Online Learning and Prediction with Eventual Almost Sure Guarantees

A DISSERTATION SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERSITY OF HAWAIʻI IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

ELECTRICAL ENGINEERING

August 2021

By
Changlong Wu

Dissertation Committee:

Narayana Santhanam, Chairperson
Anders Host-Madsen
Anthony Kuh
June Zhang
Bjoern Kjos-Hanssen

To my parents

Acknowledgements

I would like to thank my advisor Prof. Narayana Prasad Santhanam for introducing me the beautiful problems in the finite error prediction paradigm. I'm also grateful for his support and encouragement in my whole Ph.D period both professionally and personally.

I would also like to thank Prof. Anders Host-Madsen, Prof. Anthony Kuh, Prof. June Zhang and Prof. Bjoern Kjos-Hanssen for serving as my committee members. Special thanks to Prof. Kjos-Hanssen for carefully reading an earlier version of this manuscript and providing valuable feedback.

I also wish to thank my labmates and collaborators: Wenxin, Kevin, Ian and Maryam.

Abstract

In this dissertation, we introduce a general prediction paradigm where a learner predicts properties of an underlying model and of future samples from the model using past observations from the model. The prediction game continues for infinite steps in an online fashion as the sample size grows with new observations (*not* necessarily *i.i.d.*). 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 of a learner is to make only finitely many errors (i.e. loss 1) with probability 1 under the generating model, no matter what the underlying model may be in the known model class.

Any model class and loss pair that admit predictors that make finitely many errors in the above fashion is called *eventually almost surely* (or $e.a.s.$ for abbreviation) predictable. Our main contributions of this dissertation are general characterizations for the $e.a.s.$-predictable class and loss pairs. Using our general characterizations, we establish tight necessary and sufficient conditions for a wide range of prediction problems to be $e.a.s.$-predictable, which include hypothesis testing, online learning and risk management theory. Moreover, our results establish striking connections between the $e.a.s.$-predictability and the notion of regularization.

While $e.a.s.$-predictable classes admit predictors with only finitely many errors, where we made the final error may yet remain unknown. In particular, we say a class and loss pair to be $e.a.s.$-learnable if it is $e.a.s.$-predictable and, in addition, we are able to identify the last error with any given confidence using a universal stopping rule. We provide general characterizations for the $e.a.s.$-learnability, which is tight in many natural settings.

While the above results bring out broad principles, to bring about a more refined development of the framework, we study three broad categories of applications in our framework: hypothesis testing, online classification and learning, and risk prediction.

Our characterization of hypothesis testing problems includes testing general properties of distributions using *i.i.d.* samples, including testing entropy properties of discrete distributions and testing properties of random matrices with Bernoulli entries. Our general results in the $e.a.s.$-prediction framework also strengthen and extend prior results in Dembo and Peres (1994) with simple and elementary proofs and provide a partial resolution to an open problem posed therein.

In our approach to online classification, a classifier obtains the training data in an online fashion and predicts labels on the next instance, but is required to only make *finitely* many errors over an infinite horizon. Extending the classical bounded error scenario by Littlestone (1988), we show that a binary labeled hypothesis class can be learned online with finitely many errors almost surely using *i.i.d.* samples from a given distribution μ iff the class is *effectively* countable w.r.t. that distribution μ. We also characterize the setting where μ is unknown and show that corrupting the labels by independent Bernoulli(η) noise does not change learnability so long as $\eta < 1/2$. We extend our results to the case where class labels need not be binary. Going past prior results on learning recursive functions in Zeugmann and Zilles (2008), we show that the class of all binary valued computable functions on naturals can be online learned with finitely many errors almost surely by a computable predictor given samples from certain non-degenerated distributions. Next, we bound the computational complexity of the predictor, and study classes of functions that can be computed in exponential and polynomial time respectively.

Lastly, we study the problem of predicting upper bounds on the next draw of an unknown probability distribution after observing a sample generated by it. We show that a prediction rule exists that violates the bounds only finitely many often almost surely if and only if a class can be decomposed into countable union of tight classes. This implies, e.g., the class of all monotone distributions can not be predicted in such a sense.

Table of Contents

Chapter 1

Motivations

Regularization is a common practitioner's approach that resolves ill-posed problems by adding additional constraints (Tikhonov 1943; Cortes and Vapnik 1995; Tibshirani 1996; Mosci et al. 2010; Bühlmann and Van De Geer 2011; Hinton et al. 2012). In practice, regularization is sometimes chosen by trial and error, with the level of regularization chosen by empirical methods such as cross validation. Some theoretical tools do consider a more disciplined approach, matching the level of regularization with the data at hand (Rothman et al. 2008; Banerjee et al. 2008; Mazumder and Hastie 2012; Blanchet et al. 2019), but in more typical cases, theoretical guarantees are absent and the level of regularization is usually tuned ad-hoc.

The main theme of this dissertation will be a framework that places the selection of regularization in sequential prediction problems in a statistical framework that studies estimators making *finitely* many errors *almost surely* (Grimmett and Stirzaker 2020, Chapter 7.3) (see also Chapter 3 for a short summary of these terms). We seek to understand sequential prediction problems admitting estimators that make only finitely many mistakes almost surely, and end up opening a new view into regularization. To illustrate the ideas, we start with the following examples.

1.1 Is a coin biased?

Suppose we have a coin that is modeled as a Bernoulli random variable with parameter $p \in [0,1]$. Our goal is to decide whether the coin is biased or not, i.e., is $p = \frac{1}{2}$ or not by tossing the coin independently. We can refine our decision after every new observation, but demand that we make *finitely* many wrong decisions (therefore, after a finite time, we are always correct) almost surely, no matter what model is in force.

At first glance, this seems to be an ill-posed problem, since with any finite sample of size n we will not be able to distinguish between $p = \frac{1}{2}$ and $p = \frac{1}{2} \pm o(\frac{1}{\sqrt{n}})$. Clearly, the problem will not be solvable uniformly over sources in $[0,1]$ with any finite sample size. The missing ingredient we need here is *regularization*. Suppose we know *in advance* that either $p = \frac{1}{2}$ or $|p - \frac{1}{2}| \geq \epsilon$ for some given $\epsilon > 0$. Such a problem is easily decidable w.h.p. by observing a sample of size $O(\frac{1}{\epsilon^2})$.

What if such an ϵ was not known in advance? To resolve this issue, for any $k \geq 1$, we define the following regularized problem: let

$$\mathcal{P}_k = \left\{ p \in [0,1] : p = \frac{1}{2} \text{ or } |p - \frac{1}{2}| \geq \frac{1}{k} \right\},$$

and for sources in $\mathcal{P}_k$, decide if the coin is biased or not. Note that, $\mathcal{P}_k \subset \mathcal{P}_{k+1}$ for all $k \geq 1$ and the union $\bigcup_{k \geq 1} \mathcal{P}_k$ includes all parameters in $[0,1]$. Moreover, each of the class $\mathcal{P}_k$ is uniformly decidable with confidence $\geq 1 - \frac{1}{k^2}$ by observing $b_k = O(k^2 \log k)$ samples. To do so, we simply decide "biased" if $|\bar{X}_{b_k} - \frac{1}{2}| \geq \frac{1}{2k}$ and decide "unbiased" otherwise, where $\bar{X}_{b_k}$ is the empirical mean of a sample of size b_k.

To eliminate the dependency on the parameter k, we use the following decision rule. Partition the decisions into phases. We are in phase k if the sample size we observed satisfies $b_k \leq n < b_{k+1}$. At phase k, we use the decision rule for $\mathcal{P}_k$ with sample size b_k to make the decision, and we *retain* the same decision the *whole phase*.

Now, for any $p \in \bigcup_{k \geq 1} \mathcal{P}_k$, if $p \neq \frac{1}{2}$ there must be some k such that $p \in \mathcal{P}_j$ for all $j \geq k$. By the Borel-Cantelli lemma, the number of errors of the decisions given by the rule after

phase k will be finite almost surely, since $\frac{1}{k^2}$ is summable. In other words, the decision will be correct eventually almost surely. If $p = \frac{1}{2}$, we know that $p \in \mathcal{P}_k$ for all $k \geq 1$. The decision will be correct after finitely samples almost surely by the Borel-Cantelli lemma again.

1.2 Is the bias of a coin rational?

We now consider a more sophisticated example introduced by Cover (1973). In this example, our goal is to decide whether the bias of a coin is rational or irrational, i.e., is $p \in \mathbb{Q}$ or not. Note that the set of rational numbers $\mathbb{Q}$ is dense in $[0, 1]$. It will not be possible to decompose all the parameters in $[0, 1]$ into countably many classes that have bounded gaps between rationals and irrationals. Perhaps surprisingly, Cover (1973) showed that it is possible to determine the rationality with *finitely* many errors almost surely for all sources in a subset $\mathcal{S} \subset [0, 1]$ that contains all rational numbers and has Lebesgue measure 1 using the law of iterated logarithm.

We now provide a *regularization* point of view of Cover's result as we did in our previous example. 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 $\geq 1 - \frac{1}{k}$. 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 rationality of parameters in $\mathcal{S}_k$ will be *uniformly* decidable with any given confidence using *bounded* number of samples. Let Φ_k be the decision rule for $\mathcal{S}_k$ that achieves confidence $1 - \frac{1}{k^2}$ using a sample of size b_k. Define $\mathcal{S} = \bigcup_{k \geq 1} \mathcal{S}_k$, we have $\mathcal{S}$ contains all rational numbers and the Lebesgue measure of $\mathcal{S}$ is 1. We now show that the sources in $\mathcal{S}$ can be decided with finitely

many errors almost surely. We partition the decisions into phase, we are in phase k if the sample size satisfies $b_k \leq n < b_{k+1}$. At phase k, we use Φ_k to make the decision at the beginning of the phase and retain the same decision the whole phase. Now, for any $p \in \mathcal{S}$, we have $p \in \mathcal{S}_k$ for some k. By the Borel-Cantelli lemma, the rule will make finitely many errors after phase k. This recovers the main result of Cover (1973).

Conversely, we can also show that for any set $\mathcal{S} \subset [0,1]$, if we would like to decide the rationality of all sources in $\mathcal{S}$ with finitely samples almost surely then a *regularization* will be necessary. This recovers the main result of Koplowitz et al. (1995). See Example 4 in Section 4.2 for more formal statements.

More generally, we will show in Chapter 4 that in many natural settings a problem is solvable with finitely many errors almost surely if and only if a decomposition of the problem into uniformly solvable regularized sub-problems is possible. Note that, the idea of aggregating sub-problems by increasing class complexity according to the confidence level dates back to the work of Linial et al. (1991). We should emphasize that the more crucial part of applying such an approach is in defining the sub-problems, which will be the major problem we will address in this dissertation.

1.3 Model complexity and accuracy

One might have noticed that the approach we introduced in the previous sections relies heavily on the assumption that we are able to obtain arbitrary confidence with a *bounded* sample size. In this section, we introduce a different way of aggregating sub-problems by trading-off the model complexity and errors when the losses are observable from the sample, along the lines of what is commonly called as *structural risk minimization* (SRM) (Vapnik and Chervonenkis 1974).

We illustrate the idea of introducing a regularizer for complexity using the problem of online learning. We consider this problem in more detail in Chapter 6, but it is instructive to consider a simple variant of this problem in the structural risk minimization setting.

Let $\mathcal{H}$ be a binary hypothesis class with functions that map from $\mathcal{X} \to \{0, 1\}$. We now consider the following online learning game. At the beginning, Nature chooses some $h \in \mathcal{H}$. At each time step n, Nature selects some $x_n \in \mathcal{X}$ and provides it to a learner. The learner then provides a prediction y_n for $h(x_n)$ potentially using the history he observed thus far. Nature reveals $h(x_n)$ after the prediction has been made. The goal of the learner is to make finitely many errors no matter what the function is chosen by Nature at the beginning and no matter how the instances $\{x_1, x_2, \cdots\}$ are selected by Nature.

Now, suppose we have countably many classes $\mathcal{H}_1, \mathcal{H}_2, \cdots$ such that each of the classes $\mathcal{H}_k$ admits a predictor that makes finitely many errors. Consider predicting when all we know is that the hypothesis is in the union $\mathcal{H} = \bigcup_{k \geq 1} \mathcal{H}_k$. Even though each of the classes admits a predictor making finitely many errors, we do not have a uniform bound on the number of errors. Nor do we know which of the subclasses the hypothesis belongs to. Therefore we do not know *a-priori* that $\mathcal{H}$ is also predictable with finitely many errors.

Consider the following strategy. Let Φ_k be the finite error predictor for $\mathcal{H}_k$. Define

$$\hat{k} = \arg\min_{k \geq 1} \{k + \text{err}_S(\Phi_k)\},$$

where $\text{err}_S(\Phi_k)$ is the number of errors (distinct labels) of Φ_k on current samples S. We then use $\Phi_{\hat{k}}$ to make the prediction for the next sample. We can show that such a rule will indeed make finitely many errors for the class $\mathcal{H}$, see Lemma 15 in Chapter 6 for more details of the proof.

1.3.1 Structural Risk Minimization

Consider the SRM setting for the well studied scenario of learning hypotheses from a countable union of finite VC dimension classes. This is a case of non-uniform learning, where the generalization error of the learned hypothesis converges to the minimum generalization error from the union, albeit at a rate that is not uniform over the full union. We contrast this with our prediction formulation to bring out a few distinctions.

Formally, suppose we have countably finite VC-dimension classes $\{\mathcal{H}_k\}_{k\in\mathbb{N}}$ and $\mathcal{H} = \bigcup_{k\in\mathbb{N}} \mathcal{H}_k$. For each class $\mathcal{H}_k$ and number $n \in \mathbb{N}$, we denote $\epsilon_k(n,\delta)$ to be a *uniform* bound on the difference between the generalization error and empirical error for all $h \in \mathcal{H}_k$ with confidence $1 - w(k)\delta$ using a sample of size n. Here, we assume $w(k) \in [0,1]$ and $\sum_{k\in\mathbb{N}} w(k) = 1$. Clearly, $\epsilon_k(n,\delta) \to 0$ as $n \to \infty$, since $\mathcal{H}_k$ has finite VC-dimension. Now, the SRM rule suggests selecting a model in the following way

$$\hat{h} = \arg\min_{h\in\mathcal{H}}\{L_S(h) + \epsilon_{k(h)}(n,\delta)\},$$

where $L_S(h)$ is the empirical error of h on S and $k(h)$ is the minimum index such that $h \in \mathcal{H}_{k(h)}$. It can be shown that, with the SRM rule, for any $h \in \mathcal{H}$, with confidence $\geq 1 - \delta$, the generalization error of $\hat{h}$ will converge to the minimal generalization error of $\mathcal{H}$ on the generating distribution (Shalev-Shwartz and Ben-David 2014, Theorem 7.5).

Even though for any $h \in \mathcal{H}$ the SRM will output a hypothesis with generalization error that goes to zero in the realizable case, it does not necessarily provide a finite error guarantee if we use the learned hypothesis $\hat{h}$ to *predict* for the next label. Indeed, as we will see in the following section, even a class with VC-dimension 1 can have no prediction rule that achieves a finite error guarantee. However, we can show that in many general settings, e.g., the supervised setting that we will introduce in Chapter 4, the SRM rule will provide a natural way for deriving finite error guarantees if we choose the regularization term that is suggested by our theory. See Theorem 5 in Section 4.1 for more details.

1.4 More examples

We now provide more examples below in different contexts to demonstrate the ideas we will develop in this dissertation.

Testing properties of first moment: Let p be a distribution over $\mathbb{N}$, and $X_1, X_2, \cdots$ be *i.i.d.* samples from p. Assume we have $\mathbb{E}[X_1] < \infty$. Our goal is to decide whether the

first moment $\mathbb{E}[X_1] \leq 100$ or not by observing the independent samples $X_1, X_2, \cdots$. We now consider a scenario where we are allowed to update our decision every time we observe a new sample. The problem is that, can we eventually stick on the right decision after some finite observations with probability 1? or will we be perpetually switching between decisions without even convergence to one?

Note that, here we only assumed that the first moment is finite, we do not have any assumptions on the higher moments. Therefore, the approach of constructing gaped sub-problems as in the Section 1.1 will not work here, since we will not be able to achieve uniform confidence with bounded samples. However, we will show in Section 5.2.1 that there exists a decision rule that makes the right decision eventually almost surely no matter what the distribution p may be.

The similar problem, i.e., deciding whether $\mathbb{E}[X_1] \geq 100$ or not, is, however, not decidable with finitely many errors almost surely! This follows by showing that the class of all distributions over $\mathbb{N}$ *cannot* be decomposed into subclasses such that we can decide if $\mathbb{E}[X_1] \geq 100$ or not uniformly for all subclasses. See Section 5.2.1 for more details on the proof.

Online learning of linear threshold functions: Let $\mathcal{H}$ be the class of linear threshold functions over $[0, 1]$, i.e., for any $h \in \mathcal{H}$ there exist some constant $a \in [0, 1]$ such that

$$\forall x \in [0, 1],\ h(x) = 1\{x \geq a\}.$$

We now consider the following randomized online learning scenario. Let $h \in \mathcal{H}$ be some underlying function which is unknown to the learner. At each time step n, Nature will generate a sample x_n from the uniform distribution over $[0, 1]$. The goal of the learner at step n is to predict $h(x_n)$ potentially using the history he observed thus far. Nature reveals the true label after the prediction has been made. The problem is that, can we make only finitely many errors (i.e. wrong predictions) in the infinite horizon with probability 1?

We will show in Section 6.3 that this is actually not possible, though the class $\mathcal{H}$ has VC-dimension 1. Note that, by the VC-theorem (Shalev-Shwartz and Ben-David 2014, Theorem 6.8), the Empirical Risk Minimization (ERM) rule will achieve expected error at step n to be upper bounded by $O(\frac{\log n}{n})$. Our impossibility result implies that such a bound cannot be improved to $O(\frac{1}{n \log^{1+\epsilon} n})$ for all $\epsilon > 0$ and for any learning rule. This holds even when the hidden constant of the big-O notation depends on the underlying function.

On the other hand, if we consider the class of linear threshold functions with only *rational* parameters, then an almost sure finite error guarantee will be possible, even though the rational numbers are dense in $[0, 1]$. This follows from the fact that the rational numbers are countable. Indeed, as we will see in Chapter 6 that an *effective* countability condition is both necessary and sufficient to achieve finite error guarantees for online learning problems of binary labels.

Estimating rank of random matrices: Let $\mathbf{X}$ be a random matrix where each entry $\mathbf{X}_{i,j}$ of $\mathbf{X}$ is a Bernoulli random variable with parameters $p_{i,j}$. We denote $\mathbb{E}[\mathbf{X}]$ to be the (deterministic) matrix with entries of $p_{i,j}$. The goal is to estimate the rank of $\mathbb{E}[\mathbf{X}]$ by observing independent realizations $X_1, X_2, \cdots$ (which are binary matrices) of $\mathbf{X}$. The estimation can be updated every time we observe new realizations. The problem is that, can we make the right estimation of the rank after finitely many observations?

Clearly, we can't use the rank of the empirical mean $\bar{X} = \frac{X_1 + \cdots + X_n}{n}$ as the estimation, since w.h.p. $\bar{X}$ will be full rank even if $\mathbb{E}[\mathbf{X}]$ is rank 1. We will show in Section 5.3 that we are able to make the right estimation after finitely many observations almost surely, even though the rank is quite sensitive to the estimations of each entry of $\mathbb{E}[\mathbf{X}]$. This follows from a general topological criterion of hypothesis testing problems as we will discuss in detail in Chapter 5.

Chapter 2

Introduction

Statistical inference in the sequential manner is a common scenario that balances the *exploration-exploitation* trade-off of learning and prediction. This is widely studied in many different settings of estimation and learning problems, including sequential hypothesis testing (Wald 1945; Kiefer and Weiss 1957; Hoeffding 1960), universal prediction (Merhav and Feder 1998; Singer and Feder 1999; Cesa-Bianchi and Lugosi 1999) and online learning (Littlestone 1988; Haussler et al. 1994; Ben-David et al. 2009). A typical generic setup assumes an underlying probabilistic model that generates samples at discrete time steps, where the model is unknown to a *learner*. The learner could observe the samples at different time steps. Meanwhile, the learner is required to take causal actions at each time step, which could be, e.g., a prediction of properties of the next samples or an estimation of properties of the underlying generating model. The actions made by the learner could depend on the observation of the past samples and his prior knowledge of the generating model and is scored by prespecified loss functions. The goal of the learner is to minimize the cumulative loss incurred in the process over either a finite or infinite horizon.

Common results in prior literature are either bounding the exact losses made by the learner or showing that the losses are not much worse compared with reference learners that may have additional knowledge (Shalev-Shwartz 2011). However, a uniform bound on the losses often requires strong prior assumptions on the generating models, e.g., in the online classification of binary functions we will need a bounded Littlestone dimension (Littlestone

1988). Such assumptions are often still required even considering competitive setups, if a uniform guarantee is desired (Ben-David et al. 2009). This may not be realistic if we have little prior knowledge of the underlying model, e.g., learning of a universal physical law from measurements of the environment.

In this dissertation, we relax the uniform consistency narrative to include more relaxed non-uniform consistencies, i.e., guarantees that could depend on the underlying models. In particular, we are interested in whether a sequential prediction problem could admit a predictor that makes *finitely* many errors, instead of bounding the number of errors uniformly. This approach is related to multiple lines of research in the literature, two of which we outline below.

The first is the learning of *recursive functions* (a.k.a. inductive inference) that was introduced in Gold (1967). Here one assumes the generating model to be a *deterministic* binary computable sequence $\mathbf{x} \in \{0,1\}^\infty$, i.e., there exists a computable function h such that $h(n) = \mathbf{x}_n$ for all $n \in \mathbb{N}$ where $\mathbf{x}_n$ is the nth bit of $\mathbf{x}$. The strategy of the learner is restricted to be computable as well. A natural problem of this setup is to predict the next bit by observing a prefix of $\mathbf{x}$. It is shown by Barzdiņš and Freivald (1972) that a finite error computable predictor exists if and only if there exists a computable function $g : \mathbb{N} \to \mathbb{N}$ such that the time complexity of h that computes $\mathbf{x}$ is *eventually dominated* by g. Note that this encompasses a very broad range of complexity classes, e.g., when $\mathbf{x}$ is polynomial time computable. We refer the reader to the survey by Zeugmann and Zilles (2008) for more details on this topic.

The second line of research involves *randomized* observations, and one starting point for questions like this is Cover (1973). In Cover's paper, 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 continued in (Kulkarni and Tse 1994; Koplowitz et al. 1995; Leshem 2006; 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 in (Santhanam and Anantharam 2015), the authors obtain a stopping rule that additionally indicates when such a prediction has made the last mistake.

The main contribution of this dissertation is a general framework that subsumes the problem setups mentioned above into a unified analysis framework. This allows us to tackle a wide range of sequential prediction problems of large scale. Informally, we consider the following general prediction setup. The 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 the beginning of the game. 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 said to be *eventually almost surely* (or *e.a.s.*) predictable, if there is a strategy such that the learner makes only *finitely* many errors with probability 1 no matter what the underlying $p \in \mathcal{P}$ is. Moreover, we say a collection $(\mathcal{P}, \ell)$ is *e.a.s.-learnable* if, in addition, we are able to identify the last error with any given confidence using a universal stopping rule.

2.1 Implications on related research directions

Our framework of finitely many errors connects with some well studied research directions. It is worthwhile to examine the implications of our work on these well studied areas. In doing so, both the scope of our work as well as its breadth and context will become more explicit.

In many statistical problems, we are often not able to obtain guarantees that hold uniformly over all models in a class, due to the lack of prior knowledge of the model and the complexity of the model class. Non-uniform, or pointwise consistency is a paradigm that allows guarantees to be dependent on the underlying models. This is widely investigated in many research communities, including, learning theory, information theory and statistics. Of these, we will survey some of the major lines of research that are related to this dissertation.

2.1.1 Non-uniform PAC learning

Perhaps the best known non-uniform consistency setup in the learning theory literature is the non-uniform PAC learning (Blumer et al. 1989; Benedek and Itai 1994) and the concept of *Structural Risk Minimization* (SRM) (Vapnik and Chervonenkis 1974; Vapnik 2013). In this setup, one aims to obtain a *distribution-free* sample complexity that could depend on the underlying hypothesis. It is shown by Linial et al. (1991); Benedek and Itai (1994) that a hypothesis class can be learned in such a sense in the *realizable* setting if and only if the class can be decomposed as a countable union of finite VC-dimensional classes. Shawe-Taylor et al. (1998) generalized the result to the *agnostic* setup using SRM. There has been research in the *distribution-dependent* setting as well, where one aims to quantify the *learning rates* of a hypothesis class w.r.t. the sample size (Schuurmans 1997; Antos and Lugosi 1998; Bousquet et al. 2020a). In particular, Bousquet et al. (2020a) recently showed that under mild assumptions on the realizability of the underlying generating model, the learning rates can only be asymptotically to e^{-n}, $\frac{1}{n}$ or arbitrarily slow.

In (Shawe-Taylor et al. 1998), the authors introduced a notion of *Luckiness*. Instead of fixing a prior ordering of hypotheses in the standard non-uniform PAC learning framework, the luckiness framework allows one to select an ordering of the hypotheses according to the sample observed. More formally, the Luckiness framework specifies a function $L : \mathcal{H} \times \mathcal{Z}^* \to \mathbb{R}$, where $\mathcal{H}$ is a hypothesis class and $\mathcal{Z}$ is the sample space. For any $z \in \mathcal{Z}^*$ and $g, h \in \mathcal{H}$, we will interpret $L(g, z) \geq L(h, z)$ to be that g has higher priority

than h. It is shown by Shawe-Taylor et al. (1998) that under certain assumptions on the luckiness function, one will be able to obtain a generalization bound w.r.t. the luckiness function and the underlying hypothesis. In Herbrich and Williamson (2002), the authors extended the results by defining the luckiness function directly on the algorithms. More precisely, for any learning algorithm $\mathcal{A} : \mathcal{Z}^* \to \mathcal{Y}^{\mathcal{X}}$ and $z \in \mathcal{Z}^*$, the algorithmic luckiness function maps $(\mathcal{A}(z), z)$ to the real numbers $\mathbb{R}$. Similar results on the generalization bounds were obtained for the algorithmic luckiness framework.

2.1.2 Minimum Description Length principle

In the information theory and statistics literature, the Minimum Description Length principle by Rissanen (1978, 1983) is a general concept for resolving problems involving large model classes. Here the principle of universal compression (Fitingof 1967; Davisson 1973; Shtar'kov 1987; Gallager 1979; Davisson and Leon-Garcia 1980; Rissanen 1984) is used to balance the complexity and power of model classes.

Suppose each of $\mathcal{P}_i$, $i \geq 1$, is a collection of probability models and let π be a prior on natural numbers (on the index i). For any sample S we observed, the MDL rule suggests selecting a model class in the following way

$$\hat{i} = \arg\min_{i \in \mathbb{N}} \{-\log \pi(i) + L(D|\mathcal{P}_i)\},$$

where $L(D|\mathcal{P}_i)$ is the *stochastic complexity* of the data D. As a general principle, the stochastic complexity of the data is usually obtained as the codelength of universal encodings (Rissanen 1994) of D from $\mathcal{P}_i$. Roughly speaking, we visualize

$$L(D|\mathcal{P}_i) = L(D|\hat{p}_i) + R(\mathcal{P}_i)$$

where $L(D|\hat{p}_i)$ is the length of the best possible encoding of D obtained by using any distribution in $\mathcal{P}_i$, and $R(\mathcal{P}_i)$ is the *redundancy* (Davisson 1973) of the universal encoding

of $\mathcal{P}_i$. Several refinements and variations exist, see Grünwald (2007) for a comprehensive survey.

Different types of universal codes may be used, commonly used ones include the worst-case optimal Normalized Maximum Length codes (Shtar'kov 1987) and Bayesian mixtures with a non-informative prior (Davisson 1983). Several other variants exist as well: (Ryabko 1984) considered the universal compression of Markov processes with unknown memory using a MDL 2-stage code. Barron and Cover (1991) studied the density estimation using the idea of MDL principle, where they introduced the notion of index of resolvability for bounding the performance of the MDL rule. See also (Rissanen 1984; Barron et al. 1998; Rissanen 2001; Grunwald 2004) for more results on this line. The MDL principle is widely applied in machine learning, see for example, (Quinlan and Rivest 1989; Zemel 1994; Barron 1991; Squires et al. 2017; Grünwald and Roos 2019; Proença and van Leeuwen 2020) for a sample of results on this line. We refer the reader to the book by Grünwald (2007) for more applications in different contexts.

Santhanam et al. (2014) studies compression and the choice of model in our finitely-many error framework. In our framework, we adapt a sequential view of the problem, where we observe samples of the data one by one, over an infinite horizon. While the MDL framework described above treats the data D as a batch observation and is silent on what happens when we see more data, MDL can also be adapted on online settings, most naturally when using Bayesian mixture codes (Poland and Hutter 2005).

While our framework may often suggest a different way of choosing the model class than in MDL, the key difference is not in the choice per-se. In the framework envisioned in this dissertation, we are more concerned about the following question:

For what problems does it happen that at some finite point, the choice of the model class is locked in, and no longer varies no matter how much more data we see?

See Santhanam et al. (2014) for results on compression that fit into the framework of the dissertation, building on work in (Ziv 1972a,b; Neuhoff et al. 1975; Kieffer 1978; Feder and Merhav 1996). In this dissertation, we however focus on online prediction, classification and

hypothesis testing, building on problems in (Wald 1945; Gold 1967; Barzdiņš and Freivald 1972; Cover 1973; Littlestone 1988; Dembo and Peres 1994; Haussler et al. 1994; Kulkarni and Tse 1994; Koplowitz et al. 1995; Merhav and Feder 1998; Ben-David et al. 2009; Naaman 2016).

2.1.3 Jeffreys-Lindley paradox in hypothesis testing

In the hypothesis testing literature, our setup is related to one of the ways of understanding the Jeffreys-Lindley paradox Lindley (1957). The Jeffreys-Lindley paradox is a point of divergence between the frequentist and Bayesian viewpoints in hypothesis testing that persists even when samples become arbitrarily large.

Suppose we have a Bernoulli random variable with parameter $\theta \in [0, 1]$, we would like to test whether $\theta = \frac{1}{2}$ or $\theta \neq \frac{1}{2}$. Denote H_0 to be the null hypothesis that $\theta = \frac{1}{2}$. Suppose we compare the hypotheses at any fixed significance level (p-value in frequentist literature), and let t be the value of the normalized test statistic (sample mean multiplied by $\sqrt{n}$) at the boundary of rejecting the null hypothesis. The Bayes factor at t goes to infinity (for reasonably chosen priors on $[0, 1]$) when the sample size n goes to infinity, weighing the odds heavily towards the null hypothesis. Therefore, at t the frequentist approach rejects the null hypothesis at the prescribed significance level, while the Bayes approach heavily favors the null hypothesis, see (Lindley 1957; Robert 2014; Cousins 2017; Naaman 2016).

In particular, Naaman (2016) and others recognize the source of the paradox as keeping the significance level of the test fixed despite having access to more data, and discuss its implications in methodologies that automatically refine the significance level as more data becomes available. More formally, under some assumptions, Naaman (2016) shows varying the critical value of the test statistic with sample size n as $\Phi^{-1}(1 - O(n^{-r}))$, where Φ is the CDF of standard normal distribution and r is a constant > 1, both the type I and type II errors can be made finite almost surely. Moreover, the author points out that with that tuning, the frequentist and Bayesian methods will match in the limit.

