each time step n, the learner makes a prediction Y_n based on the current observation X_1^{n-1} generated according to p. Nature then generates X_n based on p and X_1^{n-1}.

The learner fails at step n if $\ell(p, X_1^n, Y_n) = 1$. The learner targets a strategy that minimizes the cumulative loss in the infinite horizon, without knowledge of the model the environment chooses at the beginning.

The loss in general can be a function of the probability model in addition to the sample observed, and our prediction on the sample. When the loss depends on the probability model, there may be no direct way to estimate the loss incurred at say, step n, from observations of the sample X_1^{n-1} even after the prediction Y_n is made. We call such setups the *unsupervised setting* borrowing from learning theory. A special case is the *supervised* setting, where we define the loss to be a function from $\mathcal{X}^* \times \mathcal{Y}$ to $\{0, 1\}$.

Definition 1 ($\eta-$predictability)**.** *A collection $(\mathcal{P}, \ell)$ is η-predictable, if there exists a prediction rule $\Phi : \mathcal{X}^* \to \mathcal{Y}$ and a sample size n such that for all $p \in \mathcal{P}$,*

$$p\left(\sum_{i=n}^{\infty} \ell(p, X_1^i, \Phi(X_1^{i-1})) > 0\right) \leq \eta,$$

i.e. the probability that the learner makes errors after step n is at most η uniformly over $\mathcal{P}$.

Definition 2. *A collection $(\mathcal{P}, \ell)$ is said to be eventually almost surely (e.a.s.)-predictable, if there exists a prediction rule Φ, such that for all $p \in \mathcal{P}$*

$$p\left(\sum_{n=1}^{\infty} \ell(p, X_1^n, \Phi(X_1^{n-1})) < \infty\right) = 1.$$

We need a technical definition that will help simplify notation further.

Definition 3. *A* nesting *of $\mathcal{P}$ is a collection of subsets of $\mathcal{P}$, $\{\mathcal{P}_i : i \geq 1\}$ such that $\mathcal{P}_1 \subset \mathcal{P}_2 \subset \ldots$ and $\bigcup_{i \geq 1} \mathcal{P}_i = \mathcal{P}$.*

The following lemmas characterize immediate connections between the above definitions.

Lemma 1. *Let $\mathcal{P}$ be a collection of models, $\{\mathcal{P}_i, i \geq 1\}$ be a nesting of $\mathcal{P}$. If for all $\eta > 0$ and $i \in \mathbb{N}^+$, $(\mathcal{P}_i, \ell)$ is η-predictable for some loss ℓ. Then $(\mathcal{P}, \ell)$ is e.a.s.-predictable.*

Proof. From the definition of $\eta-$predictability, we can choose an increasing sequence $\{b_i, i \geq 1\}$, and predictors Φ_i for $\mathcal{P}_i$ respectively as follows. For all i and for all $p \in \mathcal{P}_i$, the probability Φ_i makes errors after step b_i is at most 2^{-i}.

The predictor Φ is then constructed from $\{\Phi_i, i \geq 1\}$ as follows: use predictor Φ_i when the length T of the observed sample satisfies $b_i \leq T < b_{i+1}$.

Let $p \in \mathcal{P}_k \subset \mathcal{P}$. Because the collections $\mathcal{P}_i$ are nested, for all $i \geq k$, $p \in \mathcal{P}_i$. During the phase Φ coincides with Φ_i, the probability of Φ making an error is $\leq 2^{-i}$. The result follows using the Borel-Cantelli lemma. □

Lemma 2. *Let $\mathcal{P}$ be a collection of probability measures and let ℓ be a loss function. Suppose $\{\mathcal{P}_i : i \geq 1\}$ is a nesting of $\mathcal{P}$ such that for all $i \geq 1$, $(\mathcal{P}_i, \ell)$ is η-predictable for some $\eta > 0$.*

Then there exists a relabeling of the sets in a nesting $\{\mathcal{P}'_i : i \geq 1\}$ of $\mathcal{P}$ such that $(\mathcal{P}'_i, \ell)$ is η-predictable with sample size i. Namely, there is an estimator Φ_i such that for all $p \in \mathcal{P}'_i$, the probability Φ_i incurs non-zero ℓ-loss on samples with size larger than i is $\leq \eta$.

Proof. Since $\mathcal{P}_i$ is $\eta-$predictable, there exists a number n'_i and an estimator Φ_i such that for all $p \in \mathcal{P}_i$, the probaiblity Φ_i incurs non-zero $\ell-$loss on samples larger than n'_i is $\leq \eta$. We can therefore choose an increasing sequence $\{n_i, i \geq 1\}$ and $n_i \geq n'_i$. Note that each $\mathcal{P}_i$ is $\eta-$predictable with samples size n_i. For $n_k \leq i < n_{k+1}$, set $\mathcal{P}'_i = \mathcal{P}_k$. $\{\mathcal{P}'_i : i \geq 1\}$ is the desired nesting and the lemma follows. □

3 Characterization

We first characterize *e.a.s.*-predictability in supervised setting. This result is reminiscent of how practitioners do regularization, where one restricts $\mathcal{P}$ to match the amount of data available. When $\mathcal{P}$ is e.a.s.-predictable, the implication is that such a regularization approach will settle at some point on a complexity—more data from that point on will not really lead to making the model more complex. While the theorem below can be proved in multiple ways, we provide an approach that reflects the intuition above.

Theorem 1. *Consider a collection $\mathcal{P}$ with a loss $\ell : \mathcal{X}^* \times \mathcal{Y} \to \{0, 1\}$ (i.e. the supervised setting). $(\mathcal{P}, \ell)$ is e.a.s.-predictable iff for all $\eta > 0$, there exists a nesting $\{\mathcal{P}^\eta_n : n \geq 1\}$ of $\mathcal{P}$ such that for all $n \geq 1$, $(\mathcal{P}^\eta_n, \ell)$ is η-predictable.*

Proof. Suppose $\mathcal{P}$ is *e.a.s.*-predictable, we show that $\mathcal{P}$ can be decomposed into a nesting of $\eta-$predictable collections. By Definition 2, there exists predictor Φ such that for all $p \in \mathcal{P}$, Φ makes finitely many errors with probability 1. For $\eta > 0$, we define

$$\mathcal{P}^\eta_n = \{p \in \mathcal{P} \mid p(\Phi \text{ makes errors after time } n) < \eta\},$$

so that for all n, $\mathcal{P}^\eta_n$ is $\eta-$predictable by definition. Further, by definition, $\forall n \in \mathbb{N}^+$, $\mathcal{P}^\eta_n \subset \mathcal{P}^\eta_{n+1}$. To see that the union of $\mathcal{P}^\eta_n$ over all n is $\mathcal{P}$, consider the event

$$A_k = \left\{ X_1^\infty \mid \sum_{n=k}^{\infty} \ell(p, X_1^n, \Phi(X_1^{n-1})) > 0 \right\}.$$

For all $p \in \mathcal{P}$, have $p(A_k) \to 0$ as $k \to \infty$. Therefore, there must be some k such that $p(A_k) < \eta$, and for such a number k we have $p \in \mathcal{P}^\eta_k$. Therefore, $\mathcal{P} = \bigcup_{n\in\mathbb{N}} \mathcal{P}^\eta_n$.

To prove that the decomposition is sufficient, suppose that for all $j \in \mathbb{N}$, there exists a nesting $\left\{\mathcal{P}^j_n : n \geq 1\right\}$ of $\mathcal{P}$ such that $\mathcal{P}^j_n$ is 2^{-j} predictable. Furthermore, from Lemma 2, we can choose a decomposition such that $\mathcal{P}^j_n$ is 2^{-j} predictable with sample size n. Therefore, there exist predictors $\Phi_{n,j}$ such that for all $p \in \mathcal{P}^j_n$

$$p(\Phi_{n,j} \text{ makes errors after time } n) \leq 2^{-j}.$$

We construct a predictor Φ for $\mathcal{P}$ as follows. At each time step T, let $I(n, j)$ be the indicator that $\Phi_{n,j}$ makes no error on X_1^{T-1} after time n. Let

$$(k, i) = \underset{(n,j)\in\mathbb{N}\times\mathbb{N}}{\operatorname{argmin}} \{j + n \mid I(n, j) = 1\}. \tag{1}$$

The prediction is defined to be $\Phi(X_1^{T-1}) = \Phi_{k,i}(X_1^{T-1})$.

We claim that the predictor Φ will make only finitely many errors with probability 1 for all models in $\mathcal{P}$. Fix some $p \in \mathcal{P}$. Let $n_j = \min\{n \mid p \in \mathcal{P}^j_n\}$. Define the event

$$A_j = \{\Phi_{n_j,j} \text{ makes errors after time step } n_j\}.$$

We have $\sum_{j=1}^\infty p(A_j) \leq \sum_{j=1}^\infty 2^{-j} < \infty$.

Therefore, the Borel-Cantelli lemma implies that there is a set with probability 1, such that on every semi-infinite sequence X_1^∞ in that set, there is a J such that $\Phi_{n_J,J}$ makes no errors after step n_J. By construction of Φ, for X_1^∞ in the set of probability 1 above, we will therefore never choose an estimator $\Phi_{n,j}$ with $n + j > n_J + J$ in step (1). If some $\Phi_{n,j}$ with $n + j \leq n_J + J$ makes infinitely many errors, it will no longer appear in the feasible set in (1) after some time step $T \geq n$. Since there are only finitely many predictors $\Phi_{n,j}$ with $n + j \leq n_J + J$, the procedure will eventually choose some predictor that makes finitely errors. □

Remark 1. *As mentioned in the prelude to the theorem, one could view the the decomposition of the class in Theorem 1 as an regularization. The smaller $n + j$ is, the less the complexity the $\mathcal{P}_n^j$ has. Alternately, one may view the selection of the model classes in equation (1) as a structural risk minimization (Shalev-Shwartz and Ben-David, 2014, Chapter 7.2) that considers both the empirical losses and complexity of the class into account.*

The above characterization does not carry over for arbitrary loss functions if we cannot gauge the loss from the data. Instead, we prove the following analog.

Theorem 2. *Consider a collection $\mathcal{P}$ with a loss $\ell : \mathcal{P} \times \mathcal{X}^* \times \mathcal{Y} \to \{0, 1\}$ (i.e. the unsupervised setting). $(\mathcal{P}, \ell)$ is e.a.s.-predictable if there exists a nesting $\{\mathcal{P}_i : i \geq 1\}$ of $\mathcal{P}$ such that for all $\eta > 0$, $(\mathcal{P}_i, \ell)$ is η-predictable.*

Conversely, if $(\mathcal{P}, \ell)$ is e.a.s.-predictable, then for all $\eta > 0$, there is a nesting $\{\mathcal{P}_i^\eta : i \geq 1\}$ of $\mathcal{P}$ such that for all i, $(\mathcal{P}_i^\eta, \ell)$ is $\eta-$predictable.