While intuitively satisfying, the resolution of Naaman (2016) only considers setting the critical values of the first moment. The work in this dissertation is broader, but has similar implications on hypothesis testing of arbitrary properties of the underlying distribution. The above observation on tuning the critical value is naturally placed into a larger framework of similar refinements in problems beyond hypothesis testing, and in addition guarantees finitely many errors almost surely with weaker assumptions.

2.1.4 More setups

In this subsection, we will collate other research directions that are related to the concepts in this dissertation.

Other data dependent guarantees Related to our $e.a.s.$-learnability setup, there has been recent research that considers data-dependent guarantees as well. Cohen et al. (2020) consider the problem of estimating distributions over countably infinite support, where they derive generalization bounds of the empirical estimator using a quantity that depends on the data. In (Bousquet et al. 2020b; Hanneke and Kontorovich 2021), the authors establish tight generalization bounds for the SVM w.r.t. the geometric margin and radius of the data. In (Chan et al. 2021), the authors establish bounds for testing Markov properties w.r.t. the k-covering time.

Regret formulations A different way of resolving prediction problems when we lack prior knowledge of the generating model is to consider agnostic guarantees. In such setups, one does not assume an underlying probabilistic model that generates the samples, instead, the samples can be generated arbitrarily and even adversarially. One quantifies the goodness of a prediction rule by means of *regret* that compares the predictor errors with those made by reference experts. We refer the reader to (Vovk 1990; Littlestone and Warmuth 1994; Vovk 1998, 2001; Cesa-Bianchi and Lugosi 2006; Shalev-Shwartz 2011; Arora et al. 2012) for more details on this topic.

Computability aspects There has been recent research that considers computational learnability in the PAC model, which differs from the line of research in (Gold 1967; Solomonoff 1964a,b; Barzdiņš and Freivald 1972; Blum and Blum 1975; Zeugmann and Zilles 2008) where the learnability is defined in the limit. See (Soloveichik 2008; Ben-David et al. 2019; Agarwal et al. 2020; Ben-David et al. 2021) for more details along this line.

2.2 Organization of the dissertation

Chapter 4 defines the notations and basic concepts that are used throughout the dissertation, and develops a general theory that is universal to all problems in the $e.a.s.$-prediction paradigm. Chapter 5 studies hypothesis testing in depth, where we also provide simple and elementary alternative proofs of the results in (Cover 1973; Koplowitz et al. 1995; Dembo and Peres 1994) using our general machinery in Chapter 4, and a partial resolution to an open problem posed in (Dembo and Peres 1994). Portions of the material in Chapter 4 and Chapter 5 have been published in (Wu and Santhanam 2021b). Chapter 6 takes on online learning problems in the $e.a.s.$-paradigm, and portions of this chapter have been published in (Wu and Santhanam 2021a). Chapter 7 studies the insurance problem as introduced in (Santhanam and Anantharam 2015), and has been published in (Wu and Santhanam 2019). We conclude the dissertation and discuss several future directions in Chapter 8.

Chapter 3

Mathematical preliminaries

For the reader's convenience, we collect some background material for the dissertation in this chapter. The results in this chapter are all quite standard, and are included here primarily for completeness sake.

3.1 Basic topological concepts

Let Ω be a metric space with metric d. A set $A \subset \Omega$ is said to be *open* if for any $x \in A$ there exists a neighborhood $N_r(x) = \{y \in \Omega : d(y, x) < r\}$ with $r > 0$ such that $N_r(x) \subset A$. We say A is *closed* if $\Omega \backslash A$ is open. A is *dense* in Ω if for any $x \in \Omega$ and $r > 0$ there exists $y \in A$ such that $d(x, y) \leq r$. For example, the set C of all continuous functions over $[0, 1]$ with the supremum norm,

$$d_\infty(f, g) = \sup_{0 \leq x \leq 1} |f(x) - g(x)|$$

forms a metric space. The set of all polynomials is dense in C under the metric d_∞ (a result known as the Stone-Weierstrass theorem (Rudin 1964, Theorem 7.26)). We say A is *compact* if for any family of open sets $\{U_i\}_{i \in I}$ with $A \subset \bigcup_{i \in I} U_i$ there exists a *finite* subset $I' \subset I$ such that $A \subset \bigcup_{i \in I'} U_i$. We refer the reader to (Rudin 1964, Chapter 2) for some basic facts about the above concepts.

We say Ω is *separable* if there exists a *countable* subset $G \subset \Omega$ such that G is dense in Ω (Srivastava 2008, Chapter 2). We say the metric space Ω is complete under metric d

if for any Cauchy sequence (Rudin 1964, Definition 3.8) $x_1, x_2, \cdots \in \Omega$ there exists $x \in \Omega$ such that $d(x_n, x) \to 0$ as $n \to \infty$. We will refer a separable and complete metric space to be a *Polish space*.

For any topological space Ω, we define the infinite product topology over Ω^∞ to be the smallest topology generated by open sets of form $A_1 \times A_2 \times \cdots A_N \times \Omega^\infty$ where $A_i \subset \Omega$ are open sets. We refer to (Srivastava 2008, Chapter 2) for more properties about infinite product topology.

The Borel σ-algebra $\mathcal{F}$ of a topological space Ω is the smallest σ-algebra generated by the *open* sets in Ω (Srivastava 2008, Chapter 3). The *cylinder* σ-algebra over Ω^∞ is defined to be Borel σ-algebra over the infinite *product topology* of Ω^∞ (Srivastava 2008, Chapter 2). See also (Srivastava 2008, Chapter 3) for basic properties about Borel σ-algebra over infinite product space.

In this dissertation, we will often define probabilities over semi-infinite strings of binary digits or natural numbers. The set of even binary semi-infinite strings is uncountable. It is helpful to think of these binary strings as binary expansions of numbers in [0,1] (with the caveat that the mapping is not one-one), and think of a finite string as the set of all infinite strings to beginning with it, thus any finite string maps to an interval in [0,1]. Therefore, one can define the σ-algebra over the semi-infinite binary sequences with Borel σ-algebra over $[0, 1]$. This is equivalent to the *cylinder* σ-algebra over the infinite product topology on $\{0, 1\}^\infty$ with discrete topology over $\{0, 1\}$. Whenever we do not specify a σ-algebra, we will use the *cylinder* σ-algebra as the underlying σ-algebra of the product space.

3.2 Basic probability theory

Let $A_1, A_2, \cdots$ be a sequence of events over the same probability space $(\Omega, \mathcal{F})$, and $1\{A_n\}$ be the indicator random variable of event A_n. We are interested in the behaviour of the random variable

$$S = \sum_{n=1}^{\infty} 1\{A_n\},$$

where $1\{A_n\}$ is the indicator function of A_n.

We say the events $A_1, A_2, \cdots$ happen *finitely often* if $S < \infty$. For any probability measure p over $(\Omega, \mathcal{F})$, we say the events $\{A_n\}$ happen finitely often with probability η, if

$$p(S < \infty) = \eta.$$

Moreover, we say $\{A_n\}$ happen *finitely* often *almost surely* w.r.t. p if $p(S < \infty) = 1$.

Borel Cantelli Lemma The following lemma is well known in the literature, and the partial converse is due to Erdős and Rényi (1959).

Lemma 1 (Borel-Cantelli lemma)**.** *For any probability measure p over $(\Omega, \mathcal{F})$, if we have*

$$\sum_{n=1}^{\infty} p(A_n) < \infty,$$

then $p(S < \infty) = 1$.

Conversely, if for any $i \neq j \in \mathbb{N}$ we have $p(A_i \cap A_j) = p(A_i)p(A_j)$ (i.e., the events $\{A_n\}$ are pairwise independent), and

$$\sum_{n=1}^{\infty} p(A_n) = \infty,$$

then $p(S < \infty) = 0$.

Proof. Let

$$M_n = \sum_{i=1}^{n} p(A_i),$$

and $M = \lim_{n\to\infty} M_n$. By Markov inequality, if $M < \infty$, we have for all $n \geq 1$

$$p(S \geq n) \leq \frac{M}{n}.$$

Let B_n be the event that $\{S \geq n\}$, we have $\forall n \in \mathbb{N}$, $B_{n+1} \subset B_n$ and $\{S = \infty\} \equiv \cap_{n\geq 1} B_n$. Therefore,

$$p(S = \infty) = \lim_{n\to\infty} p(B_n) \leq \lim_{n\to\infty} \frac{M}{n} = 0.$$

To prove the converse, we define

$$S_n = \sum_{i=1}^{n} 1\{A_i\}.$$

Clearly, we have $\mathbb{E}_p[S_n] = M_n$. We now bound the variance of S_n as follows,

$$\begin{aligned}
\mathrm{Var}_p[S_n] &= \mathbb{E}_p[(S_n - M_n)^2] \\
&= \mathbb{E}_p\left[\left(\sum_{i=1}^{n} 1\{P_i\} - p(A_i)\right)^2\right] \\
&= \sum_{i=1}^{n} \mathbb{E}_p[(1\{A_i\} - p(A_i))^2], \text{ by pairwise independence} \\
&= \sum_{i=1}^{n} p(A_i)(1 - p(A_i)) \\
&\leq \sum_{i=1}^{n} p(A_i) = M_n, \quad \text{by } p(A_i) \geq 0.
\end{aligned}$$

Now, Chebyshev's inequality gives that

$$p(|S_n - M_n| \geq M_n/2) \leq \frac{4\mathrm{Var}_p(S_n)}{M_n^2} \leq \frac{4}{M_n}.$$

This implies,

$$p(S < M_n/2) \leq p(S_n \leq M_n/2) \leq \frac{4}{M_n}.$$

Since $M_n \to \infty$, we have

$$p(S < \infty) = \lim_{n\to\infty} p(S < M_n) = 0.$$

□

The coupling Lemma and Le-Cam's two point method Assume the probability space Ω is Polish with Borel σ-algebra $\mathcal{F}$. For any distributions p_1, p_2 over the same

probability space $(\Omega, \mathcal{F})$, the total variation distance between p_1, p_2 is defined to be

$$||p_1 - p_2||_{TV} = \sup_{A \in \mathcal{F}} |p_1(A) - p_2(A)|.$$

The following lemma is well known in the literature, see (Den Hollander 2012, Section 2) for a proof.

Lemma 2 (Coupling lemma)**.** *Let (X, Y) be any coupling of (p_1, p_2), then*

$$||p_1 - p_2||_{TV} \leq \Pr[X \neq Y],$$

where the probability is over the joint distribution of (X, Y). Conversely, there exists a coupling (X', Y') of (p_1, p_2) such that

$$||p_1 - p_2||_{TV} = \Pr[X' \neq Y'],$$

and X', Y' are independent conditioned on the event $\{X' \neq Y'\}$ provided the event has positive measure. We say that (X', Y') is a perfect coupling *of (p_1, p_2).*

The following lemma bounds the total variation of n-fold *i.i.d.* distributions from the marginal distributions, which is "folklore" in the literature.

Lemma 3. *Let p_1, p_2 be distributions over the same probability space, and p_1^n, p_2^n be the n-fold* i.i.d. *distributions of p_1, p_2 respectively. Then*

$$||p_1^n - p_2^n||_{TV} \leq 1 - (1 - ||p_1 - p_2||_{TV})^n.$$

In particular, we have $||p_1^n - p_2^n||_{TV} \leq n||p_1 - p_2||_{TV}$.

Proof. Consider a perfect coupling (X_1, X_2) of (p_1, p_2). We have

$$\Pr[X_1 \neq X_2] = ||p_1 - p_2||_{TV}.$$

Let (X_1^n, X_2^n) be the n-fold *i.i.d.* copy of (X_1, X_2). We have

$$\begin{aligned}\Pr[X_1^n = X_2^n] &= \Pr[X_1 = X_2]^n, \text{ by independence} \\ &= (1 - ||p_1 - p_2||_{TV})^n.\end{aligned}$$

Now, we have

$$\begin{aligned}||p_1^n - p_2^n||_{TV} &\leq \Pr[X_1^n \neq X_2^n], \text{ by coupling lemma} \\ &= 1 - \Pr[X_1^n = X_2^n] \\ &= 1 - (1 - ||p_1 - p_2||_{TV})^n.\end{aligned}$$

The last part follows since $(1-x)^n \geq 1 - nx$ for all $x \in [0,1]$ and $n \geq 1$. □

The following lemma is due to Le Cam (1973).

Lemma 4 (Le Cam's Two Point Method)**.** *Let p_1, p_2 be two distributions over the same probability space $\mathcal{X}$. Then for any estimator $\Phi : \mathcal{X} \to \{1, 2\}$ we have*

$$\max_{i \in \{1,2\}} \Pr_{X \sim p_i}(\Phi(X) \neq i) \geq \frac{1 - ||p_1 - p_2||_{TV}}{2}.$$

Proof. Let (X_1, X_2) be a perfect coupling of (p_1, p_2). We have

$$\begin{aligned}\Pr_{X \sim p_1}(\Phi(X) \neq 1) + \Pr_{X \sim p_2}(\Phi(X) \neq 2) &= \mathbb{E}_{(X_1, X_2)}[1\{\Phi(X_1) \neq 1\} + 1\{\Phi(X_2) \neq 2\}] \\ &\geq \mathbb{E}_{(X_1, X_2)}[1\{X_1 = X_2\}] \\ &= 1 - ||p_1 - p_2||_{TV},\end{aligned}$$

where the first inequality follows from

$$1\{X_1 = X_2\} = 1 \Rightarrow 1\{\Phi(X_1) \neq 1\} + 1\{\Phi(X_2) \neq 2\} = 1,$$

and the last equality follows since (X_1, X_2) is a perfect coupling. The lemma follows from elementary inequality $2\max\{a, b\} \geq a + b$. □

3.2.1 Weak convergence of probability measures

The following material will be used in Chapter 5. Let Ω be a Polish space with metric d and $\mathcal{F}$ be the corresponding Borel σ-algebra over Ω. For any sequence of probability measures $\mu_1, \mu_2, \cdots$ and μ over $(\Omega, \mathcal{F})$, we say μ_n *weakly* converges to μ if for any *uniformly continuous* and *bounded* measurable function $g : \Omega \to \mathbb{R}$, we have

$$\int g d\mu_n \to \int g d\mu \text{ as } n \to \infty.$$

The following lemma is well known about weak convergence, see (Van Gaans 2003, Theorem 3.2) for a proof.

Lemma 5 (Portmanteau Theorem)**.** *Let $(\Omega, \mathcal{F})$ be a Polish space with Borel σ-algebra, $\mu, \mu_1, \mu_2, \cdots$ be probability measures over $(\Omega, \mathcal{F})$. Then the following statements are equivalent:*

1. *μ_n weakly converges to μ;*
2. *$\limsup_{n\to\infty} \mu_n(C) \leq \mu(C)$ for all* closed *$C \subset \Omega$;*
3. *$\liminf_{n\to\infty} \mu_n(U) \geq \mu(U)$ for all* open *$U \subset \Omega$;*
4. *$\mu_n(A) \to \mu(A)$ for every Borel set $A \subset \Omega$ such that $\mu(\partial A) = 0$, where ∂A is the* boundary *of A.*

Note that for probability measures over $\mathbb{R}^d$, μ_n weakly converges to μ if and only if $F_n(\mathbf{x}) \to F(\mathbf{x})$ at all continuity points $\mathbf{x}$ of F, where F_n and F are the CDFs of μ_n and μ respectively (Hunter 2006, Theorem 2.32).

For any probability measures μ, ν over $(\Omega, \mathcal{F})$, the Prokhorov metric is defined to be

$$d_P(\mu, \nu) = \inf\{\alpha > 0 : \mu(A) \leq \nu(A_\alpha) + \alpha \text{ and } \nu(A) \leq \mu(A_\alpha) + \alpha,\ \forall A \in \mathcal{F}\},$$

where $A_\alpha = \{x : d(x, A) < \alpha\}$ and $\emptyset_\alpha = \emptyset$ for all $\alpha > 0$. It can be shown that convergence under Prokhorov metric is consistent with the weak convergence (Van Gaans 2003, Theorem 4.1). Moreover, the topological space induced by Prokhorov metric is *separable* (Van Gaans 2003, Proposition 4.4).

For any collection $\mathcal{P}$ of probability measures over the same probability space $(\Omega, \mathcal{F})$, we say the $\mathcal{P}$ is tight if for any $\epsilon > 0$ there exists a *compact* set $S_\epsilon \subset \Omega$ such that $\forall \mu \in \mathcal{P},\ \mu(S_\epsilon) \geq 1-\epsilon$. We have the following lemma which follows directly from Prokhorov's theorem (Van Gaans 2003, Theorem 5.2), and will be used in Chapter 5.

Lemma 6. *Let $\mu, \mu_1, \mu_2, \cdots$ be a sequence of measures over the same probability space $(\Omega, \mathcal{F})$ such that μ_n weakly converges to μ. Then $\{\mu, \mu_1, \mu_2, \cdots\}$ is tight.*

Proof. We show that the collection $\{\mu, \mu_1, \mu_2, \cdots\}$ is compact under weak convergence topology. For any open covering of $\{\mu, \mu_1, \mu_2, \cdots\}$, there must be an open set U such that $\mu \in U$. By weak convergence we have there exists some N such that for all $n \geq N$ we have $p_n \in U$ as well. Therefore, $\{\mu, \mu_1, \mu_2, \cdots\}$ is compact. The lemma follows by Prokhorov's theorem, which asserts that tightness is equivalent to compactness for weak convergence topology. □

We also refer the reader to the books by Parthasarathy (2005) and Billingsley (2013) for more discussion about weak convergence.

Chapter 4

A general framework

This chapter develops the basic tenets of the eventual almost sure prediction paradigm, and studies relationships among the variants of this formulation. Section 4.1 defines $e.a.s.$-predictability and the concept of η-predictability. Informally, the η-predictability defines a notion of model classes where we are able to predict with *bounded* number of errors and with confidence $\geq 1 - \eta$ uniformly over all models in the class. In effect, η-predictability characterizes inference with confidence $\geq 1 - \eta$ from a finite sample.

Our first main result is Theorem 1 which shows that the $e.a.s.$-predictability of a model class is in many cases equivalent to a decomposition of the class into nested η-predictable subclasses. Therefore, this has the context of matching sample size to complexity—as the sample size increases, we deal with larger subsets of the model class. Put another way, this implies that such $e.a.s.$-predictable estimators can be obtained via regularization based on the sample at hand. We illustrate how Theorem 1 can be used to break seemingly impossible tasks as posed in Cover (1973) into essentially trivial problems in Example 4. In subsequent chapters, Theorem 1 is refined into different contexts—hypothesis testing, online learning and insurance tasks.

Theorem 1, while broadly applicable, does not always characterize $e.a.s.$-predictability in the most general setting—indeed we show counterexamples for the same. We then refine Theorem 1 for several special cases. Theorem 5 completely characterizes the $e.a.s.$-predictability in the *supervised setting*, i.e., a setting when the losses are observable from

samples, using methods reminiscent of structural risk minimization. In Theorem 6, we establish a different tight characterization of the *e.a.s.*-predictability for *pure estimation* problems with *i.i.d.* sampling and finite prediction domain. This result forms the basis of the almost sure hypothesis testing setup of Dembo and Peres (1994) that we will study in detail in Chapter 5.

Finitely many errors does not imply that there exists a way, a stopping rule, that identifies when we are past the point of the final error. Indeed, in many cases, such a stopping rule is impossible. Section 4.4 captures this refinement with the notion of *e.a.s.*-learnability. Informally, *e.a.s.*-learnability of a model class requires that for any given confidence parameter $\eta > 0$, we are able to find a prediction rule and stopping rule such that the probability of the prediction rule making errors after the stopping rule has stopped is upper bounded by η. We provide a characterization of the *e.a.s.*-learnability in Theorem 10 through a notion of *identifiability*. This characterization is tight when the processes are *i.i.d.*.

Lastly, we introduce other natural variations of the *e.a.s.*-predictability framework in the Section 4.5, including a notion of weak *e.a.s.*-predictability and the notion of predictability with finite expected loss. We show that predictability with finite expected loss implies *e.a.s.*-predictability, and that *e.a.s.*-predictability implies weak *e.a.s.*-predictability. Converses do not hold, and these are essentially different formulations—we also provide examples that illustrate these nuances. We conclude the chapter with several open problems that concern the exact characterizations of the above variations.

4.1 Basic setup of our paradigm

Let $\mathcal{X}$ be a Polish space (separable completely metrizable topological space) with Borel σ-algebra and $\mathcal{P}$ be a collection of probability measures over the cylinder σ-algebra 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 $p \in \mathcal{P}$. We will denote $X_i^j = (X_i, X_{i+1}, \cdots, X_j)$ in the sequel.

Prediction is modeled as a meassurable 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* is a function $\ell : \mathcal{P} \times \mathcal{X}^* \times \mathcal{Y} \to \{0, 1\}$ such that $\ell(p, \cdot, \cdot)$ is measurable for all $p \in \mathcal{P}$. 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 $Y_n = \Phi(X_1^{n-1})$ 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 that Nature chooses at the beginning.

The loss in general can be any 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 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 (η-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 $\mathcal{P}$.

Example 1. *Let $\mathcal{P}$ be the class of all i.i.d. processes with marginal distributions supported on $[-B, B]$ for some $B > 0$. We define the loss to be, for all $p \in \mathcal{P}$*

$$\ell_\epsilon(p, X_1^{n-1}, Y_n) = 1\left\{|\mathbb{E}[X_1] - Y_n| > \epsilon\right\}.$$

By Chebyshev's inequality we know that the collection $(\mathcal{P}, \ell_\epsilon)$ is η-predictable for all $\eta > 0$ and $\epsilon > 0$ by predicting the empirical mean of X_1^{N-1} at all steps $\geq N$ for any constant $N > \frac{B^2}{\epsilon^2 \eta}$.

Definition 2 (*e.a.s.*-predictable)**.** *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 7. *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. Then $(\mathcal{P}, \ell)$ is e.a.s.-predictable.*

Proof. From the definition of η-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. □

Refined notions of nestings Motivated by the above lemma, we introduce the following two refinements on the notion of nestings.

Definition 4 (Universal nestings)**.** *A nesting $\{\mathcal{P}_i, i \geq 1\}$ of $\mathcal{P}$ is a* universal *nesting w.r.t loss ℓ, if for all $\eta > 0$ and $i \geq 1$, $(\mathcal{P}_i, \ell)$ is η-predictable.*

Definition 5 (η-nestings)**.** *For any $\eta > 0$, a nesting $\{\mathcal{P}_i^\eta, i \geq 1\}$ of $\mathcal{P}$ is an* η-nesting *w.r.t. loss ℓ, if for all $i \geq 1$, $(\mathcal{P}_i^\eta, \ell)$ is η-predictable.*

Clearly, if a class $\mathcal{P}$ has a universal nesting then it has η-nesting for all $\eta > 0$. However, the converse is not true, as we will see in Section 4.2. Indeed, as we will see in Theorem 1 that these two concepts will form the basis for characterizing *e.a.s.*-predictability.

The following technical lemma will be useful in our following proofs.

Lemma 8. *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 a predictor 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 η-predictable, there exists a number n_i' and a predictor Φ_i such that for all $p \in \mathcal{P}_i$, the probability Φ_i incurs non-zero $\ell-$loss on samples larger than n_i' is $\leq \eta$. We can therefore choose an increasing sequence $\{n_i, i \geq 1\}$ with $n_i \geq n_i'$. Note that each $\mathcal{P}_i$ is η-predictable with sample 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. □

Example 2. *Let $\mathcal{P}$ be the class of all i.i.d. processes with marginal distributions over $\mathbb{R}$ that have finite absolute first moments. The loss ℓ_ϵ is the same as the loss in Example 1. We show that the collection $(\mathcal{P}, \ell_\epsilon)$ is e.a.s.-predictable for all $\epsilon > 0$ by using Lemma 7 above. We define*

$$\mathcal{P}_i = \{p \in \mathcal{P} : \mathbb{E}[1\{|X_p| > i\}X_p] \leq \epsilon/2\},$$

where X_p is the marginal random variable governed by probability law p. For any $p \in \mathcal{P}$, we have $p \in \mathcal{P}_i$ for some $i \geq 1$ since $\mathbb{E}[|X_p|] < \infty$, i.e., $\{\mathcal{P}_i : i \geq 1\}$ forms a nesting of $\mathcal{P}$. By Example 1, we know that $(\mathcal{P}_i, \ell_\epsilon)$ is η-predictable for any $\eta > 0$ by predicting empirical means on the truncated variables $1\{|X_p| \leq i\}X_p$. The claim follows by Lemma 7.

With a slight modification of the nesting and loss, we can actually obtain a predictor Φ such that $\Phi(X_1^{n-1}) \to \mathbb{E}[X_1]$ almost surely whenever $\mathbb{E}[|X_1|] < \infty$. Clearly, the strong law of large numbers already shows that the empirical mean is sufficient to achieve this. What is interesting here is that the almost sure consistency of our predictor follows easily from a simple application of Borel–Cantelli lemma, while the proof of strong law of large numbers is non-trivial. Similar argument could also establish almost sure consistent estimators for other functionals of random variables, e.g. the entropy of random variables over $\mathbb{N}$.

4.2 Characterization of *e.a.s.*-predictability

We first provide a general characterization of the *e.a.s.*-predictability without any assumption on the class $\mathcal{P}$ and loss ℓ.

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

Proof. The sufficiency follows directly from Lemma 7. We now prove the necessary condition. Suppose $\mathcal{P}$ is *e.a.s.*-predictable, we show that $\mathcal{P}$ can be decomposed into a nesting of η-predictable collections. By Definition 2, there exists a predictor Φ such that for all $p \in \mathcal{P}$, Φ makes finitely many errors with probability 1. For $\eta > 0$, we define

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

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

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

We have $p(A_k^p) \to 0$ as $k \to \infty$ by *e.a.s.*-predictability of Φ. Therefore, there must be some k such that $p(A_k^p) < \eta$, and for such a number k we have $p \in \mathcal{P}_k^\eta$. Therefore, $\mathcal{P} = \bigcup_{n \in \mathbb{N}^+} \mathcal{P}_n^\eta$. □

Theorem 1 shows that to prove a collection $(\mathcal{P}, \ell)$ is *e.a.s.*-predictable, it is sufficient to find a *universal nesting* of $\mathcal{P}$. To prove that a collection $(\mathcal{P}, \ell)$ is not *e.a.s.*-predictable it is sufficient to show that for some $\eta > 0$ there is no *η-nesting* of $\mathcal{P}$. We can also interpret the nesting as *regularization* of the model class $\mathcal{P}$, and more details will follow in the theorems of Section 4.3.1. We will illustrate how this can be applied to concrete problems in different setups in the following chapters.

Note that the necessary and sufficient conditions in Theorem 1 do not necessarily match in general. Indeed, the following two theorems show that the necessary (respectively sufficient) condition in Theorem 1 is not sufficient (respectively necessary).

Theorem 2. *There exists a collection $(\mathcal{P}, \ell)$, such that for all $\eta > 0$, $\mathcal{P}$ has an η-nesting, but $(\mathcal{P}, \ell)$ is not e.a.s.-predictable.*

Proof. 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 1 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 $b \in \{0, 1\}$

$$\Phi(x_1, \cdots, x_m, b, \cdots, b) = b, \tag{4.1}$$

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 (4.1) 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 infinitely often almost surely by the converse of the 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. □

Theorem 3. *There exists a collection $(\mathcal{P}, \ell)$, such that $(\mathcal{P}, \ell)$ is e.a.s.-predictable, but $\mathcal{P}$ has no universal nesting.*

Proof. Let $\mathcal{P}$ be the class of all random processes over $\{0,1\}^\infty$ such that for any $p \in \mathcal{P}$ there exists a parameter $b \in \{0,1\}$ and strictly monotonically increasing integer sequence $M_1, M_2, \cdots$ with $M_1 = 1$ satisfying for all $j \geq 1$, $X_{M_j} = X_{M_j+2} = \cdots = X_{M_{j+1}-1}$, and $X_{M_j}^{M_{j+1}-1}$ is *independent* of all *other* random variables in the process, and

$$p(X_{M_j} = b) = 1 - \frac{1}{(j+1)^2}.$$

Note that a process in $\mathcal{P}$ is purely determined by the parameter b and sequence $\{M_j\}_{j\geq 1}$. We will denote a process to be p_b if the parameter is b. Clearly, by Borel-Cantelli lemma $\mathcal{P}$ is *e.a.s.*-predictable under the loss $\ell(p_b, X_1^n, Y_n) = 1\{Y_n \neq b\}$ by predicting X_{n-1} at each step n. We now show that there is no nesting $\{\mathcal{P}_i, i \geq 1\}$ of $\mathcal{P}$ such that $(\mathcal{P}_i, \ell)$ is η-predictable for all $\eta > 0$ and $i \geq 1$. Our approach is a proof by contradiction. Let $\{\mathcal{P}_i, i \geq 1\}$ be a *universal nesting* of $\mathcal{P}$ w.r.t. loss ℓ, i.e., $(\mathcal{P}_i, \ell)$ is η-predictable for all $\eta > 0$ and $i \geq 1$. We construct a distribution in $\mathcal{P}$ that is not in $\bigcup_{i\geq 1} \mathcal{P}_i$, a contradiction on our supposition that $\mathcal{P} = \bigcup_{i\geq 1} \mathcal{P}_i$.

For $j \geq 1$, let R_j be a number such that the class $\mathcal{P}_j$ is η_j-predictable with a sample of size R_j, where η_j will be specified later. W.l.o.g., we may assume R_j is strictly increasing on j. Let p_0, p_1 be two distributions in $\mathcal{P}$ that are associated with the sequence $M_1 = 1$ and $(M_j = R_{j-1} + 1)_{j\geq 2}$ with parameter $b = 0$ and $b = 1$ respectively, i.e. p_0, p_1 share the same partition of independent blocks but with different parameter for b.

Let $||p_0^{R_j} - p_1^{R_j}||_{TV} = 1 - \epsilon_j$, where $p_b^{R_j}$ with $b \in \{0,1\}$ is the distribution of p_b on the initial length-R_j binary strings. Observe that the probability of any length-R_j binary string under either p_0 or p_1 is purely a function of the number j of blocks which are all-0 (or equivalently all-1), and therefore, so does ϵ_j, the total variation distance. Specifically, ϵ_j depends only on j, not the sequence $\{R_j\}$, or $\{\eta_j\}$, and in particular we can choose $\eta_j < \frac{\epsilon_j}{2}$.

By Lemma 4, we know that any prediction rule will make an error with probability at least $\frac{\epsilon_j}{2}$ on either p_0 or p_1 at time step $R_j + 1$ if they both belong to $\mathcal{P}_j$. Since $\epsilon_j/2 > \eta_j$, we conclude that for all $j \geq 1$ at least one of p_0 or p_1 cannot be in $\mathcal{P}_j$.

But $\mathcal{P}_j \subset \mathcal{P}_k$ for all $k \geq j$. Let t be the smallest number that contains one of p_0 or p_1. $\mathcal{P}_t$ cannot contain both, per the argument above, it follows that $\mathcal{P}_k$ contains that distribution for all $k \geq t$. However, the argument above implies that the distribution missing from $\mathcal{P}_t$ cannot be in $\bigcup_{k\geq 1} \mathcal{P}_k$, contradicting the assumption that $\{\mathcal{P}_i, i \geq 1\}$ is a nesting of $\mathcal{P}$. $\square$

4.3 Specialized settings

While the necessary and sufficient conditions in Theorem 1 do not match up in general as shown above, the characterization of *e.a.s.*-predictability can be tightened in several natural settings. We split our further analysis into two settings—one where the loss can be observed (the supervised setting) and where the loss cannot be observed (unsupervised setting). We provide a tight characterizations for the supervised setting, and for *i.i.d.*generated data when the loss is only a function of the source and the prediction in the unsupervised setting.