Proof. The sufficiency follows directly from Lemma 1. The proof of the necessary condition is identical to the necessity proof of Theorem 1. □

The gap in the Theorem above cannot be closed. Indeed, the following example shows that the necessary (respectively sufficient) condition in Theorem 2 is not sufficient (respectively necessary).

Example 1. *Let $\mathcal{P}$ be a class of binary (taking values 0 or 1) random processes that converge to either 0 or 1 in probability. Formally, $\mathcal{P}$ is the collection of all probability measures p_b, $b \in \{0, 1\}$, defined on the Borel σ-algebra of $\{0, 1\}^\infty$, that satisfies*

$$\lim_{n\to\infty} p_b(X_n = b) = 1.$$

The task of the prediction Y_n is to predict the parameter b associated with the process, and takes values in $\{0, 1\}$. The loss ℓ associated with the prediction is defined to be

$$\ell(p_b, X_1^n, Y_n) = 1\{Y_n \neq b\}.$$

We now show that the condition deemed necessary in Theorem 2 holds for $(\mathcal{P}, \ell)$. To see this, let $\mathcal{P}_i^\eta$ be the class of processes $p_b \in \mathcal{P}$ such that for all $n \geq i$

$$p_b(X_n = b) \geq 1 - \eta.$$

The η-predictability of $\mathcal{P}_i^\eta$ follows because $p_b(X_i = b) \geq 1 - \eta$, and a predictor that predicts X_i for all time steps $\geq i$ will incur loss 0 past time step i whenever $X_i = b$.

We show that the collection $(\mathcal{P}, \ell)$ is, however, not *e.a.s.-predictable. To see this, suppose such a prediction rule Φ exists. We first observe that there exists a number $N(m)$ such that for all finite binary sequences $x_1, \cdots, x_m$ of length m and all $b \in \{0, 1\}$*

$$\Phi(x_1, \cdots, x_m, b, \cdots, b) = b, \tag{2}$$

whenever the number of b's is larger than $N(m)$. This holds because each of the $2 \cdot 2^m$ semi-infinite strings $x_1, \cdots, x_m, b, \ldots$ corresponds to a process in $\mathcal{P}$ that assigns probability 1 to that string. If (2) did not hold, Φ would make an infinite number of errors on one of these processes, contradicting the e.a.s.-predictability of Φ on $(\mathcal{P}, \ell)$.

We now construct the following process p_0 in $\mathcal{P}$ that will break Φ. Let $M_0 = 0$, $M_1 = N(0) + 1$, and recursively define $M_n = N(M_{n-1}) + 1$. The process p_0 is partitioned into independent sample blocks, where the nth block ranges from X_{M_n+1} to $X_{M_{n+1}}$ such that $X_{M_n+1} = X_{M_n+2} = \cdots = X_{M_{n+1}}$ and

$$p_0(X_{M_n+1} = 0) = 1 - \frac{1}{n}.$$

Let A_n be the event that $X_{M_n+1} = 1$. We have A_n happens infinity often almost surely by converse Borel-Cantelli lemma, since $\sum_n p_0(A_n) = \sum \frac{1}{n} = \infty$ and A_n's are independent. By construction, Φ makes errors in sample block n if A_n happens, hence Φ makes infinitely many errors almost surely. But clearly $p_0 \in \mathcal{P}$, contradicting the e.a.s.-predictability of Φ.

To see that the sufficient condition of Theorem 2 is not necessary, consider the processes we constructed above with M_n arbitrary and $p(X_{M_n+1} = b) = 1 - \frac{1}{n^2}$. The e.a.s.-predictability follows by Borel-Cantelli lemma but the sufficient condition of Theorem 2 does not hold for any decomposition of $\mathcal{P}$ (using a diagonalization argument).

Nevertheless, we can provide conditions that are both necessary and sufficient for e.a.s.-predictability for several natural unsupervised settings.

Theorem 3. *Let $\mathcal{P}$ be a model collection of* i.i.d. *measures over $\mathcal{X}^\infty$, ℓ is a loss function that only depends on the prediction and underlying source but not samples, i.e. $\ell : \mathcal{P} \times \mathcal{Y} \to \{0, 1\}$. If $|\mathcal{Y}|$ is finite, then $(\mathcal{P}, \ell)$ is e.a.s.-predictable iff there exists a nesting $\{\mathcal{P}_n\}_{n\geq 1}$ of $\mathcal{P}$ such that for all $\eta > 0$, $(\mathcal{P}_n, \ell)$ is η-predictable.*

Proof. Applying Theorem 2 to $(\mathcal{P}, \ell)$, we know that if the universal nesting exists, then $(\mathcal{P}, \ell)$ is e.a.s.-predictable. Theorem 2 also guarantees that if $(\mathcal{P}, \ell)$ is e.a.s.-predictable, then for all $\eta > 0$, there is a nesting $\{\mathcal{P}_i^\eta\}$ where $\mathcal{P}_i^\eta$ is $\eta-$predictable.

We prove the theorem by showing that if there exists some $\eta > 0$ for which there is a nesting $\{\mathcal{P}_i\}$ where $(\mathcal{P}_i, \ell)$ is η-predictable, then this nesting is also universal for all $\eta > 0$, *i.e.* $(\mathcal{P}_n, \ell)$ is also η-predictable for all $\eta > 0$.