In this section, the distinction between η-nesting and universal nestings will become clear as well—we will also encounter examples that allow η-nestings for all $\eta > 0$, but which lack a universal nesting even in the supervised setting.

We first prove the following theorem, which shows that if $\mathcal{P}$ is *countable* then existence of universal nesting $\Leftrightarrow$ *e.a.s.*-predictability $\Leftrightarrow$ existence of η-nesting for all $\eta > 0$.

Theorem 4. *Let $\mathcal{P}$ be a* countable *class and ℓ be an arbitrary loss. Then there exists a universal nesting of $\mathcal{P} \Leftrightarrow (\mathcal{P}, \ell)$ is e.a.s.-predictable $\Leftrightarrow$ there exist η-nesting of $\mathcal{P}$ for all $\eta > 0$.*

Proof. Let $p_1, p_2, \cdots$ be an arbitrary enumeration of $\mathcal{P}$.

We first show that if $(\mathcal{P}, \ell)$ is *e.a.s.*-predictable then there is a universal nesting for $\mathcal{P}$. Define $\mathcal{P}_k = \{p_1, \cdots, p_k\}$. We show that $(\mathcal{P}_k, \ell)$ is η-predictable for all $\eta > 0$ and $k \geq 1$. To see this, let Φ be the *e.a.s.*-prediction rule for $(\mathcal{P}, \ell)$. For any given $\eta > 0$ and $i \geq 1$, there exists a number $N_{i,\eta}$ such that Φ makes errors after step $N_{i,\eta}$ w.p. $\leq \eta$ when p_i is in force. Now, for any $\mathcal{P}_k$, we know that Φ makes no errors after step $\max_{1 \leq i \leq k}\{N_{i,\eta}\}$ w.p. $\geq 1 - \eta$ no matter what source $\in \mathcal{P}_k$ is in force. Therefore $(\mathcal{P}_k, \ell)$ is η-predictable for all $\eta > 0$ and $k \geq 1$.

We now prove that if there exists η-nesting of $\mathcal{P}$ for all $\eta > 0$ then $(\mathcal{P}, \ell)$ is *e.a.s.*-predictable. To see this, let $\{\mathcal{P}_k^j, k \geq 1\}$ be the 2^{-j}-nesting of $\mathcal{P}$. By Lemma 8, we can assume that there exists prediction rule Φ_k^j such that Φ_k^j makes no errors after step k w.p. $\geq 1 - 2^{-j}$ for all $p \in \mathcal{P}_k^j$. We can assume, w.l.o.g., that $\mathcal{P}_k^j$ contains only the sources in $\{p_1, \cdots, p_k\}$. For all $j \geq 1$, let K_j be the minimum index of class in $\{\mathcal{P}_k^j\}$ such that $\mathcal{P}_{K_j}^j$ contains all sources in $\{p_1, \cdots, p_j\}$. We now construct the *e.a.s.*-prediction rule as follows. We partition the prediction into phases, initially at phase 0. We are in phase $j \geq 1$ if the time step n satisfies $K_j \leq n < K_{j+1}$. At phase $j \geq 1$ we use $\Phi_{K_j}^j$ to make the prediction and predict arbitrarily at phase 0.

Suppose the underlying source is p_t, we know that $\mathcal{P}_{K_j}^j$ contains p_t for all $j \geq t$. Meaning that, the probability we make errors at phase $j \geq t$ is at most 2^{-j}. The theorem follows by the Borel-Cantelli lemma. □

4.3.1 Supervised setting

We now characterize *e.a.s.*-predictability in *supervised* setting, i.e. the loss $\ell : \mathcal{X}^* \times \mathcal{Y} \to \{0, 1\}$ is independent of the underlying source p (or equivalently the loss can be gauged

from the samples). We show that the existence of η-nesting for all $\eta > 0$ is necessary and sufficient to achieve *e.a.s.*-predictability in the supervised case.

This result is to be interpreted in the context of regularization, where one restricts $\mathcal{P}$ to match the amount of data available. Recall that when $\mathcal{P}$ has a 2^{-j}-nesting, the class $\mathcal{P}$ is broken into a nesting $\left\{\mathcal{P}_i^j, i \geq 1\right\}$ of classes, where each $\mathcal{P}_i^j$ is 2^{-j}-predictable with sample size i. Given any sample, we say $\mathcal{P}_i^j$ is consistent with the sample if the 2^{-j}-prediction rule for $\mathcal{P}_i^j$ makes no error after step i on the sample. Note that, since we have assumed the loss is supervised, the consistency can be verified from data. We define the complexity of the class $\mathcal{P}_i^j$ to be $i + j$. Now, for any given sample, we will match the sample with a class $\mathcal{P}_i^j$ that is consistent with the sample and has *minimum* complexity $i + j$. As we see more and more samples the competing class $\mathcal{P}_i^j$ may change over time. However, when $\mathcal{P}$ is *e.a.s.*-predictable, the implication is that such a regularization approach will stabilize at some point. After a certain finite point, the competing class $\mathcal{P}_i^j$ will no longer change.

This provides an assurance that matching the sample with restricted classes is a fundamentally valid strategy to settle on the accurate answer. While the theorem below can be proved in multiple ways, we provide an approach that reflects the intuition from the regularization perspective.

Theorem 5. *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 η-nesting of $\mathcal{P}$.*

Proof. The necessity follows directly from Theorem 1. We now prove that the existence of η-nesting for all $\eta > 0$ is also sufficient. Suppose that for all $j \in \mathbb{N}$, there exists a nesting $\left\{\mathcal{P}_n^j : n \geq 1\right\}$ of $\mathcal{P}$ such that $\mathcal{P}_n^j$ is 2^{-j}-predictable for all $n \geq 1$. Furthermore, from Lemma 8, we can choose a decomposition such that $\mathcal{P}_n^j$ is 2^{-j} predictable with sample size n. Therefore, there exist predictors $\Phi_{n,j}$ such that for all $p \in \mathcal{P}_n^j$

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

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

$$(k, i) = \underset{(n,j)\in\mathbb{N}\times\mathbb{N}}{\operatorname{argmin}} \{j + n \mid I(n, j) = 1\}. \tag{4.2}$$

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 $n_j = \min\{n \mid p \in \mathcal{P}_n^j\}$. 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 (4.2). 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 (4.2) 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 decomposition of the class in Theorem 5 as a regularization. The measure of complexity here is $n + j$. Here j equals the negative log of the confidence probability of the $2^{-j}-$nesting $\{\mathcal{P}_n^j, n \geq 1\}$ and n denotes that we restrict the class to $\mathcal{P}_n$, i.e., it is a measure of the complexity of the class considered. The smaller $n + j$ is, the less the complexity the $\mathcal{P}_n^j$ has. Alternatively, one may view the selection of the model classes in equation (4.2) 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.*

While *e.a.s.*-predictability is completely characterized by the existence of η-nesting for all $\eta > 0$ in the supervised setting, *e.a.s.*-predictability does not imply the existence of a *universal nesting* as the following example demonstrates. A corollary of this is that the existence of η-nesting for all η is not equivalent to the existence of a universal nesting even in the supervised setting—the two are different concepts.

Example 3. *Let $\mathcal{P}$ be the class of the following processes. For any infinite sequence $M_1 < M_2 < \cdots < M_n < \cdots \in \mathbb{N}^\infty$, we define process p^M over $\{0,1\}^\infty$ as follows:*

$$X_{M_n} \sim \textit{Bernoulli}\left(\frac{1}{2}\right) \textit{ independently for all } n \in \mathbb{N},$$

and $X_t = 0$ for all other time step t. The loss ℓ is defined as follows:

$$\ell(p^M, X_1^t, Y_t) = 0 \textit{ iff there exists at least one } 1 \textit{ in } X_1^t.$$

Note that the loss only depends on the sample but not the underlying source and predictions. Clearly, the loss is observable from the sample, thus the setup is supervised. Furthermore, $(\mathcal{P}, \ell)$ is e.a.s.-predictable since for any source p^M, $p^M(\exists n \geq 1,\ X_{M_n} = 1) = 1$.

Let $\mathcal{P}_1 \subset \mathcal{P}_2 \subset \cdots \subset \mathcal{P}$ be an arbitrary nesting of $\mathcal{P}$. We show that $\{\mathcal{P}_k,\ k \geq 1\}$ cannot be a universal nesting of $\mathcal{P}$ by contradiction.

Suppose the above were a universal nesting, so each $\mathcal{P}_k$ in the nesting is η-predictable for all $\eta > 0$. For all k, let B_k be the sample size for $\mathcal{P}_k$ to achieve $\frac{1}{2^{k+1}}$-predictability. Now let $p^B \in \mathcal{P}$ be the source that is associated with sequence $\{B_k + 1\}_{k \in \mathbb{N}^+}$.

We claim that $p^B \notin \mathcal{P}_k$ for all $k \in \mathbb{N}^+$. To see this, note that $p^B(\forall i \leq k,\ X_{B_i+1} = 0) = \frac{1}{2^k}$ which in turn implies that with probability at least $\frac{1}{2^k}$, we make an error at step $B_k + 1$. If for any k, p^B were in $\mathcal{P}_k$, $\mathcal{P}_k$ would not be $\frac{1}{2^{k+1}}$-predictable. Therefore $p^B \notin \bigcup_k \mathcal{P}_k$. However, we also know that $p^B \in \mathcal{P}$, meaning that $\{\mathcal{P}_k, k \geq 1\}$ cannot be a nesting of $\mathcal{P}$.

4.3.2 Unsupervised setting

While η-nestings (for all η) characterize the supervised setting, we will see now that universal nestings become necessary to achieve *e.a.s.*-predictability in certain unsupervised settings. In the process, we uncover another setting where we obtain a full characterization of η-predictability—the *pure estimation* case with *finite* prediction domain in the *i.i.d.* setting, where the loss is a function of only the source and prediction. This includes, e.g., predicting the rationality of the mean of the marginal (Cover 1973). Note that in this case, there may be no feedback in general on the loss incured by a predictor, since the loss is a function of the unobserved source generating the data. Therefore, we classify this scenario under the unsupervised category.

Theorem 6. *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 universal nesting $\{\mathcal{P}_i,\ i \geq 1\}$ of $\mathcal{P}$ such that for all $\eta > 0$, $(\mathcal{P}_i, \ell)$ is η-predictable.*

Proof. Applying Theorem 1 to $(\mathcal{P}, \ell)$, we know that if the universal nesting exists, then $(\mathcal{P}, \ell)$ is *e.a.s.*-predictable. Theorem 1 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 $(\mathcal{P}_i^\eta, \ell)$ is η-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.* $(\mathcal{P}_i, \ell)$ is also η-predictable for all $\eta > 0$.

To do so, we show that if $(\mathcal{P}_i, \ell)$ is $(\frac{1}{|\mathcal{Y}|} - \epsilon)$-predictable for any $\epsilon > 0$ then it is η-predictable for all $\eta > 0$. Suppose N is the sample size such that a predictor Φ makes no errors 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 $\frac{1}{|\mathcal{Y}|} - \epsilon/2$ frequency among $Y_1, \ldots, Y_M$. But at least one element of $\mathcal{Y}$ appears more than $1/|\mathcal{Y}|$ frequency among $Y_1, \ldots, Y_M$, and any such an element must incur 0 loss. The theorem follows. □

Remark 2. *Note that the finiteness of the prediction domain $\mathcal{Y}$ is essential for the proof of the above Theorem to work. We leave it as an open problem to determine if the theorem above can be extended to countable domain $\mathcal{Y}$ and general loss $\ell : \mathcal{P} \times \mathcal{Y} \to \{0, 1\}$.*

We give another example, which illustrates how Theorem 1 can be used to derive prior known results.

Example 4. *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 1 above and an argument evocative of regularization.

Define the loss $\ell(p, X_1^n, Y_n) = 0$ iff Y_n gives the correct 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 1, 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 $S_k \subset S_{k+1}$ and

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

Since $\mathcal{B}$ is e.a.s.-predictable, Theorem 1 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, η-predictability of $\mathcal{B}_k$ implies that we must have

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

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

Suppose we want $\mathcal{S}$ to contain all rational numbers in $[0,1]$. Then it follows (see Appendix A 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 a meager or Baire first category set (Rudin 2006, Chapter 2.1), proving (Koplowitz et al. 1995, Theorem 2).

While Theorem 1 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 $\ell(p, X_1^n, Y_n) = 1\{p \notin \mathcal{H}_{Y_n}\}$. Dembo and Peres (1994) showed that if the distributions in $\mathcal{H}_1 \cup \mathcal{H}_2$ have densities *and* there exists some $r > 1$ such that the r-th norm of the densities are 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 Chapter 5 for definition) under any metric consistent with weak convergence topology (see Chapter 3.2.1). Dembo and Peres (1994) asked whether the condition $r > 1$ 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 left to Chapter 5.

4.4 Capturing the final error

While *e.a.s.*-predictability is an attractive setup when considering rich model classes, we would like to see if a predictor 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 $\tau = 1$ as the rule has stopped waiting.

Definition 6 (*e.a.s.*-learnable)**.** *A collection $(\mathcal{P}, \ell)$ is said to be eventually almost surely (e.a.s.)-learnable, if for any $\eta > 0$, there exists a universal prediction rule Φ_η together with a stopping rule τ_η, such that for all $p \in \mathcal{P}$*

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

and

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

Clearly, *e.a.s.*-learnability implies *e.a.s.*-predictability.

Theorem 7. *Any e.a.s.-learnable* $(\mathcal{P}, \ell)$ *is e.a.s.-predictable.*

Proof. Suppose $(\mathcal{P}, \ell)$ is e.a.s.-learnable. Then for each i, we let Φ_i and τ_i be the predictor and stopping rule pair respectively that *e.a.s.*-learns $(\mathcal{P}, \ell)$ with $\eta = 1/2^i$. By definition, we have that the probability Φ_i makes an error after τ_i stops (i.e. $\tau_i = 1$) is $\leq \frac{1}{2^i}$. Let Φ_0 be an arbitrary predictor.

Now, there are countably many stopping rules (one for each natural number $i \geq 0$) and each such rule stops at a finite time with probability 1, we conclude that with probability 1 all of them would have stopped simultaneously at some finite time by a union bound.

We initialize $t = 1$ (t will stand for the stage). As we see more of the sample, at any stage t, we predict using the prediction rule Φ_{t-1}, till τ_t halts (i.e. $\tau_t = 1$). At that point, we move to stage $t + 1$. For $t \geq 2$, the probability of making an error in stage t is $\leq 2^{-t}$. Invoking the Borel-Cantelli lemma, we conclude that we make errors in finitely many stages almost surely, and the Theorem follows. □

However, *e.a.s.*-predictability does not imply *e.a.s.*-learnability.

Example 5. *Let* $\mathcal{P}$ *be the class of all* i.i.d. *Bernoulli random processes with parameters in* $[0, 1]$. *For any* $p \in \mathcal{P}$, *we would like to determine if the parameter of the process equals 1/2 or not (prediction is 1 if parameter is 1/2, 0 else), namely, the loss is* $\ell(p, X_1^n, Y_n) = 1\{Y_n = 1\{p = \frac{1}{2}\}\}$, *where in a slight abuse of notation, we use p to denote both the iid Bernoulli source and its parameter. In Chapter 1.1, we have shown that* $(\mathcal{P}, \ell)$ *is e.a.s.-predictable. We now show that* $(\mathcal{P}, \ell)$ *is not e.a.s.-learnable.*

Suppose otherwise, let Φ, τ *be the prediction rule and stopping rule that e.a.s.-learns* $(\mathcal{P}, \ell)$ *with* $\eta = \frac{1}{4}$. *Consider the Bernoulli 1/2 source. Since* τ *stops finitely almost surely on all sources, there exists a number N such that* τ *has stopped before step N w.p.* $\geq \frac{3}{4}$ *when the Bernoulli 1/2 source is in force.*

Let A be the event that τ has stopped before N and Φ takes value 1 at step $N+1$. Observe that $p(A) \geq \frac{1}{2}$ by union bound. Let p' be any source other than the Bernoulli 1/2 source such that $||p' - p||_{TV} \leq \frac{1}{8N}$. We have $p'(A) \geq \frac{1}{2} - \frac{1}{8} > \frac{1}{4}$ from Lemma 3. But Φ incurs an error at step $N+1$ on any sequence in A whenever p' is in force, which contradicts that Φ incurs errors after the stopping rule has halted on a set of probability $< 1/4$.

4.4.1 Unions of learnable classes

Even finite unions of *e.a.s.*-learnable classes need not be *e.a.s.*-learnable, as the stopping rule for one class need not even stop with probability 1 on sources of the other. It is more interesting to consider nested unions of classes. Clearly finite nested unions of *e.a.s.*-learnable classes are trivially *e.a.s.*-learnable, but countable unions of nested *e.a.s.*-learnable classes may not necessarily be *e.a.s.*-learnable. To see this, consider the subclasses $\mathcal{P}_k$ we constructed for the example in Chapter 1.1. We know that each of the $\mathcal{P}_k$ is *e.a.s.*-learnable but $\bigcup_{k\geq 1} \mathcal{P}_k$ is not *e.a.s.*-learnable per Example 5 above.

The following theorem however shows that a countable union of nested *eas*-learnable classes is always *e.a.s.*-predictable.

Theorem 8. *Let $\mathcal{P}$ be a class of distributions, and $\{\mathcal{P}_i,\ i \geq 1\}$ be a nesting of $\mathcal{P}$. If for all $i \geq 1$, $(\mathcal{P}_i, \ell)$ is e.a.s.-learnable, then $(\mathcal{P}, \ell)$ is e.a.s.-predictable.*

Proof. The proof is similar to the proof of Lemma 7. For any k, i, let Φ_i^k and τ_i^k be the prediction rule and stopping rule respectively that achieves *e.a.s.*-learnability for $(\mathcal{P}_i, \ell)$ with parameter $\eta = \frac{1}{k^2 2^i}$. We now construct the *e.a.s.*-prediction rule for $(\mathcal{P}, \ell)$ as follows. We partition the prediction into phases, at the beginning, we are at phase 0 and the prediction is arbitrary. We will go from phase $k-1$ to phase k if at least one of τ_i^k with $i \geq k$ has stopped. Denote I_k to be the index of the stopping rule in $\{\tau_i^k\}_{i\geq k}$ that stops earliest during phase $k-1$. We will use the prediction rule $\Phi_{I_k}^k$ to make the prediction during phase k. Suppose the underlying distribution is $p \in \mathcal{P}_j$ for some j. We know that for each phase $k \geq 1$ w.p. 1 there must be some stopping rule in $\{\tau_i^k\}_{i\geq 1}$ with $i \geq j$ that stops finitely

almost surely. Using a union bound, we therefore conclude that all the phases will be finite simultaneously almost surely. Now, by definition of *e.a.s.*-learnability, the probability of making errors at phase $k \geq j$ is upper bounded by $\frac{1}{k^2}$ using a union bound. By the Borel-Cantelli lemma, we will make errors in finitely many phases almost surely, proving that $\mathcal{P}$ is *e.a.s.*-predictable. □

Remark 3. *Note that the nesting property is required for Theorem 8 to hold, that is, it is possible that countable unions of non-nested e.a.s.-learnable classes are not e.a.s.-predictable. To see this, let* $r_1, r_2, \cdots$ *be an arbitrary enumeration of rationals in* $[0,1]$*. We denote* $\mathcal{B}_{i,j} = \{x \in [0,1] : x \text{ is irrational and } |x - r_i| \geq \frac{1}{j}\}$*, and* $\mathcal{P}_{i,j}$ *be the class of* i.i.d. *Bernoulli processes with parameters in* $\mathcal{B}_{i,j}$*. Denote* $\mathcal{P} = \bigcup_{i,j \geq 1} \mathcal{P}_{i,j}$*. Let* ℓ *be the rationality testing loss of Cover (1973) as we introduced in Example 4. It is easy to see that* $(\mathcal{P}_{i,j}, \ell)$ *is e.a.s.-learnable (in fact it is* η*-predictable for all* $\eta > 0$*). Moreover, the stopping rule for each* $\mathcal{P}_{i,j}$ *stops finitely almost surely on all sources in* $\mathcal{P}$*. However,* $(\mathcal{P}, \ell)$ *is not e.a.s.-predictable. This follows from Example 4 and the fact that the set of irrational numbers in* $[0,1]$ *is not of Baire first category (Rudin 2006, Chapter 2.1).*

4.4.2 Characterization of *e.a.s.*-learnability

The characterization of *e.a.s.*-learnability is captured by the notion of *identifiability* as follows.

Definition 7. *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 exists a 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* $p \in \mathcal{U} \backslash \mathcal{V}$*.*

In other words, the rule almost surely stops on sources in $\mathcal{V}$, but does not stop on sources in $\mathcal{U}\backslash\mathcal{V}$ with the prescribed confidence. If we want to distinguish whether a source is in $\mathcal{V}$ (the null hypothesis) or in $\mathcal{U}\backslash\mathcal{V}$ (alternate), we can do so with asymptotically zero type-I error and arbitrarily small type-II error, hence the term identifiability.

Example 6. *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 $\frac{2\log(2^{n+1}/\eta)}{\epsilon_n^2}$ 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 7. 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.

Theorem 9. *Let $\mathcal{U}$ be a collection of i.i.d. processes over $\mathbb{N}^\infty$, $\mathcal{V} \subset \mathcal{U}$. Then $\mathcal{V}$ is identifiable in $\mathcal{U}$ iff the marginals of $\mathcal{V}$ are relatively open in the marginals of $\mathcal{U}$ under total variation distance.*

Proof. Suppose there is a limit point p of $\mathcal{U}\backslash\mathcal{V}$ in $\mathcal{V}$, we show that $\mathcal{V}$ is not identifiable in $\mathcal{U}$. Let $p_1, p_2, \cdots, p_n, \cdots \in \mathcal{U}\backslash\mathcal{V}$ such that $||p_n - p||_{TV} \to 0$ as $n \to \infty$, where we have abused the notation p, p_n to denote the marginal distributions of processes p, p_n as well. For any $\eta \leq \frac{1}{4}$, we denote τ_η as the stopping rule in Definition 7 that identifies $\mathcal{V}$ in $\mathcal{U}$. Let A_m be the event that τ_η stops before step m. We have $p(A_m) \to 1$ as $m \to \infty$. Taking M be a number such that $p(A_M) \geq \frac{3}{4}$. Now, since $||p_n - p||_{TV} \to 0$, we have some p_N such that $||p_N - p||_{TV} \leq \frac{1}{4M}$. Therefore, we have $|p(A_M) - p_N(A_M)| \leq \frac{1}{4}$, where we have used the inequality $||p^M - p_N^M||_{TV} \leq M||p - p_N||_{TV}$ if p^M, P_N^M are the M-fold *i.i.d.* distributions of p, p_N respectively. However, this would imply that τ_η stops on p_N finitely with probability $\geq \frac{3}{4} - \frac{1}{4} \geq \frac{1}{4}$, a contradiction.

Suppose now there are no limit points of $\mathcal{U}\backslash\mathcal{V}$ in $\mathcal{V}$. We know that for any point $p \in \mathcal{V}$, there exists a open ball $\mathcal{B}_p$ of p such that $t_p \triangleq \inf\{||p - q||_1 : p \in \mathcal{B}_p \text{ and } q \in \mathcal{U}\backslash\mathcal{V}\} > 0$.

Since the topological space of distributions over $\mathbb{N}$ is separable, there exist countably many distributions $p_1, p_2, \cdots \in \mathcal{V}$ such that $\mathcal{V} = \bigcup_{i\in\mathbb{N}} \mathcal{B}_{p_i}$. Let $\mathcal{V}_n = \bigcup_{i=1}^{n} \mathcal{B}_{p_i}$, we show that $\mathcal{V}_n$ can be distinguished with $\mathcal{U}\backslash\mathcal{V}$ with arbitrary high confidence by observing bounded samples. We denote r_i to be the radius of ball $\mathcal{B}_{p_i}$ and $t_i = t_{p_i}$. Denote $n_i = \min\{n : p_i(X \geq n) \leq t_i/4\}$. Let q be the underlying distribution, we can estimate the empirical frequency $\hat{q}$ of q on $\{1, 2, \cdots, n_i\}$ with error at most $t_i/4$ and arbitrary high confidence using bounded samples. The key observation is that, for any $q \in \mathcal{U}\backslash\mathcal{V}$ we have $||q - p_i||_1 \geq r_i + t_i$. Meaning that we can distinguish distributions from $\mathcal{B}_{p_i}$ and $\mathcal{U}\backslash\mathcal{V}$ by checking whether $||\hat{q} - p_i||_1 \leq r_i + t_i/2$ with arbitrary high confidence using bounded samples. A similar argument as in Example 6 completes the proof. □

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

Theorem 10. *Let $\mathcal{P}$ be a collection of probability measures. Then $(\mathcal{P}, \ell)$ is e.a.s.-learnable if for all $\eta > 0$ there exists an η-nesting $\mathcal{P} = \bigcup_{n\geq 1} \mathcal{P}_n^\eta$, where in addition for all n, $\mathcal{P}_n^\eta$ 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 a 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 index of the stopping rule that stops earliest. 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 probability 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 *e.a.s.*-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 *e.a.s.*-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 of length n with probability $\leq \eta/2$ by construction, namely, $\mathcal{P}_n$ is η-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 $\geq 1-\frac{\delta}{2^m}$ and error $\leq e_m/4$, 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 exist 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. □

The tight characterization for *i.i.d.* sources can not be extended to arbitrary sources. We provide an example below that shows that the characterization in Theorem 10 for *i.i.d.*sources will not extend to Markov processes with even two states.

Example 7. *We consider the Markov processes 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 $\ell(p, X_1^n, Y_n) = 0$ *if there exists* $k \leq n$ *such that* $X_k = 1$. *Else the loss is* 1. *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 10 does not exist for $(\mathcal{P}, \ell)$. *Suppose otherwise, we have a decomposition* $\{\mathcal{P}_n\}_{n\geq 1}$ *such that each* $(\mathcal{P}_n, \ell)$ *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 7, we know that any τ must stop 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 $\mathcal{P}_k$ 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 7 is possible is because the number of states of the process is not known *a-priori* (either 1 or 2). Indeed, with arguments similar to the necessity part of Theorem 10 in the *i.i.d.* case, we can show that if $\mathcal{P}$ is a collection of irreducible finite state Markov processes with the same number of states then the necessary condition in Theorem 10 still holds. We should also emphasize that the stopping rule derived from the sufficient condition of Theorem 10 is not meant to be optimal. For specific problems, there will often be more natural stopping rules.

4.5 Variations

In addition to *e.a.s.*-predictability and *e.a.s.*-learnability that we established in the previous sections, we define the natural extensions of weakly *e.a.s.*-predictability and prediction with finite expected loss in this section, and study their relationships with *e.a.s.*-predictability.

4.5.1 Weakly *e.a.s.*-predictable

Definition 8. *A collection $(\mathcal{P}, \ell)$ is said to be* weakly *e.a.s.-predictable if for any $\eta > 0$ there exists a prediction rule Φ^η, such that for all $p \in \mathcal{P}$*

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

It is easy to see that the characterization in Theorem 1 holds for weakly *e.a.s.*-predictability as well.

Theorem 11. *If a collection $(\mathcal{P}, \ell)$ is weakly e.a.s.-predictable then for all $\eta > 0$ there exists a η-nesting of $\mathcal{P}$.*

Conversely, if there exists a universal nesting of $\mathcal{P}$, then $(\mathcal{P}, \ell)$ is weakly e.a.s.-predictable.

Proof. The sufficiency follows immediately from Theorem 1, since *e.a.s.*-predictability implies weakly *e.a.s.*-predictability. To prove the necessary condition, one could replicate the same argument as in the proof of Theorem 1. □

The following example shows that weakly *e.a.s.*-predictable does not imply *e.a.s.*-predictable.

Example 8. *For any $m \in \mathbb{N}$ we define a set $S_m \subset \mathbb{N}$ such that $|S_m| = m$, and we require $S_m \cap S_{m'} = \emptyset$ for all $m \neq m' \in \mathbb{N}$. Clearly, such a family of sets exists. Now, for any sequences $\boldsymbol{a} = (a_1, a_2, \cdots) \in \mathbb{N}^\infty$ and $\boldsymbol{b} = (b_1, b_2, \cdots) \in \mathbb{N}^\infty$ with $b_m \in S_m$ for all $m \in \mathbb{N}$, we define a source $p^{(\boldsymbol{a},\boldsymbol{b})}$ as follows.*

1. *The process $p^{(\boldsymbol{a},\boldsymbol{b})}$ has* no *outcome, i.e. the learner can't observe any sample;*
2. *The loss $\ell(p^{(\boldsymbol{a},\boldsymbol{b})}, X_1^n, Y_n) = 0$ if and only if there exist some $m \in \mathbb{N}$ such that $n \geq a_m$, $Y_n \in S_m$ and $Y_n \neq b_m$.*

Let $\mathcal{P}$ be the class of all such sources. We now show that there exist randomized prediction rules that achieve weakly e.a.s.-predictability for $(\mathcal{P}, \ell)$. For any $\eta > 0$, the predictor Φ^η is defined as follows. Let M be a number such that $M \geq \frac{1}{\eta}$. We select a number $c \in S_M$ uniform randomly at the beginning of the game, and define $\Phi^\eta(X_1^{n-1}) = c$ for all $n \geq 1$. Clearly, there exists some time step N such that $N \geq a_M$, and the probability that we make errors after step N is at most $\frac{1}{M} \leq \eta$.

We now show that any randomized predictor Φ can't achieve e.a.s.-predictability for $(\mathcal{P}, \ell)$. Let Ω be the probability space of the internal randomness of Φ, i.e., for any $\omega \in \Omega$, Φ^ω is a deterministic strategy. Let μ be the corresponding probability measure over Ω.

Let $c_1, c_2, \cdots$ be a sequence of numbers such that the probability of Φ making predictions in $\bigcup_{i \leq m} S_i$ after step c_m is upper bounded by $\frac{\eta}{2^m}$ for some $\eta < 1$. We show that such sequences exist. Otherwise, there exists some m such that Φ makes predictions on some

number $s \in S_m$ infinite often with positive probability. However, this will violate the e.a.s.-predictability of Φ, since it will make infinite errors for the source that takes $b_m = s$.

Let A be the event that, Φ makes no prediction on $\bigcup_{i \leq m} S_i$ after step c_m for all $m \in \mathbb{N}$. By union bound, we have $\mu(A) \geq 1 - \eta$. We now take $a_m = c_m$ for all $m \in \mathbb{N}$. And let b_m to be choosing uniform randomly from S_m and independently for different m. Denote Γ to be the probability space of the selection processes, and ν be the corresponding probability measure. For any $\gamma \in \Gamma$ we denote p^γ to be the source that is selected from the process.

For any $(\omega, \gamma) \in \Omega \times \Gamma$, we define $1\{\omega, \gamma\}$ to be the indicator that Φ^ω makes infinite errors on p^γ. We claim that for any $\omega \in A$ we have

$$\mathbb{E}_{\gamma \sim \nu} 1\{\omega, \gamma\} = 1.$$

W.o.l.g., we can assume that Φ^ω makes predictions only on S_{m+1} during phase m of steps from $a_m + 1$ to a_{m+1}. Now, we have the probability of making errors at phase m is at least $\frac{1}{m}$. Since the b_m is selected independently, we know that the errors at different phases are independent. The claim follows by the converse Borel-Cantelli lemma.

Now, we have shown that

$$\mathbb{E}_{\omega \sim \mu}[\mathbb{E}_{\gamma \sim \nu} 1\{\omega, \gamma\}] \geq 1 - \eta.$$

By switching order of expectation, we have

$$\mathbb{E}_{\gamma \sim \nu} \mathbb{E}_{\omega \sim \mu} 1\{\omega, \gamma\} \geq 1 - \eta.$$

This implies that there exists some γ such that the probability of Φ making infinite errors on p^γ is at least $1 - \eta > 0$. This violates the e.a.s.-predictability of Φ.

Remark 4. *The above example shows that if the prediction rule is randomized, then weakly e.a.s.-predictability does not imply e.a.s.-predictability. However, it is still open whether such a separation exists if the prediction rule is required to be deterministic.*

4.5.2 Prediction with finite expected loss

Definition 9. *A collection $(\mathcal{P}, \ell)$ is said to be predictable with finite expected loss if there exists a prediction rule Φ, such that for all $p \in \mathcal{P}$*

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

By Borel-Cantelli lemma, it is clear that if a collection is predictable with finite expected loss, then it is also *e.a.s.*-predictable. The following example shows that the converse does not hold in general.

Example 9. *Let $\mathcal{P}$ be the class with a single source p defined as follows. Let U be the uniform distribution over $[0, 1]$, we define*

$$\forall n \in \mathbb{N},\ X_n = U,$$

i.e. all the X_n have the same value that is sampled from U. The loss is defined as follows

$$\ell(p, X_1^n, Y_n) = 1 \text{ if and only if } X_n \leq \frac{1}{n}.$$

Clearly, $(\mathcal{P}, \ell)$ is e.a.s.-predictable, since with probability 1 we have $U \neq 0$, meaning that we will make no error after step $\lceil \frac{1}{U} \rceil$. However, the collection is not predictable with finite expected loss. To see this, we define A_n to be the event that an error occurred at step n. We have

$$\mathbb{E}_p\left[\sum_{n=1}^{\infty} \ell(p, X_1^n, \Phi(X_1^{n-1}))\right] = \sum_{n=1}^{\infty} \mathbb{E}_p[A_n] = \sum_{n=1}^{\infty} \frac{1}{n} = \infty.$$

Clearly, if a collection $(\mathcal{P}, \ell)$ is predictable with finite expected loss, then there exists a nesting $\{\mathcal{P}_i, i \geq 1\}$ of $\mathcal{P}$ such that $\forall i \geq 1$ $(\mathcal{P}_i, \ell)$ is predictable with *bounded* expected loss. To do so, we can simply define

$$\mathcal{P}_i = \{p \in \mathcal{P} : p \text{ has expected loss} \leq i \text{ under prediction rule } \Phi\},$$

where Φ is a prediction rule that achieves finite expected loss on $(\mathcal{P}, \ell)$. However, we are unaware if the converse is true or not. We have the following open problem.

Problem 1 (Open Problem)**.** *Suppose that there is a nesting $\{\mathcal{P}_i, i \geq 1\}$ for a class $\mathcal{P}$, such that $\forall i \geq 1$ $(\mathcal{P}_i, \ell)$ is predictable with bounded expected loss for the same loss function ℓ. Is the collection $(\mathcal{P}, \ell)$ predictable with finite expected loss?*

Remark 5. *We say a collection $(\mathcal{P}, \ell)$ is predictable with bounded expected* tail *loss if there exists a prediction rule Φ, function $\rho : \mathbb{N}^+ \to \mathbb{R}^+$ and constant N, such that for all $p \in \mathcal{P}$ and $n \geq N$, we have*

$$\mathbb{E}_p \left[\sum_{k=n}^{\infty} \ell(p, X_1^k, \Phi(X_1^{k-1})) \right] \leq \rho(n),$$

and $\rho(n) \to 0$ as $n \to \infty$. It is not hard to show that the Problem 1 is true if we assume $\forall i \geq 1$ the collection $(\mathcal{P}_i, \ell)$ is predictable with bounded expected tail loss. This is satisfied, e.g., by the setting in Theorem 6.

Note that the Problem 1 is non-trivial even when the loss is assumed to be supervised. This can be seen as follows. Suppose the underlying source is $p \in \mathcal{P}_t$. We denote Φ_t to be the predictor for $\mathcal{P}_t$ that achieves finite expected errors. For any realization $\mathbf{x} \in \mathcal{X}^\infty$, suppose the number of errors on $\mathbf{x}$ by Φ_t is s. By the structural risk minimization argument as outlined in the proof of Theorem 5, we know that the number of errors of the aggregated predictor Φ will $\leq (s+t)^2$. Our assumption only guarantees that $\mathbb{E}[S] < \infty$, where S is the random counterpart of s when $\mathbf{x} \sim p$. However, this *does not* imply that $\mathbb{E}[S^2] < \infty$ in general.

4.6 Summary

The author is the primary contributor for the work in this chapter. The work in Section 4.1, Section 4.2, Section 4.3 and Section 4.4 has been partially published in (Wu and Santhanam 2021b), and the work in Section 4.5 has not yet been published.

Chapter 5

Hypothesis Testing

In this chapter we consider hypothesis testing using *i.i.d.* samples. For any two disjoint collections $\mathcal{H}_1, \mathcal{H}_2$ of distributions over $\mathbb{R}^d$, the *hypothesis testing* problem is to decide whether some underlying source $p \in \mathcal{H}_1 \cup \mathcal{H}_2$ is from $\mathcal{H}_1$ or $\mathcal{H}_2$, by observing *i.i.d.* samples from p. Therefore, using the notations we developed in Chapter 4, the class under consideration will include the *i.i.d.* processes in $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2$, and for any $p \in \mathcal{H}$ the loss is defined to be

$$\ell(p, X_1^n, Y_n) = 1\{p \notin \mathcal{H}_{Y_n}\}.$$

With slight abuse of notation, for any distribution p over $\mathbb{R}^d$, we will use p to denote the *i.i.d.* process of p as well, and we denote p^n to be the n-fold *i.i.d.* distribution of p.

This problem was considered extensively in (Cover 1973; Dembo and Peres 1994). In Section 5.2, we first revisit and strengthen the topological characterizations of Dembo and Peres (1994) for the almost sure hypothesis testing setup. Informally, Dembo and Peres (1994) showed that the existence of $e.a.s.$-prediction rule for a hypothesis testing problem is *essentially* captured by the notion of what they call F_σ-separability. Roughly speaking, two sets of a metric space are said to be F_σ-separable if they are contained in disjoint F_σ-sets, i.e., sets that can be written as countable unions of closed sets.

We provide a different characterization of the F_σ-separability in Lemma 11 through the concept of nestings. Using such a characterization and our general results in Chapter 4, we recover all the main results in Dembo and Peres (1994) with simple and elementary proofs.

Moreover, we provide a partial resolution of an open problem of Dembo and Peres (1994) in Corollary 2. We then go beyond the characterization of Dembo and Peres (1994) by providing a complete characterization for the almost sure hypothesis testing in Theorem 14 when the support of the distributions is $\mathbb{N}$.

In Section 5.2.2, we study testing of entropy properties. Under mild conditions, we completely characterize the *e.a.s.*-predictability for testing entropy properties through a notion of dominance of tail entropy. The main technique for establishing such a result is our alternative characterization of F_σ-separability in Lemma 11 and our general characterizations in Chapter 4. We note that the law of iterated logarithm based techniques, including that of Dembo and Peres (1994) seem unlikely to establish such a result.

Using our general topological characterizations, we provide several concrete examples for testing the singularity, rank, and multiplicity of eigenvalues of random matrices in Section 5.3.

5.1 Preliminaries

For any two distributions p_1, p_2 over $\mathbb{R}^d$, the Kolmogorov-Smirnov distance (abbreviate as KS-distance) is defined to be

$$|p_1 - p_2|_{KS} = |F_{p_1}(\mathbf{x}) - F_{p_2}(\mathbf{x})|_{KS} \stackrel{\text{def}}{=} \sup_{\mathbf{x}\in\mathbb{R}^d} |F_{p_1}(\mathbf{x}) - F_{p_2}(\mathbf{x})|,$$

where F_{p_i} is the CDF of p_i for $i \in \{1, 2\}$.

Clearly, the KS-distance satisfies the triangle inequality. To see this, we have

$$\begin{aligned}
|p_1 - p_2|_{KS} + |p_2 - p_3|_{KS} &= \sup_{\mathbf{x}\in\mathbb{R}^d} |F_{p_1}(\mathbf{x}) - F_{p_2}(\mathbf{x})| + \sup_{\mathbf{x}\in\mathbb{R}^d} |F_{p_2}(\mathbf{x}) - F_{p_3}(\mathbf{x})| \\
&\geq \sup_{x\in\mathbb{R}^d} |F_{p_1}(\mathbf{x}) - F_{p_2}(\mathbf{x})| + |F_{p_2}(\mathbf{x}) - F_{p_3}(\mathbf{x})| \\
&\geq \sup_{x\in\mathbb{R}^d} |F_{p_1}(\mathbf{x}) - F_{p_3}(\mathbf{x})| = |p_1 - p_3|_{KS}.
\end{aligned}$$

The following lemma is well known in the literature, which relates convergence under KS-distance with weak convergence (see Chapter 3.2.1 for definition), see e.g. (Athreya and Lahiri 2006, Theorem 9.1.4).

Lemma 9 (Polya's Theorem)**.** *Let $p_1, p_2, \cdots$ and p be distributions over $\mathbb{R}^d$ that are absolutely continuous w.r.t. Lebesgue measure. Then $\lim_{n\to\infty} |p_n - p|_{KS} = 0$ iff p_n weakly converges to p.*

Clearly, if $\lim_{n\to\infty} |p_n - p|_{KS} = 0$ then p_n weakly converges to p. However, the converse is not necessarily true if the CDF of p is not continuous. To see this, let p_n be the distribution over $\mathbb{R}$ that assigns mass $1 - \frac{1}{n}$ on $-\frac{1}{n}$ and mass $\frac{1}{n}$ on $\frac{1}{n}$, and let p be the distribution over $\mathbb{R}$ that assigns mass 1 on 0. We know that p_n weakly converges to p but $\lim_{n\to\infty} |p_n - p|_{KS} = 1$.

For any $\mathbf{x} \in \mathbb{R}^d$, we denote $\mathbb{I}_{\mathbf{x}}$ as the indicator function of set $\prod_{i=1}^{d}(-\infty, \mathbf{x}_i]$, where $\mathbf{x}_i$ is the i'th coordinate of $\mathbf{x}$. Let $X_1, X_2, \cdots, X_n$ be *i.i.d.* random variables over $\mathbb{R}^d$, denote the CDF of the empirical distribution as follows

$$\forall \mathbf{x} \in \mathbb{R}^d,\ F_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\mathbf{x}}(X_i).$$

The following lemma shows that the empirical distribution uniformly estimates any distribution under KS-distance, and is known as Dvoretzky-Kiefer-Wolfowitz Inequality, see e.g., Massart (1990); Kiefer and Wolfowitz (1958); Naaman (2021).

Lemma 10 (Dvoretzky-Kiefer-Wolfowitz Inequality)**.** *Let $X_1, X_2, \cdots, X_n$ be* i.i.d. *samples of distribution p over $\mathbb{R}^d$, $F_n(\boldsymbol{x})$ is the CDF of the empirical distribution. Then there exists a constant C_d depends only on d such that*

$$p\left(|F_n(\boldsymbol{x}) - F_p(\boldsymbol{x})|_{KS} \geq \epsilon\right) \leq C_d \exp(-n\epsilon^2).$$

5.2 Topological criterion: revisited

Let A, B be two sets in some metric space with metric d. We say A, B are *F_σ-separable* if $A \subset A'$ and $B \subset B'$, $A' \cap B' = \emptyset$ and both A' and B' can be written as countable unions of closed sets (with topology induced by d). Such a notion was used by Dembo and Peres (1994) to characterize the *e.a.s.*-predictability of the hypothesis testing problems as we described in the beginning of this chapter. One of the main results of (Dembo and Peres 1994) is stated below using our notations in Chapter 4.

Theorem 12 ((Dembo and Peres 1994, Theorem 2))**.** *Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be disjoint collections of distributions over $\mathbb{R}^d$ that are absolutely continuous w.r.t. Lebesgue measure. Denote $\mathcal{H} = \mathcal{H}_1 \cup \mathcal{H}_2$ and let ℓ be the loss as defined in beginning of this chapter. We have*

1. *If $\mathcal{H}_1$ and $\mathcal{H}_2$ are F_σ-separable under the weak convergence topology, then $(\mathcal{H}, \ell)$ is e.a.s.-predictable.*
2. *If for any $p \in \mathcal{H}$, there exists $r > 1$ (possibly depends on p) such that*

$$\int_{\mathbb{R}^d} f_p^r(\boldsymbol{x}) d\mu < \infty,$$

where f_p is the density of p and μ is Lebesgue measure. Then $(\mathcal{H}, \ell)$ is e.a.s.-predictable iff $\mathcal{H}_1$ and $\mathcal{H}_2$ are F_σ-separable under the weak convergence topology.

Dembo and Peres (1994) asked whether the condition $r > 1$ in Theorem 12 can be removed. We now give a positive answer to this open problem by showing that the condition of the finite rth norm of the density can be relaxed to a much weaker condition as defined in the following.

Definition 10. *Let $\mathcal{H}$ be a collection of distributions over $\mathbb{R}^d$ that are absolutely continuous w.r.t. Lebesgue measure. We say $\mathcal{H}$ is* uniformly bounded *if for any $\epsilon > 0$ there exists a number M_ϵ such that*

$$\forall p \in \mathcal{H},\ p(f_p(\boldsymbol{x}) \geq M_\epsilon) \leq \epsilon,$$

where f_p is the density of p.

Before showing why such a condition is sufficient to establish Theorem 12(2), we first provide a simple alternative proof to Theorem 12(1) by using our general characterizations in Chapter 4. To begin with, we first prove the following lemma, which relates F_σ-separability with the nesting of sets.

Lemma 11. *Let A, B be two sets of some metric space with metric d. The A, B are F_σ-separable iff there exist nestings $\{A_i, i \geq 1\}$ and $\{B_i, i \geq 1\}$ for A, B respectively. such that*

$$\forall i \geq 1, \ \inf\{d(x, y) : x \in A_i, y \in B_i\} > 0.$$

Proof. By definition of F_σ-separability, we have collections of *closed* sets $\{A'_i\}$ and $\{B'_i\}$ such that $A \subset \bigcup_i A'_i$ and $B \subset \bigcup_i B'_i$. Since finite unions of closed sets are closed, we may assume $\{A'_i\}$, $\{B'_i\}$ to be nested. Define

$$A_i = \{x : x \in A'_k \text{ and } d(x, B'_k) \geq 1/i, \ k \in \mathbb{N}\}$$

and

$$B_i = \{y : y \in B'_k \text{ and } d(y, A'_k) \geq 1/i, \ k \in \mathbb{N}\}.$$

Since A'_i, B'_i are closed, we have $A \subset \bigcup_i A_i$ and $B \subset \bigcup_i B_i$. For any $x \in A_i$ and $y \in B_i$, we show that $d(x, y) \geq 1/i$, thus proving the necessary condition. By definition, we have $x \in A'_{k_1}$, $d(x, B'_{k_1}) \geq 1/i$ and $y \in B'_{k_2}$, $d(y, A'_{k_2}) \geq 1/i$ for some $k_1, k_2 \in \mathbb{N}$. W.l.o.g., we may assume $k_1 \leq k_2$. Since the collections are nested, we have $x \in A'_{k_2}$. Now, since $d(y, A'_{k_2}) \geq 1/i$, we have $d(x, y) \geq 1/i$.

To prove the sufficiency, it is sufficient to show that

$$\bigcup_{i \geq 1} \bar{A}_i \cap \bigcup_{i \geq 1} \bar{B}_i = \emptyset,$$

where $\bar{S}$ denotes the closure of set S under metric d. Otherwise, there will be some $x \in \bar{A}_i$ and $x \in \bar{B}_j$ for some i, j. W.l.og., we can assume $j \geq i$. By nesting property, we now have $x \in \bar{A}_j$ as well. However, this will imply that $d(A_j, B_j) = 0$, a contradiction. □

Remark 6. *Note that, the closeness condition of A'_i, B'_i in the proof of Lemma 11 can be replaced with the condition that there is no limit point of A'_i in B'_i (and vice versa).*

The following lemma will be useful in our following analysis.

Lemma 12. *Let $\{\{A_{i,j}\}_{i\in\mathbb{N}}, \{B_{i,j}\}_{i\in\mathbb{N}}\}_{j\in\mathbb{N}}$ be nesting such that $A_{i,j} \subset A_{i',j'}$ and $B_{i,j} \subset B_{i',j'}$ whenever $i' \geq i$ and $j' \geq j$. If $d(A_{i,j}, B_{i,j}) > 0$ for all $i, j \in \mathbb{N}$, where d is some metric. Then $A = \bigcup_{i,j\in\mathbb{N}} A_{i,j}$ and $B = \bigcup_{i,j\in\mathbb{N}} B_{i,j}$ are F_σ-separable.*

Proof. We only need to show that for all i, j, B and A contain no limit point of $A_{i,j}$ and $B_{i,j}$ respectively. Suppose otherwise that there exists some $y \in B$ such that $d(y, A_{i,j}) = 0$. We have $y \in B_{i',j'}$ for some i', j'. By nesting property, we may assume $i = i'$ and $j = j'$. However, this will imply $d(B_{i,j}, A_{i,j}) = 0$, a contradiction. □

Proof of Theorem 12(1). By Lemma 11, we have nestings $\{A_i, \geq 1\}$, $\{B_i, i \geq 1\}$ of $\mathcal{H}_1$, $\mathcal{H}_2$ respectively, such that

$$\forall i \in \mathbb{N},\ \inf\{|p_1 - p_2|_{KS} : p_1 \in A_i, p_2 \in B_i\} \geq \frac{1}{i}.$$

By Theorem 1, we only need to show that $(A_i \cup B_i,\ \ell)$ is η-predictable for all $i \geq 1$ and $\eta > 0$. Let F_n be the CDF of the empirical distribution with sample of size n. By Lemma 10, we can simultaneously make $|F_n(\mathbf{x}) - F_p(\mathbf{x})|_{KS} \leq 1/4i$ with confidence $\geq 1-\eta$ for all $p \in A_i \cup B_i$ by choosing the sample size n large enough. By triangle inequality of KS-distance, one can classify the distributions in $A_i \cup B_i$ successfully with probability at least $1-\eta$, by predicting the class that is closer to F_n under KS-distance.

The theorem follows by observing that convergence under KS-distance is equivalent to weak convergence with the absolutely continuous assumption by Lemma 9. □

Note that one may view the subclasses $\{A_i \cup B_i\}$ to be regularized subproblems that can be handled with bounded number of samples.

We now show that if the class $\mathcal{H}_1 \cup \mathcal{H}_2$ is uniformly bounded as defined in Definition 10, then the F_σ-separability will be necessary to achieve *e.a.s.*-predictability under the assumption of Theorem 12.

Theorem 13. *Let $\mathcal{H}_1, \mathcal{H}_2$ be collections of distributions that are absolutely continuous w.r.t. Lebesgue measure on $\mathbb{R}^d$, and $\mathcal{H}_1 \cup \mathcal{H}_2$ is* uniformly bounded. *Then $(\mathcal{H}_1 \cup \mathcal{H}_2,\ \ell)$ is e.a.s.-predictable if and only if $\mathcal{H}_1, \mathcal{H}_2$ are F_σ-separable under KS-distance.*

Proof. The sufficiency follows directly from Theorem 12(1). We now only prove the necessity.

By Theorem 1, *e.a.s.*-predictability implies that there exist nesting $\{A_i, i \geq 1\}$, $\{B_i, i \geq 1\}$ of $\mathcal{H}_1$, $\mathcal{H}_2$ respectively, such that for all $i \geq 1$, $(A_i \cup B_i,\ \ell)$ is $\frac{1}{8}$-predictable. By Lemma 11, it is sufficient to show that for all $i \geq 1$, there is no limit point of A_i in B_i or vice versa. Suppose otherwise, there exist $p_1, p_2, \cdots \in A_i$ and $p \in B_i$, such that $|p_n - p|_{KS} \to 0$ as $n \to \infty$. Let Φ be an arbitrary predictor that achieves $\frac{1}{8}$-predictability of $A_i \cup B_i$ with sample size i. Denote Φ^i as the prediction function at step i. By Lemma 9, p_n weakly converges to p. Therefore, there exists a compact set $S \subset \mathbb{R}^{i \cdot d}$, such that $p(S) \geq \frac{7}{8}$ and $p_n(S) \geq \frac{7}{8}$ for all $n \in \mathbb{N}$ by Lemma 6, where we have used p, p_n to also denote i-fold measures of p, p_n over $\mathbb{R}^{i \cdot d}$. By uniform boundedness of $\mathcal{H}_1 \cup \mathcal{H}_2$, there exists a number M, such that

$$\forall p \in \mathcal{H}_1 \cup \mathcal{H}_2,\ p(\{\mathbf{x} \in \mathbb{R}^{i \cdot d},\ f_p(\mathbf{x}) \geq M\}) \leq \frac{1}{16}. \tag{5.1}$$

By Lusin's theorem (Rudin 1974, Theorem 2.24), there exists a continuous function g over $\mathbb{R}^{i \cdot d}$ and a set $E \subset S$, such that $\sup_{\mathbf{x} \in E} |\Phi^i(\mathbf{x}) - g(\mathbf{x})| \leq \frac{1}{4}$ and $\mu(S \backslash E) \leq \frac{1}{16M}$ where $\mu(\cdot)$ is the Lebesgue measure over $\mathbb{R}^{i \cdot d}$. Let $\Omega = \{\mathbf{x} \in \mathbb{R}^{i \cdot d} : g(\mathbf{x}) > \frac{3}{2}\}$, we have Ω is open and $\{\mathbf{x} \in E : \Phi^i(\mathbf{x}) = 2\} \subset \Omega$. By (5.1) and $\mu(S \backslash E) \leq \frac{1}{16M}$, we have $p(S \backslash E) \leq \frac{1}{8}$ and $p_n(S \backslash E) \leq \frac{1}{8}$ for all $n \in \mathbb{N}$. By $\frac{1}{8}$-predictability, we have $p(\Omega) \geq p(\Omega \cap E) \geq \frac{7}{8} - \frac{1}{4} = \frac{5}{8}$, since $p(\bar{E}) \leq \frac{1}{4}$. By weak convergence, we have $\liminf p_n(\Omega) \geq p(\Omega)$, since Ω is open

(see Lemma 5 in Chapter 3.2.1). There exists some p_n such that $p_n(\Omega) \geq \frac{1}{2}$, which implies $p_n(\Omega \cap E) \geq \frac{1}{2} - \frac{1}{4} = \frac{1}{4} > \frac{1}{8}$, contradicting the $\frac{1}{8}$-predictability. □

Remark 7. *Note that, the assumption of uniform boundedness on $\mathcal{H}_1 \cup \mathcal{H}_2$ is necessary for the proof of Theorem 13 to work. To see this, let p'_n be the uniform distribution over $\{\frac{1}{n}, \frac{2}{n}, \cdots, \frac{n-1}{n}\}$, and p be the uniform distribution over $[0,1]$. Clearly, we have that p'_n converges to p in distribution. But p'_n is not continuous. This can be easily resolved by choosing some* continuous *distribution p_n such that $|p_n - p'_n|_{KS} \leq \frac{1}{n}$ and there exists a set S_n concentrated on $\{\frac{1}{n}, \frac{2}{n}, \cdots, \frac{n-1}{n}\}$ so that we have $p_n([0,1] \backslash S_n) \leq \frac{\eta}{2^n}$ and $p(S_n) \leq \frac{\eta}{2^n}$. Let $A = \{p_1, p_2, \cdots\}$ and $B = \{p\}$, we have p_n converges to p in distribution. However, since*

$$\sum_{n=1}^{\infty} p(S_n) \leq \eta$$

we have $(A \cup B, \ell)$ is η-predictable with only one sample by predicting the underlying distribution in A if the sample is in $\bigcup S_n$ and in B otherwise.

We have the following corollary.

Corollary 2. *Let $G : \mathbb{R}^+ \to \mathbb{R}^+$ be a monotone increasing function such that $\lim_{x\to\infty} G(x) \to \infty$, $\mathcal{H}_1, \mathcal{H}_2$ be distributions over $\mathbb{R}^d$ that are absolutely continuous w.r.t. Lebesgue measure. If for all $p \in \mathcal{H}_1 \cup \mathcal{H}_2$, we have $\mathbb{E}_{\boldsymbol{x}\sim p}[G(f_p(\boldsymbol{x}))] < \infty$, where f_p is the density of p. Then $(\mathcal{H}_1 \cup \mathcal{H}_2,\ \ell)$ is e.a.s.-predictable iff $\mathcal{H}_1, \mathcal{H}_2$ are F_σ separable with KS-distance.*

Proof. By Theorem 12 we only need to prove the necessary condition. By Lemma 11, Lemma 12 and breaking $\mathcal{H}_1 \cup \mathcal{H}_2$ into countably subcollections, one may assume, w.l.o.g., $\forall p \in \mathcal{H}_1 \cup \mathcal{H}_2$, $\mathbb{E}_{\mathbf{x}\sim p}[G(f_p(\mathbf{x}))] \leq M$ for some constant M. By Theorem 13, we only need to show that $\mathcal{H}_1 \cup \mathcal{H}_2$ is uniformly bounded. For any $p \in \mathcal{H}_1 \cup \mathcal{H}_2$, we define random variable $Y_p = G(f_p(\mathbf{x}))$. We have, by Markov inequality, $p(Y_p \geq T) \leq \frac{M}{T}$. Note that the upper bound is independent of p. By letting $T = \frac{M}{\epsilon}$, one can make the probability upper bounded

by ϵ. Since G is monotone increasing and goes to infinity, thus invertible on $\mathbb{R}^+$. We now have $p(f_p(\mathbf{x}) \geq G^{-1}(M/\epsilon)) \leq \epsilon$ for all $p \in \mathcal{H}_1 \cup \mathcal{H}_2$ and $\epsilon > 0$. □

Remark 8. *Note that, Theorem 12(2) follows directly from Corollary 2 by simply taking $G(x) = x^{r-1}$ for some $r > 1$. However, the conditions in Corollary 2 are much weaker, since $G(x)$ can be an arbitrary monotone function so long as $G(x) \to \infty$ when $x \to \infty$. One may, e.g., take $G(x) = \log(x)$, which is asymptotically upper bounded by x^{r-1} for all $r > 1$.*

We now have the following conjecture, which asserts that a condition such as the uniform boundedness is necessary for Theorem 12(2) to hold.

Conjecture 1. *There exist collections $\mathcal{H}_1, \mathcal{H}_2$ of distributions over $\mathbb{R}^d$ that are absolutely continuous w.r.t. Lebesgue measure, such that $(\mathcal{H}_1 \cup \mathcal{H}_2,\ \ell)$ is e.a.s.-predictable but $\mathcal{H}_1, \mathcal{H}_2$ are* not *F_σ-separable under KS-distance.*

Note that, even though the characterization of hypothesis testing for continuous distributions is involved, we have the following full characterization for discrete distributions.

Theorem 14. *Let $\mathcal{H}_1, \mathcal{H}_2$ be collections of distributions over $\mathbb{N}$, then $(\mathcal{H}_1 \cup \mathcal{H}_2, \ell)$ is e.a.s.-predictable if and only if $\mathcal{H}_1, \mathcal{H}_2$ are F_σ-separable under total variation distance.*

We first prove a general lemma.

Lemma 13. *Let $\mathcal{H}_1, \mathcal{H}_2$ be classes of distributions over $\mathbb{R}^d$. If $(\mathcal{H}_1 \cup \mathcal{H}_2, \ell)$ is η-predictable for some $\eta \leq \frac{1}{4}$, then*

$$\inf\{||p_1 - p_2||_{TV} : p_1 \in \mathcal{H}_1,\ p_2 \in \mathcal{H}_2\} > 0.$$

Proof. Let i be the sample complexity for $(\mathcal{H}_1 \cup \mathcal{H}_2, \ell)$ to achieve η-predictability, i.e., there exists a predictor such that the probability of making errors after step i is upper bounded by η for all $p \in \mathcal{H}_1 \cup \mathcal{H}_2$. Suppose the lemma does not hold. We can find $p_1 \in \mathcal{H}_1$ and $p_2 \in \mathcal{H}_2$, such that

$$||p_1 - p_2||_{TV} \leq \frac{1}{4i}.$$

We have, by Lemma 3, $||p_n^i - p^i||_{TV} \leq \frac{1}{4}$. However, this implies that any predictor will make an error at step $i+1$ w.p. $\geq \frac{3}{8} > \frac{1}{4}$ by Lemma 4. This will violate the η-predictability of $(\mathcal{H}_1 \cup \mathcal{H}_2, \ell)$. □

Proof of Theorem 14. We first prove the necessity. By Theorem 1, there exists a nesting $\{\mathcal{G}_i, i \geq 1\}$ for $\mathcal{H}_1 \cup \mathcal{H}_2$ such that $(\mathcal{G}_i, \ell)$ is $\frac{1}{4}$-predictable for all $i \geq 1$. The necessity is followed now by Lemma 13.

To prove the sufficiency, by F_σ-separability, there exist nesting $\{A_i, i \geq 1\}$ and $\{B_i, i \geq 1\}$ for $\mathcal{H}_1$ and $\mathcal{H}_2$ respectively, such that

$$\forall i \geq 1, \ \inf\{||p_1 - p_2||_{TV} : p_1 \in A_i, \ p_2 \in B_i\} \geq \frac{1}{i}.$$

We now show that $(A_i \cup B_i, \ell)$ is *e.a.s.*-learnable (see Definition 6). The theorem will now follow by Theorem 8.