To do so, we show that if $(\mathcal{P}_n, \ell)$ is $(\frac{1}{|\mathcal{Y}|} - \epsilon)$-predictable for any $\epsilon > 0$ then it is η-predictable for all $\eta > 0$. Suppose N be the sample size such that a predictor Φ makes no error past N with probability $\geq 1 - \frac{1}{|\mathcal{Y}|} + \epsilon$. For any $\eta < \frac{1}{|\mathcal{Y}|} - \epsilon$ let $M = \frac{2\log(\eta/2)}{\epsilon^2}$ and consider a sample of size MN. We split this sample into M blocks of size N each, and apply the predictor Φ to each block, obtaining a prediction $Y_i \in \mathcal{Y}$ for the i'th block. Our prediction for the sample of size MN is an element of $\mathcal{Y}$ that repeats most often in $Y_1, \cdots, Y_M$.

By Hoeffding bound with probability $\geq 1 - \eta$, any element of $\mathcal{Y}$ that incurs loss 1 against the underlying distribution will appear at most $\frac{1}{|\mathcal{Y}|} - \epsilon/2$ times among $Y_1, \ldots, Y_M$. But at least one element of $\mathcal{Y}$ appears more than $1/|\mathcal{Y}|$ among $Y_1, \ldots, Y_M$, and any such element must incur 0 loss. The corollary follows. □

We give another example, which illustrated how Theorem 2 can be used to derive the previously known results.

Example 2. *The task is to predict whether the parameter p of an* i.i.d. *Bernoulli(p) process is rational or not using samples from it.*

Therefore our predictor

$$\Phi : \{0,1\}^* \to \{ \textit{ rational }, \textit{ irrational } \}.$$

In (Cover, 1973), Cover showed a scheme that predicted accurately with only finitely many errors for all rational sources, and for a set of irrationals with Lebesgue measure 1. Here we show a more transparent version of Cover's proof as well as subsequent refinements in Koplowitz et al. (1995) using Theorem 2 above and an argument evocative of regularization.

Define the loss $\ell(p, X_1^n, Y_n) = 0$ iff Y_n matches the irrationality of p. Note that the setting is what we would call the "unsupervised" case and that there is no way to judge if our predictions thus far are right or wrong.

Let $r_1, r_2, \cdots$ be an enumeration of rational numbers in $[0,1]$. Let $B(p, \epsilon)$ be the set of numbers in $[0,1]$ whose ℓ_1 distance from p is $< \epsilon$. For all k, let

$$\mathcal{S}_k = \left([0,1] \backslash \bigcup_{i=1}^{\infty} B(r_i, \frac{1}{k2^i}) \right) \cup \{r_1, \cdots, r_k\}$$

be the set that excludes a ball centered on each rational number, but throws back in the first k rational numbers. Note that the Lebesgue measure of $\mathcal{S}_k$ is $1 - \frac{1}{k}$. Now $\mathcal{S}_k$ contains exactly k rational numbers, the rest being irrational. Moreover, S_k contains no irrational number within distance $\leq 2^{-k}/k$ from any of the included rationals. Hence, the set $\mathcal{B}_k$ of Bernoulli processes with parameters in $\mathcal{S}_k$ is η-predictable for all $\eta > 0$.

From Theorem 2, we can conclude that the collection $\mathcal{B} \stackrel{\text{def}}{=} \bigcup_{k\in\mathbb{N}} \mathcal{B}_k$ is e.a.s.-predictable. Note that every rational number belongs to $\mathcal{S} = \bigcup_{k\in\mathbb{N}} \mathcal{S}_k$, and the set of irrational numbers in $\mathcal{S}$ has Lebesgue measure 1, proving (Cover, 1973, Theorem 1).

Conversely, let $\mathcal{S} \subset [0,1]$ and $\mathcal{B}$ be the Bernoulli variables with parameters in $\mathcal{S}$. We show that if $\mathcal{B}$ is e.a.s.-predictable for rationality of the underlying parameter, then $\mathcal{S} = \bigcup_{k\in\mathbb{N}} \mathcal{S}_k$ such that

$$\inf\{|r-x| : r, x \in \mathcal{S}_k,\ r \textit{ is rational, } x \textit{ is irrational}\} > 0.$$

Since $\mathcal{B}$ is e.a.s.-predictable, Theorem 2 yields that for any $\eta > 0$, the collection $\mathcal{B}$ can be decomposed as $\mathcal{B} = \bigcup_k \mathcal{B}_k$ where each $\mathcal{B}_k$ is η-predictable and $\forall k$, $\mathcal{B}_k \subset \mathcal{B}_{k+1}$. Let $\mathcal{S}_k$ be the set of parameters of the sources in $\mathcal{B}_k$. Intuitively, $\eta-$predictability of $\mathcal{B}_k$ implies that we must have

$$\inf\{|u-v| : u, v \in \mathcal{S}_k,\ u \textit{ rational, } v \textit{ irrational}\} > 0,$$

or else we would not be able to universally attest to rationality with confidence $1-\eta$ using a bounded number of samples. See Supplementary Material for a formal proof.

Suppose we want $\mathcal{S}$ to contain all rational numbers in $[0,1]$. Then it follows (see supplementary material for a proof) that the subset of irrational numbers of $\mathcal{S}_k$ must be nowhere dense. Therefore, the set of irrationals in $\mathcal{S}$ is meager or Baire first category set (Rudin, 2006, Chapter 2.1), completing the result in (Koplowitz et al., 1995).