For any $\eta > 0$, the stopping rule τ_η goes as follows. We partition τ_η into two phases, at phase one, τ_η tries to determine if the missing probability mass of the underlying distribution is less than $\frac{1}{4i}$ with confidence at least $1 - \frac{\eta}{2}$. This can be done by requesting additional samples to test if new symbols appear in the new coming samples. At phase two, we will estimate the probability mass only on the symbols we have seen so far at the end of phase one, so that the L_1 error of the estimation is upper bounded by $\frac{1}{4i}$ with confidence $\geq 1 - \frac{\eta}{2}$. This can be done since the symbols we have seen are finite. Now, the prediction at the end of phase two is as follows, if the total variation distance of the estimated distribution is closer to A_i we predict $p \in \mathcal{H}_1$, otherwise, we predict $p \in \mathcal{H}_2$. We will keep the prediction for all steps after phase two of τ_η. It is easy to verify that such a rule satisfies the requirements of *e.a.s.*-learnability and τ_η stops finitely almost surely on all distributions over $\mathbb{N}$. □

Remark 9. *Note that, the proof of Theorem 14 above does not work for distributions over* $\mathbb{R}$. *Since for distribution classes* A, B *over* $\mathbb{R}$, *if*

$$\inf\{||p_1 - p_2||_{TV} : p_1 \in A, \ p_2 \in B\} > 0,$$

we cannot conclude that $(A \cup B, \ell)$ *is e.a.s.-learnable. To see this, let* U *be the uniform distribution over* $[0,1]$*. We define* $A = \{U\}$ *and* $B = \{p \in \mathcal{D} : ||U - p||_{TV} \geq \frac{1}{2}\}$*, where* $\mathcal{D}$ *is the set of distributions over* $[0,1]$*. We now show that* $(A \cup B, \ell)$ *is not e.a.s.-learnable. Suppose otherwise, we have stopping rule* τ *and prediction rule* Φ *for* $(A \cup B, \ell)$ *such that the probability of* Φ *making errors after* τ *has stopped is upper bounded by* $\frac{1}{8}$*, and* τ *stops finitely almost surely on all distributions in* $A \cup B$*. Let* N *be a number such that the probability of* τ *stopping before step* N *and* Φ *making no error after* N *is* $\geq \frac{3}{4}$*. Clearly, such a number exists. We now consider the following sampling procedure. We first sample a set* S *of size* CN^2 *from* U *independently where* C *is a constant, then we generate* N *samples from* S *uniformly. Note that this is equivalent to sample from a mixture of* $\mathcal{S} = \{U(S) : S \subset [0,1],\ |S| = CN^2\}$ *where* $U(S)$ *denotes for the uniform distribution over* S*. Clearly, for any* $p \in \mathcal{S}$ *we have* $||p - U||_{TV} = 1$*, thus* $\mathcal{S} \subset B$*. Now, by birthday-paradox, w.p.* $\geq \frac{3}{4}$ *the samples will have no repeats by choosing* C *large enough. Conditioned on such an event, the statistics of samples from the mixture is exactly the statistics of samples from* U*. Therefore, by union bound, there must be some distribution* $p \in B$ *such that the probability of* Φ *making errors after* τ *has stopped* $\geq \frac{1}{2} > \frac{1}{8}$*. A contradiction.*

Note that, the example above does not imply that Theorem 14 does not hold for distributions over $\mathbb{R}$*. We leave it as an open problem to determine whether Theorem 14 holds for arbitrary distributions over* $\mathbb{R}$*.*

5.2.1 Testing properties of the first moment

Theorem 12 as well as our refinements in Theorem 13 and Theorem 14 provide a general criterion to characterize the *e.a.s.*-predictability for hypothesis testing problem. However, to determine whether two distribution classes $\mathcal{H}_1$, $\mathcal{H}_2$ are F_σ-separable under KS-distance (or total variation distance) is often not trivial in general. In this subsection, we will restrict our attention on the hypothesis testing of properties of the first moment. In particular, for any two sets $A, B \subset \mathbb{R}^d$, we would like to determine whether the underlying distribution

has first moment in A or B with the premise that the first moment is in $A \cup B$. We first introduce the following theorem, which is due to Dembo and Peres (1994).

Theorem 15 ((Dembo and Peres 1994, Theorem 1))**.** *Let $A, B \subset \mathbb{R}^d$ be disjoint sets. Denote $\mathcal{H}_1^r$ and $\mathcal{H}_2^r$ to be classes of all distributions over $\mathbb{R}^d$ with first moments in A and B respectively that have finite rth moment. We have:*

1. *If $r > 1$, then $(\mathcal{H}_1^r \cup \mathcal{H}_2^r, \ell)$ is e.a.s.-predictable iff A and B are F_σ-separable under L_2 distance.*

2. *If $r = 1$, then $(\mathcal{H}_1^r \cup \mathcal{H}_2^r, \ell)$ is e.a.s.-predictable iff A and B are contained in disjoint* open *sets of $\mathbb{R}^d$.*

We now provide an alternative proof of the theorem above by applying our general characterizations in Chapter 4. We need the following lemma which is an easy consequence of Bahr-Esseen inequality (von Bahr and Esseen 1965).

Lemma 14. *Let $X_1, \cdots X_n$ be* i.i.d.*random variables with $\mathbb{E}[|X_1|^r] = M < \infty$ where $r > 1$ and $\mathbb{E}[X_1] = u$, $\bar{X} = \frac{X_1 + \cdots + X_n}{n}$, then for any $\epsilon > 0$ we have*

$$\Pr\left[|\bar{X} - u| \geq \epsilon\right] \leq \frac{C_{\epsilon,M}}{n^{r-1}},$$

where $C_{\epsilon,M}$ is a constant that depends only on ϵ and M.

Proof. W.l.o.g., we assume $u = 0$. By Bahr-Essen inequality (von Bahr and Esseen 1965), we have

$$\mathbb{E}[|\bar{X}|^r] \leq 2\frac{\sum_{i=1}^n \mathbb{E}[|X_i|^r]}{n^r} = \frac{2Mn}{n^r} = \frac{2M}{n^{r-1}}.$$

The lemma follows by a simple application of Markov inequality. □

Proof of Theorem 15(1). We first prove the sufficiency. Since A, B are F_σ-separable, we have by Lemma 11 that there exist nestings $\{A_i, i \geq 1\}$, $\{B_i, i \geq\}$ of A, B respectively, such that

$$\forall i \geq 1, \inf\{||\mathbf{x} - \mathbf{y}||_2 : \mathbf{x} \in A_i, \mathbf{y} \in B_i\} \geq \frac{1}{i}.$$

We now define $\mathcal{G}_i = \{p \in \mathcal{H}_1^r \cup \mathcal{H}_2^r : \mathbb{E}[|X_p|^r] \leq i \text{ and } \mathbb{E}[X_p] \in A_i \cup B_i\}$, where X_p is the random variable governed by distribution p. Clearly, we have $\mathcal{G}_i \subset \mathcal{G}_{i+1}$ for all $i \geq 1$. We now show that $\mathcal{H}_1^r \cup \mathcal{H}_2^r = \bigcup_{i \geq 1} \mathcal{G}_i$, thus proving that $\{\mathcal{G}_i, i \geq 1\}$ forms a nesting of $\mathcal{H}_1^r \cup \mathcal{H}_2^r$. To see this, let $p \in \mathcal{H}_1^r \cup \mathcal{H}_2^r$ be an arbitrary distribution with $\mathbb{E}[X_p] \in A$ (respectively $\in B$), we have $\mathbb{E}[X_p] \in A_i$ (respectively $\in B_i$) for some i. Since X_p also has finite rth moment, we have $\mathbb{E}[|X_p|^r] \leq j$ for some j. Therefore, we have $p \in \mathcal{G}_{\max\{i,j\}}$ by nesting of A_i (respectively B_i).

By Theorem 1, we only need to show that $(\mathcal{G}_i, \ell)$ is η-predictable for all $i \geq 1$ and $\eta > 0$. By Lemma 14, we know that by letting sample size n large enough one can make

$$p(|\bar{X}_p - \mathbb{E}[X_p]| \geq \frac{1}{2i}) \leq \eta$$

for all distributions $p \in \mathcal{G}_i$. To achieve the η-predictability, we will predict at step n that the first moment is in A_i (respectively B_i) if $\bar{X}_p$ is closer to A_i (respectively B_i), and *retain* the same prediction for all time step $\geq n$.

We now prove the necessity. Let $\mathcal{G}$ be the class of all Gaussian distributions with mean vectors in $A \cup B$ and covariance matrix I (i.e., the identity matrix). We have $(\mathcal{G}, \ell)$ is *e.a.s.*-predictable, since $\mathcal{G}$ is a subset of $\mathcal{H}_1^r \cup \mathcal{H}_2^r$. Now, by Theorem 6, there exists a nesting $\{\mathcal{G}_i, i \geq 1\}$ of $\mathcal{G}$ such that for all $i \geq 1$, $(\mathcal{G}_i, \ell)$ is $\frac{1}{4}$-predictable. Define $A_i = \{\mathbf{x} \in A : \exists p \in \mathcal{G}_i \ s.t. \ \mathbb{E}[X_p] = \mathbf{x}\}$ (and B_i similarly). Clearly, $\{A_i, i \geq 1\}$ forms a nesting of A, and $\{B_i, i \geq 1\}$ forms a nesting of B. We show that there is no limit point of A_i (respectively B_i) in B (respectively A). This will complete the proof by Lemma 11.

Suppose we have $\mathbf{x}_1, \mathbf{x}_2, \cdots \in A_i$ and $\mathbf{x} \in B_i$ such that $||\mathbf{x}_n - \mathbf{x}||_2 \to 0$ as $n \to \infty$. Denote p_n to be the Gaussian distribution with mean $\mathbf{x}_n$ and p be the Gaussian distribution with mean $\mathbf{x}$. We have $||p_n - p||_{TV} \to 0$ as $n \to \infty$. By Lemma 13, this will violate the $\frac{1}{4}$-predictability of $\mathcal{G}_i$. □

We now prove Theorem 16 below, which is similar to Theorem 15(2). Theorem 16 shows that, it is not possible to determine whether the first moment of a distribution over $\mathbb{N}$ is

$\geq c$ or $< c$ for some constant $c > 0$ eventually almost surely, if no assumption is imposed on the distributions. Perhaps surprisingly, in Theorem 17, we can show that to determine whether the first moment is $\leq c$ or $> c$ for any constant $c > 0$ is actually possible with finitely many errors almost surely without any assumption on the distributions!

Theorem 16. *Let $\mathcal{H}_1$ be the class of all distributions over $\mathbb{N}$ with first moment $\geq c$ and $\mathcal{H}_2$ be the class of all distributions over $\mathbb{N}$ with first moment $< c$, where $c > 0$ is some real constant. Then $(\mathcal{H}_1 \cup \mathcal{H}_2, \ell)$ is* not *e.a.s.-predictable.*

Proof. Suppose not, we have a nesting $\{\mathcal{G}_i, i \geq 1\}$ of $\mathcal{H}_1 \cup \mathcal{H}_2$ such that for all $i \geq 1$, $(\mathcal{G}_i, \ell)$ is $\frac{1}{4}$-predictable with a sample size i, by Theorem 1 and Lemma 8. Denote $\mathcal{G}_i = A_i \cup B_i$ such that $A_i \subset \mathcal{H}_1$ and $B_i \subset \mathcal{H}_2$. Since $(\mathcal{G}_i, \ell)$ is $\frac{1}{4}$-predictable, we have for any $p \in A_i$ and $q \in B_i$, the total variation $||p - q||_{TV} \geq \frac{1}{4i}$. (Otherwise, $(\mathcal{G}_i, \ell)$ will not be $\frac{1}{4}$-predictable by Lemma 13)

We now construct a sequence of distributions $p_1, p_2, \cdots$ such that $p_k \in A_{n_k}$ for some $n_k \to \infty$ as $k \to \infty$, and distribution $p \in \mathcal{H}_2$ such that $||p - p_k||_{TV} \leq \frac{1}{8n_k}$ for all $k \geq 1$. By our discussion above, we know that $p \notin B_{n_k}$ for all $k \geq 1$. Since $\{B_i\}$ forms a nesting and $n_k \to \infty$, we have $p \notin B_i$ for all $i \geq 1$. However, this will imply that $\mathcal{H}_2 \neq \bigcup_{i \geq 1} B_i$ since $p \in \mathcal{H}_2$. This contradicts our premise that $\{B_i, i \geq 1\}$ is a nesting of $\mathcal{H}_2$, thus completing the proof.

The construction of sequence of distributions $p_1, p_2, \cdots$ and p goes as follows. Let $x_1, x_2, \cdots \in \mathbb{R}^+$ be a monotone increasing sequence such that $x_n \to c$ as $n \to \infty$ and

$$\sum_{n=1}^{\infty} |x_{n+1} - x_n| < \infty.$$

Let p_1 be an arbitrary distribution over $\mathbb{N}$ such that $\mathbb{E}[X_{p_1}] = x_1$ and the support of p_1 is upper bounded by some number $N_1 \leq x_1$. Since $p_1 \in \mathcal{H}_1$, we have $p_1 \in A_{n_1}$ for some n_1. We now recursively construct p_k for $k \geq 2$. Suppose p_k has been constructed with the property that $\mathbb{E}[X_{p_k}] = x_k$, $p_k \in A_{n_k}$ for some n_k and p_k has support upper bounded by N_k. We now construct distribution p_{k+1} as follows. Let ϵ_k be a real number and N_{k+1} be

an integer to be determined later. We define

$$p_{k+1} = (1-\epsilon_k)p_k + \epsilon_k U(\{N_k+1, \cdots, N_{k+1}\}),$$

where $U(S)$ is the uniform distribution over set S. We now select $\epsilon_k \leq \frac{1}{16*2^k n_k}$ and N_{k+1} large enough such that $\mathbb{E}[X_{p_{k+1}}] = x_{k+1}$. Such a selection exists since $x_{k+1} > x_k$ and the first moment is continuous w.r.t. the probability mass. Denote n_{k+1} to be some number such that $p_{k+1} \in A_{n_{k+1}}$ and $n_{k+1} > n_k$ by nesting properties of $\{A_i, i \geq 1\}$. We now define

$$p = p_1 \prod_{k=1}^{\infty}(1-\epsilon_k) + \sum_{k=1}^{\infty}\epsilon_k \prod_{j=k+1}^{\infty}(1-\epsilon_j)U(\{N_j+1, \cdots, N_{j+1}\}).$$

We now show that p is a valid probability distribution, i.e., we have

$$\prod_{k=1}^{\infty}(1-\epsilon_k) + \sum_{k=1}^{\infty}\epsilon_k \prod_{j=k+1}^{\infty}(1-\epsilon_j) = \lim_{k\to\infty}\prod_{j=k}^{\infty}(1-\epsilon_j) = 1,$$

the last equality follows since $\sum_{j=k}^{\infty}\epsilon_j \to 0$ as $k \to \infty$ by $\epsilon_k \leq \frac{1}{16*2^k n_k}$. We now verify that $||p - p_k||_{TV} \leq \frac{1}{8n_k}$, this follows from

$$\prod_{j=k+1}^{\infty}(1-\epsilon_j) \geq 1 - \sum_{j=k+1}^{\infty}\epsilon_j \geq 1 - \frac{1}{16n_k},$$

and $\epsilon_k \leq \frac{1}{16*2^k n_k}$. Lastly, we have to show that $\mathbb{E}[X_p] = c$. This follows from the dominated convergence theorem by notice that all the (probability mass of) distributions p_k is upper bounded by distribution (not necessarily a probability distribution)

$$p' = p_1 + \sum_{k=1}^{\infty}\epsilon_k U(\{N_k+1, \cdots, N_{k+1}\}),$$

and

$$\sum_{n=1}^{\infty} np'(X_{p'} = n) \leq \sum_{n=1}^{\infty}|x_{n+1} - x_n| < \infty.$$

□

Remark 10. *Note that the necessary part of Theorem 15(2) follows from similar arguments as in Theorem 16. The sufficiency part of Theorem 15 is a simple application of strong law of large numbers.*

Theorem 17. *Let $\mathcal{H}_1$ be the class of all distributions over $\mathbb{N}$ with first moment $\leq c$ and $\mathcal{H}_2$ be the class of all distributions over $\mathbb{N}$ with first moment $> c$, where $c > 0$ is some real constant. Then $(\mathcal{H}_1 \cup \mathcal{H}_2, \ell)$ is e.a.s.-predictable.*

Proof. The idea of proving this theorem is to construct an estimator $\hat{\Theta}_n$ such that for any distribution p over $\mathbb{N}$, we have

$$\hat{\Theta}_n \leq \mathbb{E}[X_p] \text{ and } \hat{\Theta}_n \to \mathbb{E}[X_p], \text{ almost surely.}$$

Suppose such an estimator exist, we define the predictor to be $Y_n = 1$ if $\hat{\Theta}_n \leq c$ and $Y_n = 2$ otherwise. To see why such a predictor works, suppose $\mathbb{E}[X_p] < c$ (respectively $> c$), we have $\hat{\Theta}_n$ will eventually $< c$ (respectively $> c$), since $\hat{\Theta}_n \to \mathbb{E}[X_p]$ almost surely. If $\mathbb{E}[X_p] = c$, we know that the predictor will also make the right prediction eventually almost surely, since $\hat{\Theta}_n \leq \mathbb{E}[X_p] = c$ eventually almost surely.

We now construct such an estimator $\hat{\Theta}_n$. We partition the estimation into phases, each phase has one estimation. At phase k, we will request $6k^4 \log k$ samples and estimate the probability mass $p(X_p = i)$ for all $i \leq k$ to be

$$\hat{p}_i = \bar{p}_i - \frac{1}{k^2},$$

where $\bar{p}_i$ is the frequency of i that appears in the sample. Now, by Chernoff bound and union bound, we have

$$p(\forall i \leq n, p_i - \frac{2}{k^2} \leq \hat{p}_i \leq p_i) \geq 1 - \frac{1}{k^2},$$

where $p_i = p(X_p = i)$. The estimates $\hat{\Theta}_n$ at phase k is given by

$$\sum_{i=1}^{n} n\hat{p}_i.$$

By Borel-Cantelli lemma, we know that with probability 1, there exists a number K possibly depends on the realization, such that $\mathbb{E}[X_p] - \frac{1}{k} \leq \hat{\Theta}_n \leq \mathbb{E}[X_p]$ for all phase $k \geq K$. This completes the proof. □

5.2.2 Testing properties of entropy

In this subsection, we will restrict the distributions to be supported on $\mathbb{N}$. For any distribution p over $\mathbb{N}$, the entropy of p is defined to be

$$H(p) = \sum_{i=1}^{\infty} p_i \log \frac{1}{p_i},$$

where $p_i = p(X_p = i)$ (we will use such an abbreviation in the sequel). We define the tail entropy of p of order k to be

$$H_k(p) = \sum_{i=k}^{\infty} p_i \log \frac{1}{p_i}.$$

Let $A, B \subset \mathbb{R}^+$ be two disjoint sets, we would like to decide whether the entropy of some underlying distribution is in A or B by observing *i.i.d.* samples from the distribution. For any class $\mathcal{H}$ of distributions over $\mathbb{N}$, we will denote $\mathcal{H}_A = \{p \in \mathcal{H} : H(p) \in A\}$ and $\mathcal{H}_B = \{p \in \mathcal{H} : H(p) \in B\}$.

For any class $\mathcal{H}$, we say the tail entropy of $\mathcal{H}$ is *dominated* by some function $\rho : \mathbb{N} \to \mathbb{R}^+$ if for all $p \in \mathcal{H}$ there exists a number N_p, such that for all $k \geq N_p$ we have

$$H_k(p) \leq \rho(k).$$

The main result of this subsection characterizes what entropy properties can be tested with finitely many errors.

Theorem 18. *Let $\rho : \mathbb{N} \to \mathbb{R}^+$ be any monotone decreasing function such that $\rho(k) \to 0$ as $k \to \infty$. $\mathcal{H}$ is the class of* all *distributions with tail entropy dominated by ρ. Then for any disjoint set $A, B \subset \mathbb{R}^+$, $(\mathcal{H}_A \cup \mathcal{H}_B, \ell)$ is e.a.s.-predictable iff A, B are F_σ-separable under L_2 distance.*

Proof. We first prove the sufficiency. Let

$$\mathcal{H}'_i = \{p \in \mathcal{H} : \forall k \geq i,\ H_k(p) \leq \rho(k)\}.$$

By dominance of tail entropy, we have that $\{\mathcal{H}'_i, i \geq 1\}$ forms a nesting of $\mathcal{H}$. Now, since A, B are F_σ-separable, we have nestings $\{A_i, i \geq 1\}$ and $\{B_i, i \geq 1\}$ for A, B respectively, such that

$$\inf\{|x - y| : x \in A_i,\ y \in B_i\} \geq \frac{1}{i}.$$

Define

$$\mathcal{H}_i = \{p \in \mathcal{H}'_{N(i)} : H(p) \in A_i \cup B_i\},$$

where $N(i)$ is choosing so that $\forall i \geq 1,\ N(i) < N(i+1)$ and $\rho(N(i)) \leq \frac{1}{8i}$. Clearly, such a sequence $N(i)$ exists since $\rho(k) \to 0$ as $k \to \infty$. We now show that $\{\mathcal{H}_i, i \geq 1\}$ forms a nesting of $\mathcal{H}_A \cup \mathcal{H}_B$. The nested properly follows directly from nesting $\{\mathcal{H}'_i\}$, $\{A_i\}$ and $\{B_i\}$. For any $p \in \mathcal{H}_A \cup \mathcal{H}_B$, we assume, w.l.o.g, $H(p) \in A$. Suppose $p \in \mathcal{H}_{k_1}$ and $H(p) \in A_{k_2}$. Let $N(k) \geq k_1$ for some k, we have $p \in \mathcal{H}_{\max\{N(k),k_2\}}$. Therefore, $\bigcup_{i \geq 1} \mathcal{H}_i = \mathcal{H}_A \cup \mathcal{H}_B$. By Theorem 1, it is sufficient to show that $(\mathcal{H}_i, \ell)$ is η-predictable for all $i \geq 1$ and $\eta > 0$. By construction, we know that for all $p \in \mathcal{H}_i$ we have

$$\forall k \geq i,\ H_k(p) \leq \rho(k).$$

Let $M \geq i$ be a number, large enough, so that $\rho(M) \leq \frac{1}{8i}$. Now, since the entropy function is continuous over bounded support, we can estimate the entropy of $p \in \mathcal{H}_i$ sufficiently accurate so that the estimation differs only $\frac{1}{8i}$ for the entropy restricted on support $[M]$.

Since the tail only contribute $\frac{1}{8i}$ to the whole entropy and $d(A_i, B_i) \geq \frac{1}{i}$, one will be able to decide whether $H(p) \in A_i$ or $H(p) \in B_i$ with arbitrary confidence. The sufficiency follows.

To prove the necessity, we construct a class $\mathcal{H}'$ of distribution over $\mathbb{N}$ with the following properties:

1. The range $\{H(p) : p \in \mathcal{H}'\} = \mathbb{R}^+$.

2. For any $h \in \mathbb{R}^+$, there exists one and only one $p_h \in \mathcal{H}'$ such that $H(p_h) = h$, i.e. the distributions are parameterized by its entropy.

3. For any $h_1, h_2, \cdots \in \mathbb{R}^+$ such that $\lim_{k\to\infty} h_k = h$ for some $h \in \mathbb{R}^+$, we have $||p_h - p_{h_k}||_{TV} \to 0$ as $k \to \infty$, i.e. the class $\mathcal{H}'$ is continuous w.r.t. the entropy under total variation distance.

4. The support of all distributions in $\mathcal{H}'$ are finite.

We now show that if such a class exists, then the necessary condition holds. By property (4) of the class, we know that the tail entropy of $\mathcal{H}'$ is dominated by any ρ, i.e. $\mathcal{H}' \subset \mathcal{H}$. By Theorem 1, we have a nesting $\{\mathcal{G}_i, i \geq 1\}$ for $\mathcal{H}'$ such that $\forall i \geq 1$, $(\mathcal{G}_i, \ell)$ is $\frac{1}{4}$-predictable. Let $A_i = \{x \in A : \exists p \in \mathcal{G}_i \ s.t. \ H(p) = x\}$ and $B_i = \{x \in B : \exists p \in \mathcal{G}_i \ s.t. \ H(p) = x\}$. By property (1) and (2), we know that $\{A_i, i \geq 1\}$ and $\{B_i, i \geq 1\}$ form nesting for A and B respectively. It is sufficient to show that for all $i \geq 1$ there is no sequences $h_1, h_2 \cdots \in A_i$ such that $\lim_{k\to\infty} h_k = h \in B_i$. By property (3), this implies $||p_h - p_{h_k}||_{TV} \to 0$ as $k \to \infty$. A similar argument as in Theorem 15(1) completes the proof.

We now construct the class $\mathcal{H}'$. Let $\mathcal{H}'_0$ be the class that contains one distribution $p^{(0)}$ which assigns probability 1 on 0. We now construct $\mathcal{H}'_k$ for $k \geq 1$ as follows. Denote p_*^{k-1} to be the uniform distribution over $\{0, 1, \cdots, k-1\}$, and δ_k to be the distribution that assigns probability 1 on k. Define

$$\mathcal{H}'_k = \{(1-\epsilon)p_*^{k-1} + \epsilon\delta_k : 0 < \epsilon \leq \frac{1}{k+1}\}.$$

For any distribution $p_\epsilon^k \in \mathcal{H}'_k$ with parameter ϵ, we have

$$H(p_\epsilon^k) = h(\epsilon) + (1-\epsilon)\log k,$$

where $h(\epsilon)$ is the binary entropy function. Let $g_k(\epsilon) = h(\epsilon) + (1-\epsilon)\log k$, we show that $g_k(\epsilon)$ is monotone increasing on $(0, \frac{1}{k+1}]$. Taking derivative on ϵ, we have

$$g'_k(\epsilon) = \log\left(\frac{1}{\epsilon} - 1\right) - \log k,$$

which is non-negative for all $\epsilon \in (0, \frac{1}{k+1}]$. It is easy to verify that the range of entropy of distributions in $\mathcal{H}'_k$ is exactly $(\log k, \log(k+1)]$. We now define

$$\mathcal{H}' = \bigcup_{k \geq 0} \mathcal{H}'_k.$$

It is easy to verify that all the properties of $\mathcal{H}'$ are satisfied. □

We now prove the following theorem, which shows that a finite first moment is sufficient to have a uniform dominance on the tail entropy.

Theorem 19. *If the distributions in $\mathcal{H}$ have finite first moment, then the tail entropy of $\mathcal{H}$ is dominated by $\frac{\log^2 k}{k}$.*

Proof. Let $p \in \mathcal{H}$ and let $M = \mathbb{E}[X_p]$. We show that

$$H_k(p) = O\left(\frac{\log k}{k}\right),$$

where O hides some constant that only depends on M. This will then imply that there exists some number N_p, such that for all $k \geq N_p$, $H_k(p) \leq (\log^2 k)/k$ as desired.

By the Markov inequality, we have

$$\epsilon_k \stackrel{\text{def}}{=} p(X_p \geq k) \leq \frac{M}{k}.$$

Let $\tilde{p}_n = \frac{p_n}{\epsilon_k}$ for all $n \geq k$, i.e. $\tilde{p}$ is a distribution over $\mathbb{N}$ derived from p and whose support is simply the set of integers $\geq k$. We therefore have

$$H_k(p) = \epsilon_k H(\tilde{p}) + \epsilon_k \log \frac{1}{\epsilon_k}.$$

Let $M_k = \mathbb{E}[X_{\tilde{p}}]$ be the first moment of $\tilde{p}$. Clearly, we have

$$k \leq M_k \leq \frac{M}{\epsilon_k}.$$

Since the geometric distribution maximizes entropy among all distributions with first moment M_k (Cover and Thomas 1991), we have

$$H(\tilde{p}) \leq M_k h(1/M_k),$$

where $h(x)$ is the binary entropy function. Thus

$$\epsilon_k H(\tilde{p}) \leq M h(1/M_k) \leq M h(1/k) = O\left(\frac{\log k}{k}\right),$$

where the second inequality we have used the fact that $h(x)$ is monotonically increasing for $x \leq \frac{1}{2}$. A simple estimation also yields $\epsilon_k \log \frac{1}{\epsilon_k} = O\left(\frac{\log k}{k}\right)$. The theorem follows. □

5.3 Applications to matrix properties

We now apply our framework of testing functional properties to matrix properties. To begin with, we first prove the following theorem, which is a direct corollary of Theorem 15(1).

Theorem 20. *Let $\mathcal{H}$ 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 $(\mathcal{H}, \ell_A)$ is e.a.s.-predictable if A is closed in $[0,1]^d$.*

Proof. This follows directly from Theorem 15(1), since $A, \bar{A}$ are F_σ-separable if A is closed, where $\bar{A}$ is the complement of A in $[0,1]^d$. □

For any function $f : [0,1]^d \to \{0,1\}$, we will be able to 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 $(\mathcal{H}, \ell_{A_i})$ 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 $(\mathcal{H}, \ell_{A_f})$ 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 $\mathcal{H}$ 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 $(\mathcal{H}, \ell_A)$ is *e.a.s.*-predictable where $A \subset [0,1]^{d \times d}$ is the subset corresponding to the property.

Theorem 21. *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 are closed. The theorem follows by Theorem 20. □

Theorem 22. *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 needs to check the maximum non-singular square submatrix is of dimension k. Thus the property can be expressed as a function of finite singularity tests. By Theorem 21 we know that the property is still *e.a.s.*-predictable. □

The above theorem is surprising because all of $\mathbf{X}_1, \cdots, \mathbf{X}_n$ will have rank $\geq n - 2\log n$ with high probability even when the entries are Bernoulli($\frac{1}{2}$) (i.e. $\mathbb{E}[X]$ has rank 1). To see

this, for each $\mathbf{X}_i$, note that the probability of being full rank for the first $n - 2\log n$ rows is

$$\prod_{i=2\log n}^{n}(1-2^{-i}) \geq 1 - \sum_{i=\log 2n}^{n} 2^{-i} = 1 - O\left(\frac{1}{n^2}\right),$$

and the probability that none of $\mathbf{X}_i$, $1 \leq i \leq n$ have rank $\leq n - 2\log n$ is $\geq 1 - O(1/n)$. Yet, Theorem 22 shows that it is still possible to infer the rank of $\mathbb{E}[\mathbf{X}]$ from the realizations $\mathbf{X}_1, \mathbf{X}_2, \ldots$.

To unravel the proof in Theorem 22, we also give a more algorithmic way of estimating the rank that is correct eventually almost surely. Let n be the sample size we observed, and $\hat{p}_{i,j}$ be the empirical mean of the (i,j)th entry. The estimation of rank is given by the following optimization problem.

$$\begin{aligned} &\min_{X} \ \operatorname{rank}(X) \\ &s.t.\ \forall i,j,\ |X_{i,j} - \hat{p}_{i,j}| \leq \frac{\log n}{\sqrt{n}}. \end{aligned}$$

We show that the estimation given by the rule above converges to $\operatorname{rank}(\mathbb{E}[\mathbf{X}])$ w.p. 1. Note that, by the selection of the threshold of the constraints at $\log n/\sqrt{n}$ and using the Chernoff bound, the probability that the matrix $\mathbb{E}[\mathbf{X}]$ is not in the feasible set of the optimization above is at most $O\left(\frac{1}{n^2}\right)$. By the Borel-Cantelli lemma, $\mathbb{E}[\mathbf{X}]$ will be in the feasible set eventually almost surely. Therefore, the rule will never over estimate the rank in the limit.

Denote $k = \operatorname{rank}(\mathbb{E}[\mathbf{X}])$. There exists a full rank $k \times k$ submatrix $\mathbf{Y}$ of $\mathbb{E}[\mathbf{X}]$. By the continuity of the determinant and since $\det(\mathbf{Y}) \neq 0$, there is a neighborhood $\mathcal{N}$ of $\mathbf{Y}$ such that any $\mathbf{T} \in \mathcal{N}$ has rank k. Since $\log n/\sqrt{n} \to 0$, there is some constant N depending on the neighborhood $\mathcal{N}$ such that for all $n > N$, the feasible set of the optimization restricted to coordinates of $\mathbf{Y}$ will be a subset of $\mathcal{N}$. Combining this with the observation that $\mathbb{E}[\mathbf{X}]$ can be out of the feasible set only finitely many times, we conclude that the rule will eventually output k almost surely.

Theorem 23. *To determine if* $\mathbb{E}[\boldsymbol{X}]$ *has eigenvalues 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 20. □

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

Problem 2 (Open Problem)**.** *Is determining whether a matrix is diagonalizable e.a.s.-predictable?*

5.4 Summary

The author is the primary contributor for the work in this chapter, which has been partially published in (Wu and Santhanam 2020) and (Wu and Santhanam 2021b).

Chapter 6

$e.a.s.$-Prediction in online learning

6.1 Introduction

In this chapter we will study the $e.a.s.$-prediction paradigm in the online learning setup. Let $\mathcal{H}$ be a class of binary functions from $\mathcal{X} \to \{0, 1\}$. The online learning setup is a game between two parties: Nature and a Learner. At the beginning of the game, Nature chooses a hypothesis $h \in \mathcal{H}$. The game proceeds in discrete time steps. At time step n, Nature provides an instance $x_n \in \mathcal{X}$ to the Learner—here, we study cases where the sample is generated from a known distribution, an unknown distribution, or even in an adversarial fashion. The Learner then predicts $\hat{y}_n$ as its best guess of what $h(x_n)$ may be, potentially using the history (samples and revealed labels) observed thus far. Subsequently, Nature reveals the true label $h(x_n)$, and the Learner incurs a binary loss $1\{\hat{y}_n \neq h(x_n)\}$—where a loss of 1 implies that an error has been made. The goal of the learner is a strategy that minimizes the number of errors over an infinite horizon—we are looking to characterize those setups where some prediction scheme can always make a finite number of errors with probability 1.

We note that this chapter studies the $e.a.s.$-prediction paradigm exclusively in the online learning task—we do not consider stopping rules in this chapter. However a bulk of existing literature in the online classification task does not yet make a distinction on whether the Learner is aware she has learned the hypothesis or not. Therefore, for more convenient

comparisons with existing literature on this topic, in this chapter alone, we say a scheme "$e.a.s.$-online learns" $\mathcal{H}$ to mean $e.a.s.$-predicts $\mathcal{H}$ with a binary loss.

Such a setup dates back to Littlestone (1988), and it was shown in his seminal paper (Littlestone 1988) that the number of errors made by the Learner is upper bounded by what is known as the Littlestone dimension $\text{Ldim}(\mathcal{H})$ of $\mathcal{H}$. This dimension is independent of the number of time steps the game goes on for. Moreover, Littlestone (1988) also showed that for any class $\mathcal{H}$ that has Littlestone dimension d, the learner must make at least d errors in the worst case.

However, the Littlestone dimension can be too restrictive a measure to handle rich classes that may not even have finite Littlestone dimension. Such classes are still interesting from an online learning perspective, e.g., learning of the class of all polynomial time computable binary functions. In this chapter, we address such problems by relaxing the requirements of *bounded* number of errors to *finite* number of errors (i.e., the number of errors is finite for any underlying hypothesis, but no uniform bound is required). Doing so allows us to include broader classes of interests to be studied meaningfully in the online learning setup.

Our first main result reveals that a binary-labeled class $\mathcal{H}$ can be learned with finitely many errors almost surely by observing an *i.i.d.* sampling from a known distribution μ over $\mathcal{X}$ if and only if $\mathcal{H}$ is *effectively* countable w.r.t. μ— meaning that there exists a countable class $\mathcal{H}' \subset \mathcal{H}$ such that for any $h \in \mathcal{H}$ there exists $h' \in \mathcal{H}'$ we have h, h' differ a zero measure set under μ. It is possible that an uncountable set $\mathcal{H}$ of hypotheses over $\mathcal{X}$ is still effectively countable w.r.t all measures μ over $\mathcal{X}$, therefore, effective countability is a strictly weaker requirement than countability.

We compare our setup with other classical learning models, specifically with PAC learning. As can be anticipated, VC-dimension does not really capture this problem setup. There are classes that have unbounded VC-dimensions, yet can be learned with finitely many errors. For example, let $\mathcal{H}$ be a class of all functions from $\mathbb{R} \to \{0, 1\}$ such that $\forall h \in \mathcal{H}$ there are only *finitely* $x \in \mathbb{R}$ with $h(x) = 1$. Now $\mathcal{H}$ does not have finite VC-dimension. Yet, it is easy to see that $\mathcal{H}$ is trivially $e.a.s.$-predictable with classification

loss—we simply predict 0 on all x unless x has been revealed to have label 1 previously. With this algorithm, we only make one error on each element of $\mathbb{R}$ with label 1 under the hypothesis in force, but these are known to be finitely many. We also show that there are classes with VC-dimension 1, e.g., linear threshold functions over $[0, 1]$, that cannot be predicted in an online fashion even when allowed finitely many errors.

We then extend our result to the multiclass label scenarios. Here we show that the effective countability condition is still sufficient to achieve the finite error guarantee for arbitrary multi-label classes and arbitrary binary loss functions. The converse holds when the class labels are countable, and the loss considered is the classification loss. In the general loss case even with ternary labels, we show however that the effective countability condition is not necessary to achieve a finite error guarantee.

In Section 6.5, we compare our setup with the classical learning of recursive functions (a.k.a. inductive inference) by considering computable prediction rules. We show that the class of all total computable binary functions can be learned online with finitely many errors almost surely using a computable learner by observing *i.i.d.* samples from certain non-degenerate (i.e., has infinite support) distributions over $\mathbb{N}$ (see the discussion after Theorem 30). We also study the case where the predictor is not just computable, but has bounded computational resources, e.g., polynomial time computable.

In Section 6.6, we study the scenario when independent Bernoulli noise is present. We show that a finite error guarantee is still achievable for effectively countable classes even with independent Bernoulli(η) noise where $\eta < \frac{1}{2}$.

Finally, we relax the finite error guarantee to the guarantee that the average errors per time step must go to zero. We show that the class of all measurable functions from $\mathbb{R} \to \{0, 1\}$ admits a predictor that achieves average errors per time step going to zero with probability 1. We then consider the cases when the error rates are controlled. Such a setup was recently considered by Bousquet et al. (2020a), where they provide a complete characterization on the error rates but with a weaker realizable assumption. We compare their result with our setup in Section 6.7.

6.2 Problem setup

Let $\mathcal{X}$ be the instance space that is endowed with some fixed *separable* σ-algebra $\mathcal{F}$. We say $\mathcal{F}$ is separable, if there exists a countable set $\mathcal{G}$ of measurable sets in $\mathcal{F}$, such that for all probability distributions μ on $\mathcal{F}$ and for any measurable set $A \in \mathcal{F}$ and $\epsilon > 0$, there exists $G_\epsilon \in \mathcal{G}$ with

$$\mu(A \Delta G_\epsilon) \leq \epsilon,$$

where Δ is symmetric difference. Clearly, the Borel σ-algebra over $\mathbb{R}^d$ is separable (e.g., we can take $\mathcal{G}$ to be the *algebra* generated by the set of all cuboids in $\mathbb{R}^d$ with vertices at rational coordinates). We will also assume the single point set in $\mathcal{F}$ is measurable.

Let $\mathcal{Y}$ be a label space that will often be assumed to be finite or countable. A binary loss is a function $\ell : \mathcal{Y} \times \mathcal{Y} \to \{0, 1\}$ that is symmetric, i.e., $\ell(y_1, y_2) = \ell(y_2, y_1)$ and reflexive, i.e., $\ell(y_1, y_1) = 0$ for all $y_1, y_2 \in \mathcal{Y}$. The learning strategy is a function

$$\Phi : (\mathcal{X} \times \mathcal{Y})^* \times \mathcal{X} \to \mathcal{Y},$$

that is measurable over the cylinder σ-algebra on $(\mathcal{X} \times \mathcal{Y})^\infty \times \mathcal{X}$.

We consider the following learning game between Nature and the Learner that proceeds in discrete time steps. Let $\mathcal{H}$ be a class of measurable functions from $\mathcal{X} \to \mathcal{Y}$, and μ be a probability measure over $\mathcal{X}$, known to both parties. At the beginning, Nature chooses a hypothesis $h \in \mathcal{H}$. At each time step n, nature independently samples $X_n \sim \mu$ and provides it to the learner. The learner then outputs an estimate Y_n of the true label $h(X_n)$, potentially using the history $\{(X_1, h(X_1)), \cdots, (X_{n-1}, h(X_{n-1}))\}$ thus far. Nature then provides the true label $h(X_n)$ to the learner, and the learner incurs the binary loss $\ell(Y_n, h(X_n)$. The learner makes an *error* at time step n if $\ell(Y_n, h(X_n)) = 1$.

Denote $Z_i = (X_i, h(X_i))$ to be the instance-label pair at time step i, and $Z_1^i = (Z_1, \cdots, Z_i)$ be the history observed by the learner upto time step i.

Definition 11. *A class $\mathcal{H}$ is said to be e.a.s.-online learnable w.r.t. distribution μ, if there exists a learning strategy Φ_μ such that*

$$\Pr\left(\sum_{n=1}^{\infty} \ell(\Phi_\mu(Z_1^{n-1}, X_n), h(X_n)) < \infty\right) = 1,$$

for all $h \in \mathcal{H}$ with $X_1, X_2, \cdots \overset{i.i.d.}{\sim} \mu$.

Remark 11. *Note that, here the e.a.s.-online learnability is equivalent to the e.a.s.-predictability we introduced in Chapter 4. This is* different *from the stopping rule required e.a.s.-learnability as defined in Definition 6, and the change in notation is due to the convenience in the online learning literature.*

We also introduce the following stronger version of *e.a.s.*-online learnability when the distribution is unknown to the learner.

Definition 12. *A class $\mathcal{H}$ is said to be strongly e.a.s.-online learnable, if there exists a learning strategy Φ such that*

$$\Pr\left(\sum_{n=1}^{\infty} \ell(\Phi(Z_1^{n-1}, X_n), h(X_n)) < \infty\right) = 1,$$

for all $h \in \mathcal{H}$ and μ with $X_1, X_2, \cdots \overset{i.i.d.}{\sim} \mu$.

Note that the learning strategy Φ has to be universal for all μ in the strong *e.a.s.*-online learnability setup. It is an open problem if *e.a.s.*-online learnable for all μ implies strongly *e.a.s.*-online learnable.

In both cases above, we will say that the strategy Φ *e.a.s.*-online learns (or strongly *e.a.s.*-online learns) $\mathcal{H}$ w.r.t. μ. For notational convenience, we drop the reference to μ where doing so leads to no ambiguity.

The following notion will be used frequently in this chapter.

Definition 13. *A class* $\mathcal{H}'$ effectively covers *a class* $\mathcal{H}$ *w.r.t. distribution* μ *and loss* ℓ*, if for all* $h \in \mathcal{H}$ *there exists* $h' \in \mathcal{H}'$ *such that*

$$\mu\{x : \ell(h(x), h'(x)) = 1\} = 0.$$

We say $\mathcal{H}$ *is* effectively countable *(respectively size* n*) w.r.t.* μ *and* ℓ*, if there exists* $\mathcal{H}'$ *that is countable (respectively size* n*), such that* $\mathcal{H}'$ *effectively covers* $\mathcal{H}$ *w.r.t.* μ *and* ℓ*.*

We begin with the following lemma, a version of Structural Risk Minimization for the online learning problems.

Lemma 15. *Let* $\mathcal{H}_1, \mathcal{H}_2, \cdots$ *be countably function classes that share the same instance space and loss function. If* $\mathcal{H}_n$ *is e.a.s.-online learnable for all* $n \in \mathbb{N}$ *w.r.t. distribution* μ*, then*

$$\mathcal{H} = \bigcup_{n \in \mathbb{N}} \mathcal{H}_n$$

is also e.a.s.-online learnable w.r.t. μ*.*

Proof. Let Φ_k be the strategy that *e.a.s.*-online learns class $\mathcal{H}_k$. We define a strategy for $\mathcal{H}$ as follows. At each time step n, denote by $e(n, k)$, the number of errors that Φ_k would make if it were employed in the first $n - 1$ time steps. Let

$$J_n = \operatorname*{argmin}_{k \in \mathbb{N}} \{e(n, k) + k\}. \tag{6.1}$$

We then use Φ_{J_n} to make the prediction at time step n.

We now show that this strategy indeed makes finitely many errors with probability 1 no matter what h is chosen from $\mathcal{H}$. Assume $h \in \mathcal{H}_t$ for some $t \in \mathbb{N}$. Let $B \subset \mathcal{X}^\infty$ be the set such that Φ_t would make finite errors on all realizations in B. We have $\mu(B) = 1$ because Φ_t *e.a.s.*-online learns $\mathcal{H}_t$. Fix any $\mathbf{x} \in B$ and denote s to be the number of errors Φ_t would make on $\mathbf{x}$. We have

$$e(n, t) + t \leq s + t$$

for all n and from (6.1), note that therefore for all n, J_n, the index chosen is $\leq s+t$.

Furthermore, for any Φ_i with $i \in \{1, \cdots, s+t\}$, once Φ_i makes more than $s+t$ errors, it will no longer be chosen in (6.1). Therefore, we will make at most $(s+t)^2$ errors, and thus finitely many errors on $\mathbf{x}$. Therefore, our strategy also makes finite errors with probability 1 when h is selected. The lemma follows. □

Remark 12. *Note that Lemma 15 holds for strong e.a.s.-online learning as well, since the construction of learning strategy does not depend on the knowledge of underlying distribution.*

We prove the following technical lemma that relates deterministic sampling with randomized sampling. Informally, Lemma 16 shows that for any given ordering of the natural numbers and any gap sequence, we can construct a distribution p over $\mathbb{N}$ such that with probability 1, after finitely many steps the first appearance of numbers in the *i.i.d.* sampling of p will follow the prescribed ordering with the given gaps. We will use it in Theorem 30 to show that the class of all computable binary functions can be *e.a.s.*-online learned using a computable learner with *i.i.d.* sampling from certain distributions supported on the entirety of $\mathbb{N}$.

Lemma 16. *Let $a_1, a_2, \cdots$ be an arbitrary ordering of $\mathbb{N}$, and $s_1, s_2, \cdots \in \mathbb{N}^+$ be an arbitrary sequence. Then there exists a distribution p over $\mathbb{N}$, such that*

$$p\left(\exists N : \forall n \geq N,\ T_n - T_{n-1} \geq s_{n-1}\right) = 1$$

where T_n is the first time a_n appears in an i.i.d. sample from p.

Proof. Let $p_n = p(X = a_n)$. Since $T_n - T_{n-1} < s_{n-1}$ implies that a_n appears earlier than the s_{n-1}'th appearance of a_{n-1}, we have

$$p(T_n - T_{n-1} < s_{n-1}) \leq 1 - \left(\frac{p_{n-1}}{p_n + p_{n-1}}\right)^{s_{n-1}}. \tag{6.2}$$

Taking

$$p_n = \frac{C}{(n!)^2 \prod_{i=1}^{n-1} s_i},$$

where C is the normalization constant. We have that (6.2) is further upper bounded by $O(\frac{1}{n^2})$. Since $\frac{1}{n^2}$ is summable, an application of the Borel-Cantelli Lemma completes the proof. □

6.3 Binary labels

In this section, we will consider the case when the label is binary and the loss is $\ell(y_1, y_2) = 1\{y_1 \neq y_2\}$. We will refer to this loss as the *classification loss* in the sequel. Note that since our loss function is binary valued on binary labels, the only losses can either be trivial or classification loss. We say a distribution over $\mathcal{X}$ is *non-degenerate* if it has infinite support.

The following theorem fully characterizes the *e.a.s.*-online learnability in the binary label case.

Theorem 24. *For any probability measure μ, a class $\mathcal{H}$ with binary labels is e.a.s.-online learnable w.r.t. distribution μ iff $\mathcal{H}$ is* effectively *countable w.r.t. μ.*

Proof. To see that effectively countable implies *e.a.s.*-online learnable, note that for any class that contains *effectively* 1 hypothesis is trivially *e.a.s.*-online learnable. Therefore, any class $\mathcal{H}$ with effectively countably many hypotheses is *e.a.s.*-online learnable by appealing to Lemma 15.

For the necessary part of the proof, we will assume w.l.o.g. that for all $h_1 \neq h_2 \in \mathcal{H}$ we have $\Pr_{X\sim\mu}(h_1(X) \neq h_2(X)) > 0$. Note that any hypothesis class can be reduced to one satisfying the above property by choosing a representative function in each equivalence class defined by the relation $h_1 \sim h_2 : \Pr_{x\sim\mu}(h_1(x) \neq h_2(x)) = 0$.

We now prove that if $\mathcal{H}$ admits a function $\Phi : (\mathcal{X} \times \{0,1\})^* \times \{0,1\} \to \{0,1\}$ such that $\forall h \in \mathcal{H}$

$$\Pr\left(\sum_{n=1}^{\infty} 1\{\Phi(Z_1^{n-1}, X_n) \neq h(X_n)\} < \infty\right) = 1, \tag{6.3}$$

where $Z_1^{n-1} = (Z_1, \cdots, Z_{n-1})$ and $Z_i = (X_i, h(X_i))$ is the instance-label pair at step i generated by μ, then $\mathcal{H}$ is countable.

Define the event

$$A_n^h = \left\{ X_1^\infty : \sum_{k=n}^{\infty} 1\{\Phi(Z_1^k, X_{k+1}) \neq h(X_{k+1})\} > 0 \right\},$$

and let

$$\mathcal{H}_n = \left\{ h \in \mathcal{H} : \Pr\left(A_n^h\right) \leq \frac{1}{4} \right\}. \tag{6.4}$$

From equation (6.3), we must have $\mathcal{H} = \bigcup_{n=1}^{\infty} \mathcal{H}_n$. Since for any $h \in \mathcal{H}$, we have $\Pr(A_n^h) \to 1$ as $n \to \infty$, there exists some k such that $\Pr(A_k^h) \leq \frac{1}{4}$, i.e., $h \in \mathcal{H}_k$. We show that $\mathcal{H}_n$ is countable for all $n \in \mathbb{N}$, which will prove the Theorem.

Central to our proof is our claim that for all $n \in \mathbb{N}$ we have

$$\inf\{\Pr_{X\sim\mu}(h_1(X) \neq h_2(X)) : h_1 \neq h_2 \in \mathcal{H}_n\} > 0. \tag{6.5}$$

Now equation (6.5) implies that $\mathcal{H}_n$ is countable. To see this, let

$$d(h_1, h_2) = \Pr_{X\sim\mu}(h_1(X) \neq h_2(X)).$$

The σ-algebra over $\mathcal{X}$ is separable, so by definition, there is a countable dense subset of the σ-algebra with respect to the distance d. Since each measurable set in the σ-algebra can be represented by a measurable function from $\mathcal{X} \to \{0, 1\}$, we conclude that there is a countable collection $\mathcal{G} = \{g_1, g_2 \cdots\}$ of measurable functions from $\mathcal{X} \to \{0, 1\}$, such that $\mathcal{G}$ is dense in the set of all measurable functions from $\mathcal{X} \to \{0, 1\}$ under the distance d defined above.

Suppose the infimum in (6.5) is $\epsilon > 0$. Since $\mathcal{G}$ is dense in the set of all binary valued measurable functions on $\mathcal{X}$, for each $h \in \mathcal{H}_n$, there is a function $g \in \mathcal{G}$ such that $d(h, g) < \epsilon/2$. By the triangle inequality, for all $h' \neq h$, we will also have $d(h', g) > \epsilon/2$.

Therefore every $h \in \mathcal{H}_n$ can be associated with a distinct $g \in \mathcal{G}$, which then implies that $\mathcal{H}_n$ is at most countable.

We now establish equation (6.5). The intuition behind this claim is that if the infimum in (6.5) were 0, then we could choose two hypotheses in $\mathcal{H}_n$ that were arbitrarily close in the distance d—and hence close enough that they are indistinguishable with sample size n on a large enough subset $A \subset \mathcal{X}^\infty$. However, these hypotheses will eventually differ in the suffix of the strings in A, therefore no predictor could agree with both of them in the suffix past step n. But if the probability of A is large enough, it contradicts the definition of $\mathcal{H}_n$ in (6.4).

Formally, suppose (6.5) did not hold. Then there exist $h_1, h_2 \in \mathcal{H}_n$ such that $0 < d(h_1, h_2) < \delta_n$, where δ_n is chosen so that $(1 - \delta_n)^n > \frac{1}{2}$.

Let $A \subset \mathcal{X}^\infty$ be the event that h_1, h_2 cannot be distinguished with n samples. By construction, we have $\Pr(A) > \frac{1}{2}$. For $j \in \{1, 2\}$, let

$$p_j = \Pr(\Phi \text{ makes error after step } n \text{ on } h_j).$$

We show that $\max\{p_1, p_2\} > \frac{1}{4}$, which will then contradict equation (6.4) since $h_1, h_2 \in \mathcal{H}_n$, thus establishing equation (6.5). To see that $\max\{p_1, p_2\} > \frac{1}{4}$, we use a probabilistic argument. Let $\mathbf{h}$ be the random variable uniformly chosen from $\{h_1, h_2\}$. We only need to show

$$\mathbb{E}_{\mathbf{h}} \mathbb{E}_{X \sim \mu^\infty}[1\{\Phi \text{ makes error after step } n \text{ on } \mathbf{h}\} \mid X \in A] \geq \frac{1}{2} \tag{6.6}$$

Let $B \subset \mathcal{X}^\infty$ be the event that there exists a instance Y after step n that first reveals $h_1(Y) \neq h_2(Y)$. We have $\Pr(B) = 1$, since $d(h_1, h_2) > 0$. Note that

$$\mathbb{E}_{X \sim \mu^\infty}[\mathbb{E}_{\mathbf{h}}[1\{\Phi \text{ makes error after step } n \text{ on } \mathbf{h}\}] \mid X \in A \cap B] \geq \frac{1}{2}, \tag{6.7}$$

since condition on any $X \in A \cap B$ event $C = \{\Phi \text{ makes error at sample } Y \text{ on } \mathbf{h}\}$ implies the event in the equation above, and because $\mathbb{E}_{\mathbf{h}}[1\{C\}] = \frac{1}{2}$. We now have

$$\mathbb{E}_{X\sim\mu^\infty}[\mathbb{E}_{\mathbf{h}}[1\{\Phi \text{ makes error after step } i \text{ on } \mathbf{h}\}] \mid X \in A] \geq \frac{1}{2}, \tag{6.8}$$

since $\Pr(B) = 1$. Finally (6.8) implies (6.6) by exchanging order of expectation, which is justified by Fubini's theorem. □

Example 10. *Let $\mathcal{X} = [0,1]$ and μ be the uniform distribution over $[0,1]$. We consider the linear threshold functions*

$$h_a(x) = 0 \text{ if } a \geq x \text{ and } h_a(x) = 1 \text{ otherwise.}$$

Denote $\mathcal{H} = \{h_a : a \in [0,1]\}$. By Theorem 24 we know that $\mathcal{H}$ is not *e.a.s.-online learnable.*

A couple of observations need to be emphasized here. Note that the VC dimension of $\mathcal{H}$ is 1. Therefore the VC-theorem (Shalev-Shwartz and Ben-David 2014, Thm 6.7) posits that there is a learning rule $\hat{h}_n$, such that for any $h_a \in \mathcal{H}$ and a sample S_n of size n, with high probability say $1 - \frac{1}{n^2}$ over S_n, we have

$$\Pr_{x\sim\mu}[\hat{h}_n(S_n, x) \neq h_a(x)] \leq O\left(\frac{\log n}{n}\right).$$

Theorem 24 implies that we cannot improve the upper bound from $O\Big(\frac{\log n}{n}\Big)$ to, say, $O\Big(\frac{1}{n \log^{1+\epsilon} n}\Big)$ with $\epsilon > 0$. If we could, note that since $\sum_{n=1}^{\infty} \frac{1}{n \log^{1+\epsilon} n} < \infty$, an application of the Borel-Cantelli lemma would imply that $\hat{h}_n$ only makes a finite number of errors no matter the hypothesis in force, thus violating Theorem 24. Note that this lower bound holds even when we allow the hidden constant of the big-O notation to be dependent on the underlying hypothesis. This differs only by a polylog term compared to the optimal $\Omega(\frac{1}{n})$ bound by Schuurmans (1997).

We observe the following simple corollary.

Corollary 3. *If a binary measurable hypothesis class $\mathcal{H} : \mathcal{X} \to \{0, 1\}$ has finite Littlestone dimension, then $\mathcal{H}$ is effectively countable w.r.t. any distribution μ over $\mathcal{X}$. In particular, if the instance space $\mathcal{X}$ is countable, then $\mathcal{H}$ is countable.*

Proof. Finite Littlestone dimension implies $\mathcal{H}$ is online learnable with a bounded number of errors and an adversarial sampling process (Shalev-Shwartz and Ben-David 2014, Chapter 21). The corollary follows by Theorem 24 and noting that the learning strategy is independent of the distribution. The second part follows from the fact that for any hypothesis class $\mathcal{H}$ over a countable domain $\mathcal{X}$, if $\mathcal{H}$ is effectively countable w.r.t. a distribution with support of the whole of $\mathcal{X}$, then $\mathcal{H}$ is countable. □

Theorem 24 also yields that if a class $\mathcal{H}$ is strongly *e.a.s.*-online learnable, then $\mathcal{H}$ is effectively countable w.r.t. any distribution. By Corollary 3 above, this implies that the restriction of $\mathcal{H}$ on any countable subset of $\mathcal{X}$ must be countable.

However, this does not mean that the every strongly *e.a.s.*-online learnable classes have to be countable. To see this, consider the class $\mathcal{T}$ of all functions $t_a : [0, 1] \to \{0, 1\}$ with $a \in [0, 1]$ that has the following form:

$$t_a(x) = \begin{cases} 1, & \text{if } x = a \\ 0, & \text{otherwise} \end{cases} .$$

It is easy to see that $\mathcal{T}$ is strongly *e.a.s.*-online learnable (in fact it has Littlestone dimension 1), but the class is uncountable.

Strongly *e.a.s.*-online learnable classes can have infinite Littlestone dimension. For example, consider the class of all functions from $\mathbb{R} \to \{0, 1\}$ that take value 1 on *finitely* many inputs and zeros otherwise. Clearly, the class has infinite Littlestone dimension and the prediction that assigns 0 to any input (unless it has seen label 1 to the input before) only makes finitely many errors no matter what the distribution μ generates that samples. See also Example 12 in Section 6.7 for a more sophisticated example.

6.3.1 Effective countability and regularization

One can interpret the effective countability of a class $\mathcal{H}$ as the ability to impose a regularization on $\mathcal{H}$.

Fix any enumeration of the almost sure equivalence classes under the distribution μ over the instance space $\mathcal{X}$, say $\mathcal{H}' = \{h'_1, h'_2, \cdots\}$. Define the complexity of any $h \in \mathcal{H}$ to be the the minimum index i such that $\mu(\{X : h'_i(X) = h(X)\}) = 1$, where $h'_i \in \mathcal{H}'$.

In the learning process, we give higher priority to the hypotheses that have less complexity. Given training sample-label pairs $(x_1, h(x_1)), \cdots, (x_n, h(x_n))$, we choose a hypothesis $\hat{h} \in \mathcal{H}'$ such that

$$\hat{h} = \arg\min_{h'_k \in \mathcal{H}'} \{\mathrm{err}(h'_k) + k\},$$

where $\mathrm{err}(h'_k) = \sum_{i=1}^{n} 1\{h'_k(x_i) \neq h(x_i)\}$ is the empirical error made by h'_k. The additive term k in $\mathrm{err}(h'_k) + k$ is the penalty that balances the complexity of the hypothesis with the empirical errors. Therefore, $\hat{h}$ balances the empirical error with complexity. We then use $\hat{h}$ to predict the next sample.

6.4 Multiclass labels

We consider the case when the output set $\mathcal{Y}$ has more than 2 elements and we assume the loss ℓ to be an arbitrary function. Before analyzing the randomized setting, we consider a simpler deterministic setting. A deterministic sampling process in the most general setting is a function $\mathcal{S} : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{X}$, which selects the next sample based on the previous sample-label pairs (but not the learner's answer). For any class $\mathcal{H}$ and process $\mathcal{S}$, one can construct an infinite $|\mathcal{Y}|$-ary tree $\mathcal{T} = (\mathcal{V}, \mathcal{E})$ with a label function L (for both vertices and edges) such that

1. Each edge has a label in $\mathcal{Y}$ and the labels of edges with the same parent are distinct.

2. Each vertex in $\mathcal{T}$ has a label in $\mathcal{X}$. The root v_0 has label $L(v_0) = \mathcal{S}(\epsilon)$, where ϵ is the empty string. For each vertex $v \in \mathcal{T}$, we denote $v_0 \to v_1 \to \cdots v_n = v$ to be the unique path from root to v. Let $z_i = (L(v_{i-1}), L(v_{i-1} \to v_i))$. The label of v is given by $L(v) = \mathcal{S}(z_1, \cdots, z_n)$.

3. Each vertex v in $\mathcal{T}$ has a child u such that $L(v \to u) = y$ if and only if $\exists h \in \mathcal{H}$ such that $h(L(v_i)) = L(v_i \to v_{i+1})$ for all $0 \leq i \leq n$ and $h(L(v)) = y$, where v_i is defined in item 2.

Clearly, every function $h \in \mathcal{H}$ will be associated with a unique infinite path $v_0 \to v_1^h \to v_2^h \to \cdots$ in $\mathcal{T}$ such that $h(L(v_i^h)) = L(v_i^h \to v_{i+1}^h)$. However, not every infinite path will have a function in $\mathcal{H}$ that is associated with it. The learning process can be viewed as traversing an infinite path in $\mathcal{T}$. Such a tree is also known as the mistake tree in the binary-label online learning literature, see (Zhang and Chaudhuri 2016).

A valuation of $\mathcal{T}$ is a function $\mathbf{v} : \mathcal{V} \to \mathbb{N}$ such that for each $v \in \mathcal{V}$ we have

1. $\mathbf{v}(v) \geq \max_{u \in C(v)}\{\mathbf{v}(u)\}$, where $C(v)$ is the children vertex set of v;

2. $\mathbf{v}(v) \geq 1 + \max_{u \in C(v)}\{\mathbf{v}(u)\}$ if $L_v^M = \{L(v \to u) : u \in C^M(v)\}$ cannot be covered by $\mathcal{Y}$, where $C^M(v)$ is the set of children vertices of v that have *maximum* value. We say a set $A \subset \mathcal{Y}$ is covered in $\mathcal{Y}$ if there exists $y \in \mathcal{Y}$ such that $\forall x \in A,\ \ell(y, x) = 0$.

Note that such a valuation may not always exist. The following theorem relates the existence of a valuation to a bounded error guarantee.

Theorem 25. *A class $\mathcal{H}$ is online learnable with $\leq B$ errors and a deterministic sampling process $\mathcal{S}$ if and only if there exists a valuation $\mathbf{v}$ on $\mathcal{T}$ such that $\mathbf{v}(v_0) \leq B$, where v_0 is the root of $\mathcal{T}$.*

Proof. If $\mathcal{T}$ has a valuation such that $\mathbf{v}(v_0) \leq B$, we simply predict the element that would cover L_v^M at vertex v in the traversing (if no such element, predict anything). Note that the learner makes an error only if the value of the child vertex is reduced by at least 1. Therefore, there will be at most B errors.

If $\mathcal{H}$ is online learnable w.r.t. $\mathcal{S}$ with at most B errors, we define the value at vertex v to be the maximum number of errors that the learning rule could make on the infinite subtree rooted at v.