While Theorem 2 may look rather innocuous, it provides a partial resolution to an open problem in Dembo and Peres (1994). Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be disjoint classes of distributions over $\mathbb{R}^d$. Let $\mathcal{H}$ be the class of all *i.i.d.* random processes with marginal distributions from $\mathcal{H}_1 \cup \mathcal{H}_2$. Dembo and Peres (1994) considered the problem of identifying whether the marginal of a *i.i.d.*random process in $\mathcal{H}$ comes from $\mathcal{H}_1$ or $\mathcal{H}_2$ by observing samples from it. The prediction domain now is $\mathcal{Y} = \{1, 2\}$ and loss is $\ell(p, X_1^n, Y_n) = 1\{p \in \mathcal{H}_{Y_n}\}$. (Dembo and Peres, 1994) showed that if the distributions in $\mathcal{H}_1 \cup \mathcal{H}_2$ have a density *and* there exists some $q > 1$ such that the q-th norm of the density is finite, then $(\mathcal{H}, \ell)$ is e.a.s.-predictable iff the distributions in $\mathcal{H}_1$ and $\mathcal{H}_2$ are F_σ*-separable* (see Supplementary material for definition) under any metric consistent with

weak convergence topology. Dembo and Peres (1994) ask whether the condition $q > 1$ can be removed. We give a positive answer to this problem as follows.

Corollary 1. *Suppose there exists a monotonically increasing function $G : \mathbb{R}^+ \to \mathbb{R}^+$ with $\lim_{x\to\infty} G(x) = \infty$ such that for any distribution $p \in \mathcal{H}_1 \cup \mathcal{H}_2$ with density $f_p(x)$, we have $\mathbb{E}_{X\sim p}[G(f_P(X))] < \infty$. Then $(\mathcal{H}, \ell)$ is e.a.s.-predictable iff the distributions in $\mathcal{H}_1$ and $\mathcal{H}_2$ are F_σ separable under weak convergence topology.*

The proof and a discussion of the Corollary above is in the Supplementary material.

4 Capturing the final error

While *e.a.s.*−predictability is an attractive setup when considering rich model classes, we would like to see if an estimator that makes finitely many errors has finished making the errors. Namely, can we obtain a stopping rule that identifies the last error? Recall that a stopping rule is a function $\tau : \mathcal{X}^* \to \{0, 1\}$, such that $\tau(\mathrm{y}) \leq \tau(\mathrm{x})$ if y is a prefix of x. We interpreted $\tau = 0$ as the waiting period, and $\tau = 1$ after the rule has stopped waiting.

Definition 4. *A collection $(\mathcal{P}, \ell)$ is said to be eventually almost surely (e.a.s.)-learnable, if for any confidence $\eta > 0$, there exists a universal prediction rule Φ_η together with a stopping rule τ_η, such that for all $p \in \mathcal{P}$*

$$p\left(\sum_k^\infty \ell(p, X_1^k, \Phi_\eta(X_1^{k-1}))\tau(X_1^k) > 0\right) < \eta,$$

and

$$p\Big(\lim_{n\to\infty} \tau_\eta(X_1^n) = 1\Big) = 1.$$

Clearly, e.a.s.-learnability implies e.a.s.-predictabililty. However, the converse is not true, see (Wu and Santhanam, 2019) for an example. We leave the the proof of the theorem below to the supplementary material.

Theorem 4. *Any e.a.s.-learnable $(\mathcal{P}, \ell)$ is e.a.s.-predictable.*

To characterize e.a.s.-learnability, we define *identifiability* as follows.

Definition 5. *Let $\mathcal{U}$ be a collection of probability measures over $\mathcal{X}^\infty$, $\mathcal{V} \subset \mathcal{U}$. The class $\mathcal{V}$ is said to be identifiable in $\mathcal{U}$ if for any $\eta > 0$ there exist stopping rule τ_η, such that*

1. $p\left(\lim_{n\to\infty} \tau_\eta(X_1^n) = 1\right) = 1$ *for* $p \in \mathcal{V}$;
2. $p\left(\lim_{n\to\infty} \tau_\eta(X_1^n) = 1\right) \leq \eta$ *for* $p \in \mathcal{U}\backslash\mathcal{V}$.

Example 3. *Let $\mathcal{U}$ be the collection of all* i.i.d. *processes with marginal distributions over $[0, 1]$, and let $\mathcal{V} \subset \mathcal{U}$ be the set of distributions whose marginal mean is* not *equal to t for some fixed $t \in [0, 1]$.*

We show $\mathcal{V}$ is identifiable in $\mathcal{U}$. To see this, let $\epsilon_n = \frac{1}{n}$. Consider the following stopping rule. At stage n, we obtain a sample of size $\frac{2\log(2^{n+1}/\eta)}{\epsilon^2}$ and check whether the empirical mean is within ϵ_n distance of t. If not, we stop, else we continue to stage $n + 1$.

We show that this stopping rule identifies $\mathcal{V}$ in $\mathcal{U}$ using Definition 5. Suppose the underlying process has marginal mean equal to t. By Hoeffding bound, with probability at most $\eta/2^n$, the empirical mean will be outside distance ϵ_n to t. Therefore, the stopping rule stops with probability at most η by a union bound. If the marginal mean does not equal t, since $\epsilon_n \to 0$, the probability that the empirical mean will be within distance ϵ_n to t is at most $\frac{\eta}{2^n}$. By Borel-Cantelli lemma, this happens only finitely many times since $\sum \frac{\eta}{2^n} < \infty$, and the stopping rule stops almost surely.