To prove this is a valid valuation, note that condition 1 can be verified easily. To see condition 2 holds as well, we resort to a proof by contradiction. If condition 2 does not hold at some vertex v, i.e., we have $\mathbf{v}(v) = \max_{u \in C(v)}\{C(v)\}$ but L_v^M cannot be covered by $\mathcal{Y}$. We can choose a path to a child in $C^M(v)$ that has label that differs from the prediction given by the learner, thus incurring $\max_{u \in C(v)} \mathbf{v}(u) + 1$ errors starting from v, contradicting our premise. The theorem follows. □

To capture the *finite* error guarantee, we will need a notion of ranking on the hypotheses in $\mathcal{H}$. A ranking of $\mathcal{H}$ is a function $\mathbf{r} : \mathcal{H} \to \mathbb{N}$. For any vertex $v \in \mathcal{V}$, we denote $\mathcal{H}_v$ to be the set of hypotheses in $\mathcal{H}$ that share the path from v_0 to v in $\mathcal{T}$, where v_0 is the root of $\mathcal{T}$. For a given ranking $\mathbf{r}$, we denote $\mathcal{H}_v^m$ to be the set of hypotheses in $\mathcal{H}_v$ that have *minimum* rank, i.e.,

$$\mathcal{H}_v^m = \{h \in \mathcal{H}_v : \mathbf{r}(h) = \min\{\mathbf{r}(\mathcal{H}_v)\}\}.$$

Now let $L_v^m = \{h(L(v)) : h \in \mathcal{H}_v^m\}$. Note that here L_v^m is different from the L_v^M that we defined above.

We say a class $\mathcal{H}$ is *rankable* w.r.t. a deterministic sampling process $\mathcal{S}$, if there exists a ranking $\mathbf{r}$ such that for each $h \in \mathcal{H}$ there exists a number N_h, such that for every vertex $v \in \mathcal{V}$ on the infinite path of h with depth larger than N_h, we have $h \in \mathcal{H}_v^m$ and L_v^m is covered in $\mathcal{Y}$.

Intuitively, the smaller the rank of a function in $\mathcal{H}$, the higher priority we will assign in the learning process. Formally, we prove the following theorem:

Theorem 26. *A class $\mathcal{H}$ is online learnable with finitely many errors w.r.t. a deterministic sampling process $\mathcal{S}$ if and only if $\mathcal{H}$ is rankable w.r.t. $\mathcal{S}$.*

Proof. If $\mathcal{H}$ is rankable, we show that $\mathcal{H}$ is online learnable with finitely many errors. The prediction rule works as follows, we predict an element that covers L_v^m at each vertex v

in the traversing (if it doesn't exist we predict anything). By definition of rankability, the learner will not make errors after step N_h if the underlying hypothesis is h.

To see the converse, suppose Φ is an online learning rule which makes only finitely many errors no matter what the hypothesis $h \in \mathcal{H}$ is. The rank of each hypothesis h is the number of errors Φ would make against h.

To conclude the proof, we will show that this is a valid ranking. Fix any h and let N_h be the time step of the last error Φ makes on h. Suppose there exists some vertex v at depth more than N_h such that L_v^m is not covered in $\mathcal{Y}$. Now the rule Φ will make more errors on at least one of hypotheses $h' \in \mathcal{H}_v^m$ than it made on h. This is because h and h' share the same prefix up to vertex v and hence have the same error pattern till then, but differ on their label in L_v^m. Thus this hypothesis h' must have a higher rank by definition, and therefore could not have been in $\mathcal{H}_v^m$. □

Note that both Theorem 25 and Theorem 26 work only for the deterministic sampling process where there is one mistake tree. However, in the randomized setup the mistake tree will be different for different realization of the sampling. A sufficient condition for making finitely many errors in the randomized setting is to have a universal ranking that works for almost all realizable mistake trees. The following theorem shows that it will happen when the class is effectively countable.

Theorem 27. *Let $\mathcal{H}$ be a set of measurable functions from $\mathcal{X} \to \mathcal{Y}$. Then $\mathcal{H}$ is e.a.s.-online learnable w.r.t. distribution μ and loss ℓ, if $\mathcal{H}$ is effectively countable w.r.t. μ and ℓ.*

Moreover, the effective countability of $\mathcal{H}$ is necessary for e.a.s.-online learnability (w.r.t. any distribution μ) if $\mathcal{Y}$ is countable and ℓ is the classification loss, i.e., $\ell(y_1, y_2) = 1\{y_1 \neq y_2\}$.

Proof. The sufficiency follows directly from Lemma 15. To prove the necessary condition, by the argument in the proof of Theorem 24, it is sufficient to show that there exists a

countable dense subset of measurable functions from $\mathcal{X} \to \mathcal{Y}$ with the following metric

$$d(f, g) = \Pr_{x \sim \mu}(f(x) \neq g(x)).$$

To do so, we use a truncation argument. W.l.o.g., we assume $\mathcal{Y} = \mathbb{N}$. We observe that there exists a class of countable measurable functions $\mathcal{P}_m$ that is dense in the measurable functions from $\mathcal{X} \to [m]$ for all $m \in \mathbb{N}$, since the σ-algebra on $\mathcal{X}$ is separable. We claim that $\bigcup_{m=1}^{\infty} \mathcal{P}_m$ is dense in the measurable functions from $\mathcal{X} \to \mathbb{N}$. Let $h : \mathcal{X} \to \mathbb{N}$ be an arbitrary measurable function and $\epsilon > 0$. There exists $m \in \mathbb{N}$ such that

$$\Pr_{x \sim \mu}(h(x) \geq m) \leq \epsilon/2.$$

Let h^m be the function such that $h^m(x) = h(x)$ if $h(x) \leq m$ and $h^m(x) = 1$ otherwise. We have $d(h^m, h) \leq \epsilon/2$. Now, choosing $h' \in \mathcal{H}_m$ with $d(h^m, h') \leq \epsilon/2$, we have $d(h, h') \leq \epsilon$. This completes the proof. □

Corollary 4. *Let $\mathcal{H}$ be the class of all continuous functions from $[0, 1] \to [0, 1]$, $\ell = 1\{|y_1 - y_2| \geq B\}$ for some constant $B > 0$. Namely, we have to predict the label of a continuous function at a given instance, and we incur a loss if the difference between the prediction and true label $\geq B$. Then $\mathcal{H}$ is e.a.s.-online learnable w.r.t. the loss ℓ and any distribution.*

Proof. By the Stone-Weierstrass theorem (Rudin 1964, Theorem 7.26), $\mathcal{H}$ can be covered by polynomials with rational coefficients under the supremum norm. Since polynomials with rational coefficients are countable, $\mathcal{H}$ is effectively coverable w.r.t. any distribution. Theorem 27 then implies that $\mathcal{H}$ is *e.a.s.*-online learnable w.r.t. any distribution. □

Remark 13. *Note that the requirement of a compact domain such as $[0, 1]$ is required for the proof of Corollary 4 to work. Indeed, if we consider the class of all continuous functions from $(0, 1] \to [0, 1]$ then a countable covering will not necessarily be possible. To see this, we define $\mathcal{H} = \{\sin(2\pi\alpha/x) : \alpha \in [0, 1]\}$. Let μ be an arbitrary distribution supported on*

the whole of the set $S = \{1/n : n \in \mathbb{N}^+\}$. *We show that* $\mathcal{H}$ *cannot be effectively covered by any countable set w.r.t.* μ *and the loss in Corollary 4 with* $B = \frac{1}{4}$. *Suppose such a covering exists, there must be a uncountable subset* $\mathcal{H}' \subset \mathcal{H}$ *such that*

$$\forall h_1, h_2 \in \mathcal{H}',\ \forall x \in S,\ |h_1(x) - h_2(x)| \leq \frac{1}{2},$$

since $\mathcal{H}$ *is uncountable. Let* $\mathcal{A}$ *be the set of parameters of functions in* $\mathcal{H}'$. *Since* $\mathcal{A}$ *is uncountable, there must be irrational numbers* $\alpha, \beta \in \mathcal{A}$ *such that* $1, \alpha, \beta$ *are linearly independent over rationals. Now, since the vector sequence* $\{(\alpha n, \beta n)\}_{n \in \mathbb{N}^+}$ *is uniformly distributed mod 1 (Kuipers and Niederreiter 2012, Example 6.1), there must be some* n *such that* $|\sin(2\pi\alpha n) - \sin(2\pi\beta n)| > \frac{1}{2}$. *This contradicts our premise on* $\mathcal{H}'$.

Remark 14. *The covering set in Corollary 4 is independent of the distribution. Therefore, we know that the class of all continuous functions in Corollary 4 is actually strongly e.a.s.-online learnable. Note that the set of all continuous functions from* $[0,1] \to [0,1]$ *is quite large and includes functions such as Cantor's function (or the* devil's staircase*). Cantor's function,* $c(x) : [0,1] \to [0,1]$, *is a function that assigns*

$$c(x) = \begin{cases} \sum_{n=1}^{\infty} \frac{a_n}{2^n},\ x = \sum_{n=1}^{\infty} \frac{2a_n}{3^n} \quad \textit{with } a_n \in \{0,1\} \\ \sup_{y \leq x,\ y \in \mathcal{C}} c(y),\ \textit{otherwise} \end{cases},$$

where $\mathcal{C} = \{\sum_{n=1}^{\infty} \frac{2a_n}{3^n} : a_n \in \{0,1\}\}$ *is Cantor's set. Yet, we have shown that the class is still strongly e.a.s.-online learnable under the loss in Corollary 4.*

Note also that we can't extend Corollary 4 to the class of all measurable functions from $[0,1] \to [0,1]$, *since even the set of all linear threshold functions over* $[0,1]$ *can't be e.a.s.-online learnable w.r.t. the loss in Corollary 4 with* $B = \frac{1}{4}$ *and uniform distribution over* $[0,1]$.

However, effective countability is not always necessary for *e.a.s.*-online learnability in the non-binary label setting ruling out strengthening Theorem 27 along the lines of Theorem 24.

We provide the following example, which shows that if we have $|\mathcal{Y}| \geq 3$ and $\mathcal{X} = \mathbb{N}$, then there exist class $\mathcal{H}$, distribution p and loss ℓ, such that $\mathcal{H}$ is *e.a.s.*-online learnable w.r.t. p and ℓ, but $\mathcal{H}$ is *not* effectively countable w.r.t. p and ℓ. Thus, the condition in Theorem 27 can't be necessary for arbitrary losses and distributions.

Example 11. *Let $\mathcal{X} = \mathbb{N}$ and $\mathcal{Y} = \{0, 1, 2\}$. We define a loss ℓ as follows: ℓ is symmetric in its two arguments and*

$$\ell(0, 1) = 0 \text{ and } \ell(0, 2) = \ell(1, 2) = 1.$$

We now construct the class $\mathcal{H}$ as follows. Since the domain is $\mathbb{N}$, we will denote each function in $\mathcal{H}$ as infinite sequences in $\{0, 1, 2\}^\infty$. Let $\mathcal{B} = \{0, 1\}^\infty$ be the class of all binary sequences. We define the following transformation T that maps $\{0, 1\}^\infty \to \{0, 1, 2\}^\infty$. For any $\boldsymbol{b} \in \mathcal{B}$, $T(\boldsymbol{b})$ is the sequence that inserts number 2 following each appearance of 0 in $\boldsymbol{b}$. For example, if $\boldsymbol{b} = 00100101011\ldots$ then $T(\boldsymbol{b})$ would read $020210202102102 11\ldots$. Now, we define

$$\mathcal{H} = \{T(\boldsymbol{b}) : \boldsymbol{b} \in \mathcal{B}\}.$$

For example, the sequence $020210202102102 11\ldots$ corresponds to a hypothesis $h \in \mathcal{H}$ such that $h(1) = 0, h(2) = 2, h(3) = 0, \cdots$.

We now define the distribution p. By Lemma 16, we have distribution p over $\mathbb{N}$ such that the i.i.d. *samples from p will appear in increasing order of $\mathbb{N}$ eventually almost surely—namely for all m large enough, the instance $m - 1$ will appear prior to instance m with probability 1. Therefore, one may assume, w.l.o.g., that the sample is generated sequentially as $1, 2, 3, \cdots$.*

To see that the class $\mathcal{H}$ is e.a.s.-online learnable w.r.t. to p and loss ℓ, for any instance $m \in \mathbb{N}$ we simply predict 2 for the label of instance m if the label of instance $m - 1$ is 0 and predict 0 if the label of $m - 1$ is 1 or 2. It is easy to check that the rule makes only finitely many errors with probability 1. Since any appearance of label 0 must be followed by label 2 and an error on label 0 or 1 incurs zero loss.

Clearly, the class $\mathcal{H}$ is uncountable *and for any two different sequences $\boldsymbol{b}_1, \boldsymbol{b}_2 \in \mathcal{B}$ we can find $l \in \mathbb{N}$ such that $\ell(T(\boldsymbol{b}_1)(l), T(\boldsymbol{b}_2)(l)) = 1$. Therefore, we have $\mathcal{H}$ is* not *effectively coverable by any countable set $\mathcal{H}'$.*

Even though the class in Example 11 is not effectively countable, it has a trivial universal ranking that ranks all functions in $\mathcal{H}$ to 0. We conclude this section with the following conjecture.

Conjecture 2. *A class $\mathcal{H}$ is e.a.s.-online learnable w.r.t μ iff there exists a universal ranking $\boldsymbol{r}$ of $\mathcal{H}$ such that, with probability* 1 *under the* i.i.d. *sampling of μ, $\mathcal{H}$ is rankable with ranking $\boldsymbol{r}$ for the mistake trees of the realizations of the sample.*

6.4.1 Rankability and regularization

As can be expected, rankability is a different angle to view the ability to have a regularization. Recall that a class $\mathcal{H}$ is rankable w.r.t. a deterministic sampling process if there exists a ranking $\mathbf{r} : \mathcal{H} \to \mathbb{N}$ such that for any hypothesis $h \in \mathcal{H}$, when we observe sufficiently long samples, the set of minimum rank hypotheses that are also consistent with the current samples must include h and (the labels on the next instance) must be coverable under ℓ. We can therefore interpret the rank as a measure of complexity on the class $\mathcal{H}$ that assigns priority on the hypotheses of $\mathcal{H}$, and our learning process balances this complexity against empirical error to choose hypotheses.

As we have shown in Example 11, a rankable class need not necessarily be effectively countable, so rankability is a more general notion of complexity than effective countability. But where the $\mathcal{H}$ is also effectively countable w.r.t. some distribution μ and loss ℓ, then any enumeration of the almost sure equivalence classes $\mathcal{H}' = \{h'_1, h'_2, \cdots\}$ will define a natural ranking on $\mathcal{H}$.

To do so, for any $h \in \mathcal{H}$ we define $\mathbf{r}(h) = \min\{k \geq 1 : \Pr_{X\sim\mu}[\ell(h(X), h_k(X))] = 0\}$. It is easy to verify $\mathcal{H}$ is rankable using rank function $\mathbf{r}$ for almost all (i.e., with measure 1 under μ) realizations of *i.i.d.* sampling from μ. For example, in Corollary 4 the rank of any

continuous function h is defined to be the minimum index (with any predefined ordering) of the rational coefficient polynomials that cover h under the loss ℓ in Corollary 4.

6.5 Computable and computationally bounded predictors

As we have mentioned in Chapter 2, our setup has a close connection with the *learning of recursive functions*, where one considers the domain to be $\mathbb{N}$ and the instances are presented sequentially in a predetermined ordering of $\mathbb{N}$. A learner predicts the label of next sample and incurs the classification loss, but in addition, the learner is required to be computable (see Definition 14 below) as well.

Definition 14. *A function $h : \mathbb{N} \to \{0, 1\}$ is* computable *if there exists a Turing machine $\boldsymbol{TM}$ such that $\forall n \in \mathbb{N}$, $\boldsymbol{TM}(n)$ halts and outputs $h(n)$, where the number n is in its binary representation.*

Clearly the class of all computable functions from $\mathbb{N} \to \{0, 1\}$ is countable, since there are only countably many Turing machines (Arora and Barak 2009, Chapter 1). We have the following simple corollary that follows from Lemma 15.

Corollary 5. *Let $\mathcal{H}$ be the set of all computable functions from $\mathbb{N} \to \{0, 1\}$, and let p be an arbitrary distribution over $\mathbb{N}$. Then $\mathcal{H}$ is e.a.s.-online learnable using* i.i.d. *samples generated from p.*

Unfortunately, the learning rule that is derived from Lemma 15 cannot be computable in general as shown in the following theorem of Barzdiņš and Freivald (1972), which we first recall in our notation.

Theorem 28 (Barzdins-Freivald, see (Zeugmann and Zilles 2008, Thm 5))**.** *Let $\mathcal{H}$ be a class of computable functions from $\mathbb{N} \to \{0, 1\}$. Then there exists a computable learner that e.a.s.-online learns $\mathcal{H}$ with sequential sampling if and only if there exists a computable function $g : \mathbb{N} \to \mathbb{N}$ such that the time complexity of each function in $\mathcal{H}$ is eventually dominated by $g(n)$.*

It is easy to show that if $\mathcal{H}$ is the class of *all* computable functions from $\mathbb{N} \to \{0,1\}$, then g in Theorem 28 cannot exist by the time hierarchy theorem (Arora and Barak 2009, Theorem 3.1). Therefore, the class of all computable functions from $\mathbb{N} \to \{0,1\}$ is not *e.a.s.*-online learnable with sequential sampling $\{1,2,\cdots\}$. We show in the following theorem that this holds even for *i.i.d.* sampling from certain distributions over $\mathbb{N}$.

Theorem 29. *There exists a distribution p over $\mathbb{N}$ such that the class of all computable functions from $\mathbb{N} \to \{0,1\}$ is not computationally e.a.s.-online learnable with* i.i.d. *sampling from p.*

Proof. Denote $\mathcal{H}$ to be the class of all computable functions from $\mathbb{N} \to \{0,1\}$. Let p be any distribution over $\mathbb{N}^+$ with $p(n) = \Theta(\frac{1}{(n!)^2})$ such that $1/p(n)$ are integers (assign extra mass to 0 if necessary) and computable. By Lemma 16, we know that, with probability 1, any sufficiently large n will appear later than the appearance of $n-1$ in the *i.i.d.* sampling of p. Moreover, with probability 1, any sufficient large n must appear no latter than $\Omega((n!)^2 \log n)$ sampling steps.

Suppose $\mathcal{H}$ is computationally *e.a.s.*-online learnable w.r.t. p. Let Φ be the computable prediction rule. We now construct a computable prediction rule using Φ that *e.a.s.*-online learns $\mathcal{H}$ with sequential sampling $\{1,2,\cdots\}$. This will contradict Theorem 28.

To do so, for any sample n and labels $h(1),\cdots,h(n-1)$, we simulate Φ for upto $n \cdot (n!)^2$ sampling steps until the sample n first appears and predict the outcome of the simulation (if the sample n did not appear within $n \cdot (n!)^2$ steps, we predict anything). Note that, here we only need to simulate the sampling process with the restricted distribution of p on $\{1,\cdots,n\}$. Clearly the rule is computable, since p is computationally samplable and $n \cdot (n!)^2$ is computable. By the *e.a.s.*-online learnability of Φ and the property of p, the rule will make finitely many errors almost surely (over the internal randomness) for any underlying function $h \in \mathcal{H}$.

Note that the prediction rule we constructed above is randomized. We now convert it to a deterministic prediction rule in the following way. For any n, there is a computable upper bound $B(n)$ on the random bits that is used by the randomized predictor. We run

the predictor on all the $2^{B(n)}$ binary sequences of length $B(n)$ and output the majority. Since the randomized prediction rule makes finitely many errors almost surely. For any $h \in \mathcal{H}$ there must be some N such that w.p. $> \frac{1}{2}$ the rule makes no errors on h after step N. Therefore, the deterministic rule must make no errors on h after step N. □

Still, we show in the following theorem that there are *non-degenerate* distributions over $\mathbb{N}$ such that the class of all computable functions is indeed computationally *e.a.s.*-online learnable using *i.i.d.* samples from such distributions.

Theorem 30. *Let $\mathcal{H}$ be the class of all computable functions from $\mathbb{N} \to \{0,1\}$. Then there exists a non-degenerate distribution (i.e., with infinite support) q such that $\mathcal{H}$ is computationally e.a.s.-online learnable w.r.t. q.*

Proof. Let $\mathbf{TM}_1, \mathbf{TM}_2, \cdots$ be a fixed enumeration of all Turing machines. The main idea is to construct a distribution q such that for almost all numbers n, n appears much later than any of $\{1, ..., n-1\}$ with probability 1. We choose the gap period between the first appearances of $n-1$ and n to be $\max\{C(h_i(n) : i \leq n\}$, where $C(h_i(n))$ is the computational time for the ith computable function with input n to stop when computing with a feasible Turing machine of smallest index. Such a distribution q exists by Lemma 16.

We now construct the computable predictor using a *back and forth* approach. The predictor goes as follows:

1. Initialize index $I, J = 1$;

2. In the idle period between the appearance of $n-1$ and n, the predictor does the following. It *simulates* the computation of $\mathbf{TM}_I$ on n with one computational step per time step. (For other samples that are encountered in that period, one simply predicts the memorized labels.)

3. At the time step of first observing n, output the result of the simulation. (Output is arbitrary if the simulation has not stopped.) If the result matches with the true label, keep I, J. Else:

a. If $I < J$, set $I = I + 1$ and $J = J$;

b. Else, set $I = 1$ and $J = J + 1$.

Since the underlying function is computable, there exists a Turing machine $\mathbf{TM}_t$ that computes it. Now, by construction the index I changes if and only if the predictor makes an error.

We show that the predictor only makes finitely many errors by a proof by contradiction. If indeed the predictor makes infinite errors, we know that I would repeatedly hit t until the sample is coming sequentially. By construction of the idle time, $\mathbf{TM}_t$ would then finish the computation and make the right prediction in the following time steps, which is a contradiction. □

Remark 15. *Note that the reason why Theorem 30 does not contradict the impossibility result implied by Theorem 28 is that, even though the gap periods we constructed in the proof of Theorem 30 are information theoretically deterministic, they are not (computationally) known to the learner, i.e., the learner would never be able to (computationally) figure out when the sample "n" will arrive.*

6.5.1 Exact online learning with computationally bounded predictors

We now consider a different scenario where we require the learner to be not only computable, but also limit its computational resources. We need the following notion. An *online learning scheme* is a triplet $(\mathcal{D}, \mathbf{R}, \Phi)$, where

1. $\mathcal{D} \in \{0,1\}^*$ is an *unlimited* database which the learner can use to store anything learned, initially an empty string;

2. Φ is the predictor which maps $\mathbb{N} \to \{0,1\}$, using the database $\mathcal{D}$ as an oracle;

3. $\mathbf{R}$ is the recorder that maps $\{0,1\}^* \times \mathbb{N} \times \{0,1\} \to \{0,1\}^*$, which updates the database $\mathcal{D}$ every time a new example and its label are revealed. Where needed, we use $\mathcal{D}_n$ to denote the state of the database after n instances and their labels have been revealed.

Definition 15 (Exact online learnable)**.** *Let $\mathcal{H}$ be a class of functions from $\mathbb{N} \to \{0,1\}$. $\mathcal{H}$ is said to be exact online learnable, if there is an* online learning scheme $(\mathcal{D}, \boldsymbol{R}, \Phi)$ *such that for all $h \in \mathcal{H}$*

$$\sum_{n=1}^{\infty} 1\{\Phi^{\mathcal{D}_n}(n) \neq h(n)\} < \infty,$$

where $\mathcal{D}$ updates after every n,

$$\mathcal{D}_{n+1} = \boldsymbol{R}(\mathcal{D}_n, n, h(n)).$$

For any computable online learning scheme $(\mathcal{D}, \mathbf{R}, \Phi)$, we will specify $\mathbf{R}$ and Φ as algorithms. The time complexity of $\mathbf{R}$, Φ and h is expressed in terms of $\log n$, i.e., the binary representation size of n. However, we use $\mathcal{D}$ as an oracle of $\mathbf{R}$ and Φ with no computational cost.

We focus on the worst case time complexity for $\mathbf{R}$ and Φ, *i.e.*, over the most inconvenient database and function we are trying to learn. We say an online learning scheme runs in *uniform exponential time*, if there exist some $c, N \in \mathbb{N}$ such that no matter what database $\mathcal{D}$ is and what function h is in force, both $\mathbf{R}$ and Φ run in $\exp(c \log n) = n^c$ time for all $n \geq N$. We define *uniform polynomial time* similarly.

Theorem 31. *Let $\mathcal{H}$ be the class of all exponential time computable functions from $\mathbb{N} \to \{0,1\}$. Then there is no computable online learning scheme $(\mathcal{D}, \boldsymbol{R}, \Phi)$ that exactly learns $\mathcal{H}$ such that both $\boldsymbol{R}$ and Φ run in uniform exponential time.*

Proof. We actually prove a stronger version of the theorem. Let $\mathcal{H}$ be the set of functions that can be computed within time n^{k+1} (hence exponential in $\log n$), we show that there is no online learning scheme $(\mathcal{D}, \mathbf{R}, \Phi)$ that exactly learns $\mathcal{H}$ such that $\mathbf{R}$ and Φ uniformly runs in $n^k/2$ time.

The proof uses a diagonalization argument. Assume to the contrary that we have such a scheme $(\mathcal{D}, \mathbf{R}, \Phi)$. We construct the following algorithm **ExpDiag**(n):

1. **Input**: n

2. **If** $n = 1$ **Output** 1.

3. Run scheme $(\mathcal{D}, \mathbf{R}, \Phi)$ on samples (k,**ExpDiag**(k)) with $k \leq n-1$.

4. **Output** $1 - \Phi^{\mathcal{D}}(n)$.

To analyze the running time of **ExpDiag**(n), let $f(n)$ be the time needed to compute **ExpDiag** with input n. We have

$$f(n) = n^k + f(n-1),$$

since $\mathbf{R}$ and Φ run in $n^k/2$ time by assumption and one may reuse the database at the recursion steps. We have

$$f(n) \leq \sum_{i=1}^{n-1} i^k \leq n^{k+1}.$$

Therefore, the function that is computed by **ExpDiag** is in $\mathcal{H}$. However, by construction **ExpDiag**$(n) \neq \Phi^{\mathcal{D}}(n)$ for all $n \geq 2$ which yields a contradiction. □

Remark 16. *Note that similar argument cannot be generalized to the functions that are computed in time* $\log^{k+1} n$ *against predictors runs in* $\log^k n$ *because establishing the database in step* 3 *will require* $\Omega(n)$ *steps in the naive way. However, we still believe this is true as in Conjecture 3 below. In Theorem 32 below we will establish such a conjecture when the database simply appends the true labels without processing.*

We make the following conjecture for polynomial time computable functions.

Conjecture 3. *Let* $\mathcal{H}$ *be the class of all polynomial time computable functions from* $\mathbb{N} \to \{0,1\}$. *Then there is no computable online learning scheme* $(\mathcal{D}, \boldsymbol{R}, \Phi)$ *that exactly learns* $\mathcal{H}$ *such that both* $\boldsymbol{R}$ *and* Φ *uniformly run in polynomial time.*

While we are unable to prove the conjecture, we can prove the following specialized version. We say the recorder $\mathbf{R}$ to be a naive recorder, if it simply appends $h(n)$ to the database $\mathcal{D}$ at the update of step n.

Theorem 32. *Let $\mathcal{H}$ be the class of all functions from $\mathbb{N} \to \{0,1\}$ whose time complexity is eventually bounded by $\log^{k+1} n$ time for some $k \geq 1$. Then there is no computable online learning scheme $(\mathcal{D}, \boldsymbol{R}, \Phi)$ that exactly learns $\mathcal{H}$ such that $\boldsymbol{R}$ is a naive recorder and Φ runs in uniform $\log^k n$ time.*

Proof. We use a standard approach that reduces the polynomial to exponential case. Let $\mathcal{H}$ be the class of all functions that can be computed in time $\log^{k+1} n$. Suppose to the contrary, $(\mathcal{D}, \mathbf{R}, \Phi)$ is a scheme that exactly learns $\mathcal{H}$ such that $\mathbf{R}$ is a naive recorder and Φ runs in $\log^k n$. For any function h' that can be computed in n^{k+1} time, we construct a function h that can be computed in $\log^{k+1} n$ time as follows

$$h(n) = \begin{cases} h'(t), \text{ if } n = 2^t \text{ for some } t \in \mathbb{N} \\ 1, \text{ otherwise} \end{cases}.$$

Consider the following predictor Φ':

1. **Input**: n and naive recording $\mathcal{D}$ of h' upto $n-1$
2. Simulate Φ with input 2^n as follows: if Φ queries database at position T such that $T = 2^t$ we query the tth position in $\mathcal{D}$. In all other positions, we know h took value 1.
3. **Output** $\Phi^{\mathcal{D}}(2^n)$.

By assumption Φ makes finitely many errors on h, thus Φ' makes finitely errors on h' as well. Clearly, we have Φ' runs in $(\log 2^n)^k = n^k$ time and it exactly learns the class of functions that can be computed in n^{k+1}. This contradicts Theorem 31. □

We note the following interesting connection with time hierarchy theorem. Denote

$$\mathcal{H}_k = \{h : \exists \mathbf{TM} \; s.t. \; \forall n \in \mathbb{N}, \mathbf{TM}(n) = h(n) \text{ and } \mathbf{C}(\mathbf{TM}, n) \leq \log^k n\},$$

where $\mathbf{C}(\mathbf{TM}, n)$ is the running time of Turing machine $\mathbf{TM}$ on input n.

Corollary 6. *Theorem 32 implies that $\mathcal{H}_{k+2} \backslash \mathcal{H}_k$ is not empty.*

Proof. Suppose not. We will construct an online learning scheme $(\mathcal{D}, \mathbf{R}, \Phi)$ that exact learns $\mathcal{H}_{k+2}$ such that $\mathbf{R}$ is naive recorder and Φ runs in $\log^{k+1} n$, which will contradict Theorem 32.

To do so, we enumerate all Turing machines $\mathbf{TM}_1, \mathbf{TM}_2, \cdots$. The predictor Φ maintains two indices t, i. At each time step n, both t and i are initialized to 1. The predictor *simulates* $\log^k i$ instructions on $\mathbf{TM}_t(i)$. If it stops within those $\log^k i$ instructions and its output matches $h(i)$, keep $t = t$ and increment $i = i + 1$. Else, move to the next machine, $t = t + 1$ and increment $i = i + 1$.

This process continues for $\frac{1}{2}\log^{k+1} n$ net time (this includes the time taken to simulate steps on all Turing machines thus far, and all relevant overheads, including that for incrementing indices). Then set $i = n$ and simulate $\mathbf{TM}_t(n)$ till the run for a net $\log^{k+1}(n)$ time. If $\mathbf{TM}_t$ stops by then, the predictor outputs $\mathbf{TM}_t(n)$, otherwise it outputs 1.

We show that this scheme is an exact online learning scheme for $\mathcal{H}_{k+2}$. For any $h \in \mathcal{H}_{k+2}$, we have an Turing machine $\mathbf{TM}_j$ that outputs $h(n)$ for all n within $\log^k n$ time, since $\mathcal{H}_{k+2} = \mathcal{H}_k$ by assumption. Note that t increases iff the predictor makes errors. Since in addition, $\frac{1}{2}\log^{k+1} n \to \infty$, we know that t will hit j eventually. Once t hits j, note that we never increment it since $\mathbf{TM}_j(i)$ outputs $h(i)$ within $\log^k i$ time for all i. If the time runs out before we complete the simulation of $\mathbf{TM}_j$, note again that t is not incremented. Finally, since the overhead on universal Turing machine is a $\log\log n$ factor on time complexity (Arora and Barak 2009, Theorem 1.13), we know that for large enough n, the algorithm above actually completes the simulation of $\mathbf{TM}_j(n)$. □

Remark 17. *The time hierarchy theorem (Arora and Barak 2009, Theorem 3.1) does not necessarily imply Theorem 32, since the predictor uses the database $\mathcal{D}$ as an oracle when computing on n.*

6.6 Noisy labels

In the previous sections we considered various setups in the eas-online learning paradigm when the labels are presented accurately. However, it is possible that the labels are corrupted by some noise. Perhaps surprisingly, we will show in this section that our eas-online learning paradigm is actually resistant to independent noises even if the samples are presented sequentially.