We now provide the following characterization of *e.a.s.*-learnability.

Theorem 5. *Let $\mathcal{P}$ be collection of probability measures. Then $(\mathcal{P}, \ell)$ is e.a.s.-learnable if for all $\eta > 0$ there exists a countable decomposition $\mathcal{P} = \bigcup_{n\geq 1} \mathcal{P}_n^\eta$ such that for all n,*

1. *$(\mathcal{P}_n^\eta, \ell)$ is η-predictable,*
2. *$\mathcal{P}_n$ is identifiable in $\mathcal{P}$.*

Moreover, the condition is necessary if the measures in $\mathcal{P}$ are i.i.d. *over $\mathcal{X}^\infty$.*

Proof. We first show that the stated conditions on an collection $\mathcal{P}$ and loss ℓ are sufficient to guarantee that $\mathcal{P}$ is e.a.s.-learnable. Namely, for all $\eta > 0$, we will find an estimator Φ and a stopping rule τ such that for all $p \in \mathcal{P}$, the probability Φ incurs non-zero loss after τ stops is $\leq \eta$.

Now the conditions stated imply that for all $\eta > 0$, there is an identifiable nesting $\mathcal{P} = \bigcup_{n\in\mathbb{N}} \mathcal{P}_n$ and a sequence of numbers $\{m_n : n \geq 1\}$ such that each $\mathcal{P}_n$ is $\frac{\eta}{2}$-predictable with sample size m_n. Since $\mathcal{P}_n$ is identifiable, there is a stopping rule σ_n that stops after a finite time on $\mathcal{P}\backslash\mathcal{P}_n$ with probability at most $\eta/2^{n+1}$ and stops finitely almost surely on $\mathcal{P}_n$.

We will assume without loss of generality that σ_n only stops on sequences of length $\geq m_n$.

The stopping rule τ for $(\mathcal{P}, \ell)$ stops if for some n, σ_n has stopped. Let N be the smallest such number. The

prediction for $(\mathcal{P}, \ell)$ is now the $\eta/2-$predictor for $\mathcal{P}_N$, which we call Φ_n.

For all n, define

$$A_n = \{X_1^\infty : \sigma_n \text{ stops on } X_1^n\}.$$

We claim that:

1. The stopping rule τ stops with probabilty 1. This is because $p \in \mathcal{P}_k$ for some k, we have σ_k stops with probability 1.
2. The probability that τ stops but Φ incurs non-zero loss is $\leq \eta$. The probability that τ stops but $p \notin \mathcal{P}_N$ is $\leq \sum_{i=1}^\infty p(A_i) \leq \eta/2$. Finally, since $\mathcal{P}_N$ is $\eta/2-$predictable with sample size m_N, the probability that Φ_N predicts incorrectly after sample size m_N (which we are guaranteed is the case since we assumed σ_N only stops on sequences with length$\geq m_N$) is $\leq \eta/2$. The claim follows by union bound.

Now, suppose $(\mathcal{P}, \ell)$ is learnable, and $\mathcal{P}$ are *i.i.d.* measures on $\mathcal{X}^\infty$. For any $\eta > 0$, consider the stopping rule $\tau_{\eta/2}$ and the estimate $\Phi_{\eta/2}$ that learns $(\mathcal{P}, \ell)$. Let

$$T_n = \left\{X_1^\infty : \tau_{\eta/2}(X_1^n) = 1\right\}$$

be the set of sequences on which τ has stopped at or before step n. Now for all n, let

$$\mathcal{P}_n = \{p \in \mathcal{P} : p(T_n) > 1 - \eta/2\}.$$

Clearly, we have $\mathcal{P}_n \subset \mathcal{P}_{n+1}$ and that $\bigcup_{n\geq 1} \mathcal{P}_n = \mathcal{P}$. For all n, and for all $p \in \mathcal{P}_n$, $\Phi_{\eta/2}$ incurs non-zero loss on samples on length n with probability $\leq \eta/2$ by construction, namely, $\mathcal{P}_n$ is $\eta-$predictable by the union bound.

We will now show that $\mathcal{P}_n$ are identifiable in $\mathcal{P}$ to wrap up the theorem.

For any assigned confidence δ, we construct a stopping rule τ_δ that stops after a finite time with probability 1 when the underlying process is from $\mathcal{P}_n$ and with probability at most δ on processes from $\mathcal{P}\backslash\mathcal{P}_n$. To do so, we choose an arbitrary sequence $\{e_m\}$ with $e_m \to 0$ as $m \to \infty$. The stopping rule is partitioned into phases. At phase m we estimate $p(T_n)$ with confidence $1 - \frac{\delta}{2^m}$ and error bounded by $e_m/4$, by considering independent sample blocks of length n. If the estimate is larger than $1 - \eta/2 + e_m$ we stop, otherwise we continue to phase $m + 1$. Now, if we have $p(T_n) > 1 - \eta/2$, then there exists some number M such that $p(T_n) > 1 - \eta/2 + 2e_m$ for all $m \geq M$. Therefore, for all $m \geq M$, with probability at most $\delta/2^m$ we will not stop at phase m. By Borel-Cantelli lemma, with probability 1 we will stop in a finite time. A similar argument yields that if $p(T_n) \leq 1-\eta/2$, then we stop with probability at most δ. □

We now provide an example that shows that the condition in Theorem 5 is not necessary even for Markov processes with 2 states.

Example 4. *We consider the Markov process with state space* $\{0, 1\}$. *Let* $\mathcal{P}$ *be the class that contains the single state* 1 *process* p_0, *and processes* p_ϵ *with transition probability* $p_\epsilon(1|0) = p_\epsilon(0|1) = \epsilon$ *for all* $\epsilon \in (0, 1)$. *We assume the initial state of* p_ϵ *to be uniformly sampled from* $\{0, 1\}$.

We define the loss $\ell(p, X_1^n, Y_n) = 1$ *if* $X_1 = 1$ *or* $X_1 = 0$ *but* $\exists k \leq n$, $X_k = 1$, *the loss is* 1 *otherwise. Note that, the loss only depends on the samples* X_1^n *but independent of the prediction* Y_n. *Thus the prediction does not affect the loss.*

We now observe that the class is e.a.s.-learnable. One simply stops if the initial state is 1, *else we wait until we see* 1.

We now show that the decomposition of Theorem 5 does not exist for $(\mathcal{P}, \ell)$. *Suppose otherwise, we have a decomposition* $\{\mathcal{P}_n\}_{n\geq 1}$ *such that each* (P_n, ℓ) *is* $1/4$*-predictable. We know that for all* $n \geq 1$ *there exists a number* ϵ_n *such that for all* $p_\epsilon \in \mathcal{P}_n$ *we have* $\epsilon \geq \epsilon_n$. *Otherwise, we will not see state* 1 *in the sample before a bounded time step if the initial state is* 0 *(which happens w.p.* $1/2$*), thus violating the* $1/4$*-predictability of* $\mathcal{P}_n$. *We now assume* $p_0 \in \mathcal{P}_k$ *for some* k. *We show that* $\mathcal{P}_k$ *is not identifiable in* $\mathcal{P}$. *Taking the parameter* $\eta = 1/4$ *in Definition 5, we know that any* τ *must stops on the all* 1 *sequence at some point* N_0, *since* $p_0 \in \mathcal{P}_k$. *Now, taking any process* p_ϵ *that is not in* P_k *with* $\epsilon < \epsilon_k$ *and small enough so that* $(1 - \epsilon)^{N_0} \geq 3/4$, *we have* τ *stops on* p_ϵ *with probability at least* $3/8 > 1/4$, *contradicting identifiability.*

Note that the reason why construction of Example 4 is possible is because the number of states of the process is not known a-priori. Indeed, with arguments similar to the necessity part of Theorem 5 in the *i.i.d.* case, we can show that if $\mathcal{P}$ is a collection of irreducible finite Markov processes with fixed number of states then the necessary condition in Theorem 5 still holds. We should also emphasize that the stopping rule derived from the sufficient condition of Theorem 5 is not meant to be optimal. For specific problems, there will often be more natural stopping rules.

5 Applications

We now provide some interesting concrete applications of the *e.a.s.*-predictability framework. To begin with,

we first prove the following theorem, which is a direct corollary of Theorem 2.

Theorem 6. *Let $\mathcal{P}$ the set of all* i.i.d. *processes with marginal distributions over $[0,1]^d$ for some $d \geq 1$. For all $A \subset [0,1]^d$, we define loss $\ell_A(p, X_1^n, Y_n) = 1\{1\{\mathbb{E}_{X\sim p}[X] \in A\} \neq Y_n\}$, where the prediction Y_n tries to decide whether $\mathbb{E}_{X\sim p}[X] \in A$ or not. We have $(\mathcal{P}, \ell_A)$ is e.a.s.-predictable if A is closed in $[0,1]^d$.*

Proof. Let $C_m = \{\mathbf{x} \in [0,1]^d : d(\mathbf{x}, A) \geq \frac{1}{m}\}$. We have $\bigcup_{m\geq 1} C_m = [0,1]^d \backslash A$, since A is closed. Let $\mathcal{P}_A$ be the processes in $\mathcal{P}$ with marginal mean in A and $\mathcal{P}_m$ be the processes with marginal mean in C_m. Define $\mathcal{P}'_m = \mathcal{P}_A \cup \mathcal{P}_m$. We have $\mathcal{P}'_m \subset \mathcal{P}'_{m+1}$ and $\mathcal{P} = \bigcup_{m\geq 1} \mathcal{P}'_m$. To apply Theorem 2, we have to show that $(\mathcal{P}'_m, \ell_A)$ is η-predictable for all $\eta > 0$. This can be easily achieved by simply checking whether the empirical mean is close to A or C_m. Since $d(A, C_m) \geq \frac{1}{m}$, one will be able to find bounded sample size so that we will make the right decision with arbitrary high confidence. □

For any function $f : [0,1]^d \to \{0,1\}$, we will be able identify a set $A_f = \{\mathbf{x} \in [0,1]^d : f(\mathbf{x}) = 1\}$. Let $A_1, \cdots, A_n \subset [0,1]^d$ be finitely many sets such that $(\mathcal{P}, \ell_{A_i})$ is e.a.s.-predictable for all $1 \leq i \leq n$. Let $g : [0,1]^d \to \{0,1\}$ be an arbitrary function, denote $f(\mathbf{x}) = g(1_{A_1}(\mathbf{x}), \cdots, 1_{A_n}(\mathbf{x}))$. It is easy to show that $(\mathcal{P}, \ell_{A_f})$ is also e.a.s.-predictable.

We now consider the following problem setup. Let $\mathbf{X}$ be a $d \times d$ random matrix such that each entry $\mathbf{X}(i,j)$ is a Bernoulli random variable. We denote $\mathbb{E}[\mathbf{X}]$ to be a (deterministic) matrix that takes expectation entry-wise on $\mathbf{X}$. Let $\mathbf{X}_1, \mathbf{X}_2, \cdots$ be *i.i.d.* realization of $\mathbf{X}$, which are binary matrices. We will try to identify the properties of $\mathbb{E}[\mathbf{X}]$ by observing the samples $\mathbf{X}_1, \mathbf{X}_2, \cdots$. Clearly, we can associate properties of $\mathbb{E}[\mathbf{X}]$ with subsets of $[0,1]^{d\times d}$. We will denote $\mathcal{P}$ to be the class of all *i.i.d.* Bernoulli random matrices process. We say a property of $\mathbb{E}[\mathbf{X}]$ is e.a.s.-predictable if $(\mathcal{P}, \ell_A)$ is e.a.s.-predictable where $A \subset [0,1]^{d\times d}$ is the subset corresponding to the property.

Theorem 7. *The singularity of $\mathbb{E}[\boldsymbol{X}]$ is e.a.s.-predictable.*

Proof. Note that the determinant is a continuous function w.r.t. the entries of the matrix. Thus subsets in $[0,1]^{d\times d}$ corresponding to the determinant being 0 is closed. The theorem follows by Theorem 6. □

Theorem 8. *To determine if $\mathbb{E}[\boldsymbol{X}]$ has rank k is e.a.s.-predictable for all k.*

Proof. Note that to check whether a matrix has rank k, one only need to check the maximum non-singular square submatrix is of dimension k. Thus the property can be expressed as function of finite singularity test. By Theorem 7 we know that the property is still e.a.s.-predictable. □

Note that the realizations $\mathbf{X}_1, \cdots, \mathbf{X}_n$ will have full rank with high probability even when the entries are Bernoulli($\frac{1}{2}$) (i.e. $\mathbb{E}[X]$ has rank 1). Theorem 8 shows that one still be able to infer the rank of $\mathbb{E}[\mathbf{X}]$ from $\mathbf{X}_1^n$.

Theorem 9. *To determine if $\mathbb{E}[\boldsymbol{X}]$ has eigenvalue of multiplicity more than 1 is e.a.s.-predictable.*

Proof. For any matrix A, consider the characteristic polynomial p_A of A. We know that the coefficients of p_A are polynomials of the entries of A. We now only need to check if $\mathrm{GCD}(p_A, p'_A) = 1$, where p'_A is the derivative of p_A. Note that this can be done by checking the resultant of p_A, p'_A is zero. Since resultant is continuous functions of the coefficients, the theorem follows using Theorem 6. □

While one should expect most of properties of matrix to be e.a.s.-predictable, we have the following open problem.

Problem 1 (Open Problem)**.** *Is determining whether a matrix is diagonalizable e.a.s.-predictable?*

6 Discussion

At first glance, it may not be clear what guaranteeing finitely many errors over an infinite horizon implies for finite sized samples. But, as we discussed in Section 3, the e.a.s.-predictability implies that one would be able to decompose the class into uniformly predictable subclasses. This provides us a natural regularization when dealing with finite sized samples. Depending on the sample we have at hand, we would only work with the subclass of sample complexity that matches our sample size. As we obtain more and more samples, we will then loosen the restrictions. e.a.s.-predictability guarantees that such a strategy would eventually succeed for any model in the class. In many cases, e.g. supervised setting, e.a.s.-predictable classes are the only situation where we can hope for a consistent regularization that leads to eventual success in the worst case.

The more restricted (yet more practical) e.a.s.-learnability setup, provides a generalization to the classical uniform sample complexity guarantees. Rather than using the sample size as a means to determine convergence properties, e.a.s.-learnability can be interpreted as allowing arbitrary functions of the sample (not just the sample size) to determine convergence properties.

Acknowledgments

This work was supported by NSF grants CCF-1619452 and by the Center or Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370.

References

Thomas M Cover. On determining the irrationality of the mean of a random variable. *Annals of Statistics*, 1(5):862–871, 1973.

Amir Dembo and Yuval Peres. A topological criterion for hypothesis testing. *Annals of Statistics*, pages 106–117, 1994.

E Mark Gold. Language identification in the limit. *Information and control*, 10(5):447–474, 1967.

Jack Koplowitz, Jeffrey E Steif, and Olle Nerman. On cover's consistent estimator. *Scandinavian Journal of Statistics*, pages 395–397, 1995.

Sanjeev R Kulkarni and David N. C. Tse. A paradigm for class identification problems. *IEEE Transactions on Information Theory*, 40(3):696–705, 1994.

Michael Naaman. Almost sure hypothesis testing and a resolution of the jeffreys-lindley paradox. *Electronic Journal of Statistics*, 10(1):1526–1550, 2016.

W. Rudin. *Functional Analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill, 2006. ISBN 9780070619883. URL `https://books.google.com/books?id=l7XFfDmjp5IC`.

Narayana Santhanam and Venkat Anantharam. Agnostic insurability of model classes. *Journal of Machine Learning Research*, 16:2329–2355, 2015. URL `http://jmlr.org/papers/v16/santhanam15a.html`.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Changlong Wu and Narayana Santhanam. Being correct eventually almost surely. In *Information Theory (ISIT), 2019 IEEE International Symposium on*, pages 1989–1993. IEEE, 2019.

Thomas Zeugmann and Sandra Zilles. Learning recursive functions: A survey. *Theoretical Computer Science*, 397(1-3):4–56, 2008.