For simplicity, we will assume the domain to be $\mathcal{X} = \mathbb{N}$ and that the label is binary. We consider the following noisy process. At each time step i, we denote $\tilde{Z}_i = (X_i, h(X_i) \oplus R_n)$ to be the noisy sample-label pair, where R_n is a binary random variable with $\Pr(R_n = 1) \leq \eta$ and is independent of the instances and labels. In addition, $\{R_n\}_{n\in\mathbb{N}}$ are independent for each n.

We prove the following result.

Theorem 33. *Let $\mathcal{H}$ be a function class from $\mathbb{N} \to \{0, 1\}$, p is a distribution over $\mathbb{N}$. If $\mathcal{H}$ is effectively countable w.r.t. p and $\eta < \frac{1}{2}$, then there exists a learning strategy Φ, such that for all $h \in \mathcal{H}$ we have*

$$\Pr\left(\sum_{n=1}^{\infty} \ell(\Phi(\tilde{Z}_1^{n-1}, X_n), h(X_n)) < \infty\right) = 1,$$

where the randomness comes from both the sample and noise.

Proof. Let $\mathcal{H}'$ be a countable class that effectively covers $\mathcal{H}$ such that any two distinct functions in $\mathcal{H}'$ differs by a positive measure set in $\mathbb{N}$. Let $\mathcal{H}'_1 \subset \mathcal{H}'_2 \subset \cdots \mathcal{H}'$ be an nesting such that $|\mathcal{H}'_k| = k$ and $\bigcup_{k\in\mathbb{N}} \mathcal{H}'_k = \mathcal{H}'$.

The prediction happens in stages. At stage 0, we initialize by choosing an arbitrary function h_0 from $\mathcal{H}'$. Let h_{k-1} be the function we have after stage $k-1$. We will use h_{k-1} to make the prediction in stage k. At stage k we try to identify the underlying function as if it were in $\mathcal{H}'_k$. To do so, we note that each function in $\mathcal{H}'_k$ differs from other functions by a positive measure subset of $\mathbb{N}$. One observes the samples sufficiently long so that there

are m_k samples (instances may be repeated) in the difference sets of each pair of functions. We define $h_k \in \mathcal{H}'_k$ to be a function that has a smaller Hamming distance to the noisy labeling of samples at the difference positions relative to all other functions in $\mathcal{H}'_k$. If no such function exists, choose an arbitrary function in $\mathcal{H}'_k$.

We now show that the strategy indeed works. To do so, we choose $m_k = \frac{3 \log k}{2(0.5-\eta)^2}$. By the Hoeffding bound and union bound, with probability at least $1 - \frac{1}{k^2}$, we will find the correct underlying function at stage k, if it is in $\mathcal{H}'_k$. Now, since the underlying function must be in some $\mathcal{H}'_t$, the predictor must find it in finite steps. And by the Borel-Cantelli lemma, we will only miss it finitely many times with probability 1 since $\sum_{k \in \mathbb{N}} \frac{1}{k^2} < \infty$. After that we will make no errors. The theorem follows. □

In the proof of Theorem 33, the exact probability $\Pr(R_n = 1)$ is not required to be known, only the upper bound η need be known. However, even the requirement of knowledge of η can be eliminated by using confidence parameters $\{\eta_k\}$ at each stage k where $\eta_k \to 0$ as $k \to \infty$. Note that Theorem 33 also holds if the instances are presented sequentially. In such a case, one observes for each instance exactly one noisy label. Therefore, there is no way to estimate any single label (which would have been possible in the *i.i.d.* case). However, the proof of Theorem 33 shows that one would be able to identify the correct function with arbitrary high confidence by leveraging noisy labels of different instances.

6.7 Other variations

Contrast to the finite error guarantee, perhaps a more natural non-uniform consistency of online learning one may think is to let the average loss per time step converges to zero. More formally, one may wish to find a predictor such that

$$\Pr\left(\lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \ell(\Phi(Z_1^{n-1}, X_n), h(X_n)) = 0\right) = 1.$$

However, such a guarantee might not be attractive if one only considers the consistency. Since we can show that the class of all measurable functions from $\mathbb{R} \to \{0,1\}$ is actually learnable in that sense.

Theorem 34. *Let $\mathcal{H}$ be the class of all measurable functions from $\mathbb{R} \to \{0,1\}$, μ is an arbitrary distribution over $\mathbb{R}$ that is unknown to the learner, ℓ is the classification loss. Then there exists a learning strategy Φ such that for all $h \in \mathcal{H}$ we have*

$$\Pr\left(\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \ell(\Phi(Z_1^{n-1}, X_n), h(X_n)) = 0\right) = 1,$$

where Z_i and X_i are defined as in Section 6.2.

Sketch of Proof. Let $\mathcal{G}$ be a countable class of function from $\mathbb{R} \to \{0,1\}$ that is dense in $\mathcal{H}$. Note that, since the Borel σ-algebra is separable, we know that such a class exists and is independent of the underlying distribution. The predictions are partitioned into phases. At phase k, the learner tries to identify a function h_k in $\mathcal{G}$ that is ϵ_k-close to the underlying function with confidence at least $1 - \frac{1}{k^2}$. This can be easily achieved by using a structural risk minimization argument. Note that we will use h_{k-1} to make the predictions at phase k, and use the samples observed at phase k to find h_k for the next phase. We now choose ϵ_k small enough so that the probability that the partial average errors jump above $1/k$ in the infinite horizon is at most $\frac{1}{k^2}$. This can be done by, e.g., taking $\epsilon_k = \frac{1}{k^3}$. The claim follows by using multiplicative Chernoff bound with a union bound and observing that $\sum_{n=1}^{\infty} \exp(-2nk) = O(\frac{1}{k^2})$. The theorem now follows by the Borel-Cantelli lemma, since $\frac{1}{k^2}$ is summable. □

One may also consider the scenario when the rate of convergence is controlled. It can be shown that if the class has finite VC-dimension, then we can have a rate of $O(\frac{\log n}{n})$ in expectation, see Haussler et al. (1994). However, this does not automatically give us an almost sure upper bound on the cumulative errors, since the errors at different time steps are *correlated*. We now show that we can indeed achieve an $O(\frac{\log^2 n}{n})$ rate almost surely.

To do so, we use a doubling trick. We partition the prediction into phases. At phase k, we will see 2^k sample-label pairs, and we generate a hypothesis h_k using the observed pairs. We then use h_k to make predictions for another 2^k steps, at the end of which we move to phase $k+1$. To see why this approach works, we denote I_n to be the indicator that an error occurred at step n. By construction, we know that the indicators in each phase are *independent*. We can therefore employ multiplicative Chernoff bound to show that with probability at least $1 - O(1/2^k)$ the number of errors at phase k is upper bounded by $O(k)$ (using empirical risk minimization to obtain h_k is sufficient to achieve this). Therefore, by the Borel-Cantelli lemma, the partial sums of the indicators $\{I_n\}$ up to phase k will be eventually upper bounded by $O(k^2)$ with probability 1. Since the first n steps will cover at most $\log(n)$ phases, the result follows. Note that the $O(\log^2 n/n)$ rate is not meant to be optimal. It is not hard to see that a similar argument could establish an $O(\log n \log\log n/n)$ rate almost surely by replacing the h_k with an optimal predictor that has errors $O(\log k/2^k)$ with probability $1 - O(1/k^2)$, using e.g., the 1-inclusion algorithm of Haussler et al. (1994). We leave it as an open problem to obtain the optimal almost sure rate with finite VC-dimension, i.e., whether an additional $\log\log n$ term is necessary for the almost sure rate.

Bousquet et al. (2020a) recently considered a similar online learning setup but with a weaker realizable assumption. Instead of requiring the function chosen by nature comes from $\mathcal{H}$, they allowed nature to choose any function f so long as

$$\inf_{h \in \mathcal{H}} \Pr_{x \sim \mu}[h(x) \neq f(x)] = 0,$$

where μ is the underlying distribution which is *unknown* to the learner. It is shown by Bousquet et al. (2020a) that, with the weaker realizable assumption, we have the expected rate (i.e., $\mathbb{E}[I_n]$) can either be asymptotically e^{-n} or $\frac{1}{n}$ or arbitrarily slow.

Note that, an exponential rate of a class in the setting of Bousquet et al. (2020a) will imply that the class is strongly *e.a.s.*-online learnable as we defined in Section 6.2. This is captured by the existence of *infinite Littlestone trees* (Bousquet et al. 2020a), i.e.,

infinite full binary mistake trees (see Section 6.4) of the class. Informally, Bousquet et al. (2020a) showed that a class $\mathcal{H}$ can be learned with finitely many errors with the weaker realizable assumption if and only if $\mathcal{H}$ has *no* infinite Littlestone tree. However, such a characterization does not hold for the strong *e.a.s.*-online learnability that we introduced in Section 6.2, where we have assumed that the function choosing by nature must come from the class. This follows from the following example, which is inspired by the Example 2.7 of (Bousquet et al. 2020a).

Example 12. *In this example, we show that there exists a class $\mathcal{H}$ of binary functions such that $\mathcal{H}$ is strongly e.a.s.-online learnable, but $\mathcal{H}$ can* not *be decomposed into countable union of subclasses such that each of the subclass has no infinite Littlestone tree.*

Let $\mathcal{X} = \{S : S \subset \mathbb{R}$ and S is countable $\}$ be the domain, where each element of $\mathcal{X}$ is a countable *subset of the reals $\mathbb{R}$. For any $x \in \mathbb{R}$, we define the following* binary *function $h_x : \mathcal{X} \to \{0, 1\}$ such that*

$$\forall S \in \mathcal{X} : h_x(S) = 1 \text{ iff } x \in S.$$

Denote $\mathcal{H} = \{h_x : x \in \mathbb{R}\}$. We now show that $\mathcal{H}$ is the desired class.

We first show that the class $\mathcal{H}$ is strongly e.a.s.-online learnable. We define the prediction rule as follows: if we haven't observed label 1 in the current samples, then predict 0 for the current instance. Otherwise, we will be able to observe $S \in \mathcal{X}$ such that $h_x(S) = 1$, where h_x is the underlying function choosing by nature at the beginning. Now, since S is countable, we know that the class of functions in $\mathcal{H}$ that is consistent with current samples must be countable, since $h_x(S) = 1$ implies $x \in S$. The strong e.a.s.-online learnability now follows from Lemma 15.

Suppose $\mathcal{H} = \bigcup_{n \in \mathbb{N}} \mathcal{H}_n$, we show that there must be some $n \in \mathbb{N}$ such that $\mathcal{H}_n$ has an infinite Littlestone tree. This follows from the fact that $\mathcal{H}$ is uncountable, therefore there must be some $\mathcal{H}_n$ that has infinite elements, it is easy to show that such a class must have an infinite Littlestone tree.

It is therefore an interesting problem to investigate whether a similar phenomenon as in (Bousquet et al. 2020a) will happen for the stronger realizable setting and in the almost sure scenario that we have introduced in this chapter.

6.8 Summary

The author is the primary contributor for the work in this chapter, which has been partially published in (Wu and Santhanam 2021a)

Chapter 7

The insurance problem

7.1 Introduction

In this chapter we will study a problem that arose in the risk management theory. Suppose we have an unknown distribution over $\mathbb{N}$, which models the loss of some insurance tasks. An insurer can access the historical losses that are sampled from the distribution. The goal of the insurer is to set up a *premium* based on the historical losses, in the hope that the next loss that is sampled from the same distribution does not exceed his premium. We now consider a scenario where the insurer do the insurance game (i.e., set up premiums) sequentially at each discrete time steps. We say the insurer *bankrupt* if the actual loss is exceeding his premium at some time step.

This problem was first introduced by Santhanam and Anantharam (2015), where the authors considered the scenario where the insurer is allowed to observe the losses that are sampled from the distribution for as long as he wants. But the insurer must decide to enter the insurance game after observing finitely many samples. The goal is to find an *entering* strategy so that the probability of bankrupting after entering the game is upper bounded by some given probability $\eta > 0$. A class $\mathcal{P}$ of distributions over $\mathbb{N}$ is said to be *insurable* if for any $\eta > 0$ there exists a universal *entering* rule such that the probability of bankrupting is upper bounded by η, no matter what the underlying distribution may be in $\mathcal{P}$.

In this chapter, we relax the requirements as introduced in (Santhanam and Anantharam 2015) to allow the insurer being bankrupted for finitely many times.

7.2 Problem setup

Throughout this chapter, we use $\mathcal{P}$ to denote both a class of random variables over $\mathbb{N}$ and the corresponding *i.i.d.* random processes, when there is no ambiguity. For any $p \in \mathcal{P}$, we denote X_p to be the random variable governed by probability law p. We use boldface letters $\mathbf{a}, \mathbf{b}$ to denote infinite real sequences, and a_n, b_n to be the n-th element of $\mathbf{a}, \mathbf{b}$. We write $\mathbf{a} \geq \mathbf{b}$ if $a_n \geq b_n$ for all $n \in \mathbb{N}$. We say $\mathbf{a}$ *dominates* $\mathbf{b}$ or $\mathbf{a} \succ \mathbf{b}$ if there exists N such that

$$\forall n \geq N \quad a_n \geq b_n.$$

Note that we do *not* use dominate to mean $a_n \geq b_n$ for all n ($\mathbf{a} \geq \mathbf{b}$).

A *tail-mass sequence* is a positive valued real sequence $\boldsymbol{\epsilon}$ with $\epsilon_1 = 1$ and $\epsilon_n \to 0$ as $n \to \infty$. For any distribution $p \in \mathcal{P}$, given a tail-mass sequence $\boldsymbol{\epsilon}$, the $\boldsymbol{\epsilon}$-percentiles of p is the sequence of numbers $\hat{\mathbf{p}}$, where

$$\hat{p}_n = \min\{i : p(X_p \geq i) \leq \epsilon_n\}. \tag{7.1}$$

While clearly $\hat{\mathbf{p}}$ depends on $\boldsymbol{\epsilon}$ and p, we drop the $\boldsymbol{\epsilon}$ and p from notation for simplicity where there is no ambiguity. A class $\mathcal{P}$ is said to be *tight* if there exists a real sequence $\mathbf{s}$ and a tail-mass sequence $\boldsymbol{\epsilon}$ as defined above, such that for all $p \in \mathcal{P}$, $\mathbf{s} \geq \hat{\mathbf{p}}$. This definition is, of course, consistent with the standard notion of tightness in probability theory.

For any class $\mathcal{P}$, the insurance task is to find a universal prediction rule $\Phi : \mathbb{N}^* \to \mathbb{N}$, such that

$$\forall p \in \mathcal{P},\ p\left(\sum_{n=1}^{\infty} 1\{\Phi(X_1^{n-1}) < X_n\} < \infty\right) = 1,$$

where $X_1, X_2, \cdots$ are *i.i.d.* samples of p. Meaning that, at each time step n, the insurer will need to set up a premium $Y_n = \Phi(X_1^{n-1})$ based on the past samples X_1^{n-1}, in the hope

that the current loss X_n does not exceed his premium Y_n. Clearly, this setup matches with our *e.a.s.*-prediction setup in Chapter 4.1 by defining the loss function to be

$$\ell(p, X_1^n, \Phi(X_1^{n-1})) = 1\{X_n > \Phi(X_1^{n-1})\}.$$

We will say, in the sequel, that a class $\mathcal{P}$ is *e.a.s.*-predictable for the insurance task, if $(\mathcal{P}, \ell)$ is *e.a.s.*-predictable as defined in Chapter 4.

7.3 Main result

Our main result of this chapter is a full characterization of the e.a.s.-predictability of the insurance task, as follows:

Theorem 35. *A model class $\mathcal{P}$ of the insurance task is e.a.s.-predictable, iff there exist countably many tight classes $\{\mathcal{P}_i, i \in \mathbb{N}\}$ such that*

$$\mathcal{P} = \bigcup_{i \in \mathbb{N}} \mathcal{P}_i.$$

We leave the full proof of this theorem to the next section. The following lemmas are easy to prove:

Lemma 17. *If a model class $\mathcal{P}$ is tight, then it is e.a.s.-predictable for the insurance task.*

Lemma 18. *A class $\mathcal{P}$ can be written as a countable union*

$$\mathcal{P} = \bigcup_{i \in \mathbb{N}} \mathcal{P}_i$$

where each $\mathcal{P}_i$ is tight iff there exists a real sequence $\boldsymbol{s}$ and a tail mass sequence $\boldsymbol{\epsilon}$ such that for all $p \in \mathcal{P}$,

$$\boldsymbol{s} \succ \hat{\boldsymbol{p}},$$

where $\hat{\boldsymbol{p}}$ is the $\boldsymbol{\epsilon}$-percentile sequence of p as in (7.1).

Note that, using a similar argument as in Lemma 15, Lemma 17 implies the sufficiency part of Theorem 35. However, the proof of the necessary part of Theorem 35 is more involved. Lemma 18 provides us a different way of looking at the condition of countable union of tight classes, which will be crucial in our proof. Note that Theorem 35 and Lemma 17 gives immediately the following non-trivial example:

Example 13. *A distribution p is said to be monotone, if $p(X_p = n) \geq p(X_p = n+1)$ for all $n \in \mathbb{N}$. We show that the class of all monotone distributions, $\mathcal{M}$, is not e.a.s.-predictable for the insurance task. By Lemma 18, We only need to show that for $\boldsymbol{\epsilon}$ with $\epsilon_n = \frac{1}{n}$, and for any sequence $\boldsymbol{s}$, there is $p \in \mathcal{M}$ such that $p(X_p > s_n) \geq \epsilon_n$ for infinitely many n. W.l.o.g., we assume*

$$\frac{1}{(s_n - s_{n-1})n(n-1)} \geq \frac{1}{(s_{n+1} - s_n)(n)(n+1)}.$$

This is because given any $\boldsymbol{s}$, we can construct another sequence $\boldsymbol{s'}$ with $\boldsymbol{s'} \geq \boldsymbol{s}$ and proceed from the next step with $\boldsymbol{s'}$. Assign

$$p(X_p = k) = \frac{1}{(s_n - s_{n-1})n(n-1)}$$

for $s_{n-1} \leq k < s_n$. One can easily verify that this distribution is monotone and $p(X_p \geq s_n) \geq \frac{1}{n} = \epsilon_n$ for all n.

7.4 Proofs

In this section, we prove the lemmas and theorems stated in the last section. Intuitively, Lemma 18 allows us to think more directly in terms of prediction strategies in order to prove Theorem 35. By virtue of Lemma 18, it is sufficient to show that $\mathcal{P}$ is *e.a.s.*-predictable for the insurance task if and only if there exists a real sequence $\mathbf{s}$ and a tail mass sequence $\boldsymbol{\epsilon}$ such that for all $p \in \mathcal{P}$, $\mathbf{s} \succ \hat{\mathbf{p}}$, where $\hat{\mathbf{p}}$ is the $\boldsymbol{\epsilon}$-percentile sequence of p as in (7.1). For ease of exposition, we will refer the above condition as $\mathbf{s}$ *universally dominates* the $\boldsymbol{\epsilon}$ percentiles of $\mathcal{P}$.

Indeed, if $\mathbf{s}$ universally dominates the $\boldsymbol{\epsilon}$-percentiles of $\mathcal{P}$, we will see that we can derive a strategy directly from $\mathbf{s}$ that only makes finitely many errors almost surely, proving $\mathcal{P}$ is *e.a.s.*-predictable.

The other direction—if $\mathcal{P}$ is *e.a.s.*-predictable, then there must be a sequence $\mathbf{s}$ that universally dominates the $\boldsymbol{\epsilon}$-percentiles of $\mathcal{P}$—is more involved. Here we prove the contrapositive, that if no sequence $\mathbf{s}$ can dominate any $\boldsymbol{\epsilon}$-percentiles of $\mathcal{P}$, then $\mathcal{P}$ is not *e.a.s.*-predictable. To do so, we use a method reminiscent of Cantor's diagonalization argument (Takeuti and Zaring 1982).

In order to proceed, we need the following two simple technical lemmas. Define a *natural* strategy $\Phi := \mathbb{N}^* \to \mathbb{N}$ as one that only depends on $m = \max\{x_1, \cdots, x_n\}$ and n for all $(x_1, \cdots, x_n) \in \mathbb{N}^*$. The following lemma shows that it is sufficient to only consider natural strategies as above to characterize *e.a.s.*-predictability.

Lemma 19. *$\mathcal{P}$ is e.a.s.-predictable iff there is a natural strategy that makes finitely many errors with probability* 1 *for all $p \in \mathcal{P}$.*

Proof. Let Φ be any strategy. Define

$$\Phi'(x_1, \cdots, x_n) = \max\{\Phi(y_1, \cdots, y_n) \mid y_1^n \in [m]^n\},$$

where $[m]^n = \{1, 2, \cdots, m\}^n$. Clearly, Φ' will not increase the number of errors in the game, and Φ' is natural. □

Lemma 20. *For any class $\mathcal{P}$, suppose there exists a real sequence $\boldsymbol{s}$ that universally dominates the $\boldsymbol{\epsilon}$-percentiles of $\mathcal{P}$. Then for any other tail mass sequence $\boldsymbol{\epsilon}'$, there exists a sequence $\boldsymbol{s}'$ that universally dominates the $\boldsymbol{\epsilon}'$-percentiles of $\mathcal{P}$.*

Proof. Let $n' = \min\{i \mid \epsilon_i \leq \epsilon'_n\}$ and define $s'_n = s_{n'}$. We claim that $\mathbf{s}'$ satisfies the requirement.

To see this, for all $p \in \mathcal{P}$ since $\mathbf{s} \succ \hat{\mathbf{p}}$, we can conclude there exist N such that

$$\forall n \geq N,\ p(X_p \geq s_n) \leq \epsilon_n.$$

Now by definition of $\mathbf{s}'$, we also have

$$\forall n' \geq N,\ p(X_p \geq s'_n) = p(X_p \geq s_{n'}) \leq \epsilon_{n'} \leq \epsilon'_n.$$

Therefore, for all $n \geq N'$, $s'_n \geq \hat{p}'_n$, where $N' = \min\{n : n' \geq N\}$ and $\hat{\mathbf{p}}'$ is the $\boldsymbol{\epsilon}'$-percentile of p. □

Proof of Lemma 18. Suppose there exists a real sequence $\mathbf{s}$ and a tail mass sequence $\boldsymbol{\epsilon}$ such that for all $p \in \mathcal{P}$, $\mathbf{s} \succ \hat{\mathbf{p}}$. Equivalently, for all $p \in \mathcal{P}$ there exists N_p such that $\forall n \geq N_p$, we have $s_n \geq \hat{p}_n$. Put another way, for all $n \geq N_p$

$$p(X_p \geq s_n) \leq \epsilon_n.$$

Now define

$$\mathcal{P}_n = \{p \in \mathcal{P} \mid N_p = n\}.$$

It is easily seen that $\mathcal{P}_n$ are tight for all $n \in \mathbb{N}$. Furthermore, we have

$$\bigcup_{n \in \mathbb{N}} \mathcal{P}_n = \mathcal{P},$$

therefore $\mathcal{P}$ can be written as a countable union of tight classes.

To prove the converse, suppose $\mathcal{P} = \bigcup_{n \in \mathbb{N}} \mathcal{P}_n$ where each $\mathcal{P}_n$ is tight. Fix any tail mass sequence $\boldsymbol{\epsilon}$. For all n, since $\mathcal{P}_n$ is tight, there exists a real sequence $\mathbf{s}^n$ such that for all $p \in \mathcal{P}_n$

$$\mathbf{s}^n \geq \hat{\mathbf{p}},$$

where $\hat{\mathbf{p}}$ are the percentiles of the tail mass sequence $\boldsymbol{\epsilon}$.

Consider the sequence $\mathbf{s}$

$$s_n = \max\{s_i^j \mid i, j \leq n\}.$$

Clearly, $\mathbf{s} \succ \hat{\mathbf{p}}$ for all $p \in \mathcal{P}$. The lemma follows. □

Proof of theorem 35. Suppose there exists a sequence $\mathbf{s}$ that universally dominates the $\boldsymbol{\epsilon}$-percentiles of $\mathcal{P}$. We first construct a scheme Φ that only makes finitely many errors by using $\mathbf{s}$.

Lemma 20 then implies that for all tail mass sequences $\boldsymbol{\epsilon}'$ there exists a sequence $\mathbf{s}'$ that universally dominates the $\boldsymbol{\epsilon}'$-percentiles of $\mathcal{P}$. w.l.o.g., we fix the tail mass sequence to be $\boldsymbol{\epsilon}$ with $\epsilon_n = \frac{1}{2^n}$ and let $\mathbf{s}$ universally dominate ϵ-percentiles of $\mathcal{P}$.

Lemma 19 allows us to consider only natural strategies. We build a strategy $\Phi : \mathbb{N}^* \to \mathbb{R}$ as follows. For any sequence $X_1 \cdots X_n$ where the sequence maximum is m, Φ assigns a value $\Phi(m, n)$ given by

$$\Phi(m, n) = s_{\max\{m,n\}}.$$

Denote by A_n the event that the strategy errs at round n. We have

$$p(A_n) = p(X_p \geq s_{\max\{m,n\}}) \leq p(X_p \geq s_n) \leq \frac{1}{2^{n-1}}.$$

Therefore, $\sum_{n=1}^{n} p(A_n) < \infty$. From the Borel–Cantelli lemma, we can conclude that Φ only makes finitely many errors, or that $\mathcal{P}$ is $e.a.s.$-predictable.

We will now show that if no sequence $\mathbf{s}$ can universally dominate any $\boldsymbol{\epsilon}$-percentiles of $\mathcal{P}$, then any strategy $\Phi : \mathbb{N}^* \to \mathbb{N}$ fails infinitely often. By virtue of Lemmas 19 and 20, it is sufficient to consider natural strategies and to fix a particular tail mass sequence $\boldsymbol{\epsilon}$.

We now take $\epsilon_n = \frac{1}{n}$. No sequence $\mathbf{s}$ universally dominates the $\boldsymbol{\epsilon}$-percentiles of $\mathcal{P}$. Fix any natural strategy $\Phi(m, n)$ that only depends on the maximum m and the size n of a sequence from $\mathbb{N}^*$. We also assume w.l.o.g. that $\Phi(m, n)$ is strictly increasing on both m, n. We now recursively define

$$t_n = \Phi(t_{n-1}, t_{n-1}),$$

with $t_1 = \Phi(0, 0) = 1$. One can prove by induction that $t_n \geq \Phi(n, n) \geq n$. Now define a sequence $\mathbf{s}$ with

$$s_n = t_{2n}.$$

Since $\mathbf{s}$ cannot universally dominate the $\boldsymbol{\epsilon}$-percentiles of $\mathcal{P}$, there exists some $p \in \mathcal{P}$ such that for infinitely many m we have

$$p(X_p > s_m) > \epsilon_m = \frac{1}{m}.$$

Define $A_{m,n}$ to be the event that the strategy Φ makes no errors after time step n, and where we do not observe a number greater than m in the first n time steps. We have $A_{m,n} \subset A_{m+1,n}$ and $A_{m,n} \subset A_{m,n+1}$. Define

$$A_n = \lim_{m \to \infty} A_{m,n}.$$

Clearly,

$$A = \bigcup_n A_n = \lim_{n \to \infty} A_n$$

is the event that Φ errs finitely many times.

We now show that $p(A_n) \leq \frac{1}{\sqrt{e}}$ for all $n \in \mathbb{N}$. To do so, we first note that there are infinitely many $m > n$ with $p(X_p > s_m) > \epsilon_m = \frac{1}{m}$. Fix any such $m > 2n$, and we will show that

$$p(A_{m,n}) \leq \frac{1}{\sqrt{e}}.$$

Let $y_{n+1} = \Phi(m, n)$ be the $n+1$th prediction, we know that the probability the *insurer* not fail at round $n+1$ is at most

$$\begin{aligned} &p(X_p \leq y_{n+1} = \Phi(m,n)) \leq p(X_p \leq \Phi(m,m)) \\ &\overset{(a)}{\leq} p(X_p \leq t_m) \overset{(b)}{\leq} p(X_p \leq s_m) \leq 1 - \frac{1}{m}, \end{aligned}$$

where (a) holds since $t_m \geq \Phi(m,m)$, (b) follows by $s_m = t_{2m} \geq t_m$.

Now, if the Φ does not err in round $n+1$, we must have $X_{n+1} \leq y_{n+1} \leq \Phi(m,m) \leq t_m$. Thus the maximum number observed in $n+1$ steps now is t_m, and the probability that Φ

does not err in round $n+2$ is at most

$$p(X_p \leq \Phi(t_m, n+1)) \overset{(a)}{\leq} p(X_p \leq \Phi(t_m, t_m))$$
$$\overset{(b)}{\leq} p(X_p \leq t_{m+1}) \overset{(c)}{\leq} p(X_p \leq s_m) \leq 1 - \frac{1}{m}$$

where for (a) we assume $m \geq n+1$, hence $t_m \geq n+1$, (b) is by definition, and (c) follows since $s_m = t_{2m} \geq t_{m+1}$. One can repeat this sequence of arguments for $m-n$ further steps, continuing to satisfy the constraint on m in both steps (a) and (c). Doing so we have

$$p(A_{m,n}) \leq \left(1 - \frac{1}{m}\right)^{m-n} \leq \left(1 - \frac{1}{m}\right)^{m/2} \leq \frac{1}{\sqrt{e}}.$$

Since there are infinitely many such m, we obtain

$$p(A_n) = \lim_{m \to \infty} p(A_{m,n}) \leq \frac{1}{\sqrt{e}}.$$

Therefore, we have

$$p(A) = \lim_{n \to \infty} p(A_n) \leq \frac{1}{\sqrt{e}} < 1.$$

Thus Φ errs infinitely many times with probability at least $1 - \frac{1}{\sqrt{e}}$. Theorem 35 now follows. □

Remark 18. *We should remark that the probability $\frac{1}{\sqrt{e}}$ is not optimal in the proof above and can easily be improved to* 0*. We choose the above argument for ease of presentation.*

Remark 19. *Note that, Theorem 35 can also be proved using our general characterization in Theorem 1. By showing that if a class $\mathcal{P}$ is η-predictable with $\eta < \frac{1}{4}$ then $\mathcal{P}$ can be decomposed as a countable union of tight and relatively open (under variation distance) subclasses. However, the proof of this claim is also non-trivial and uses similar recursive construction as we presented in the proof above. Together with Theorem 10, this also implies that a class is* insurable *iff it can be decomposed as countable union of tight and relatively open subclasses, recovering the main result in (Santhanam and Anantharam 2015).*

7.5 Summary

The author is the primary contributor for the work in this chapter, which has been published in (Wu and Santhanam 2019).

Chapter 8

Discussion

In this dissertation, we introduced a general prediction paradigm where a predictor is required to make finitely many errors with probability 1 in the infinite horizon. We provided general characterizations of the existence of such prediction rules. In particular, our approach for deriving the characterizations provides equivalence between the finite error guarantee and the decomposition of the class into nested unions of *uniformly* predictable subclasses. In this nested decomposition, we match the available sample with a specific level of subclass in the nesting.

We note that this is an abstract and general way to think of appropriateness of regularization. As we relax the regularization when we see more samples, the question is whether the increased flexibility in choosing candidate models leads to perpetually modifying our estimates substantially. Or, in other cases, relaxing regularization leads to modification of the estimated model, but only to a point—after a finite point, no amount of additional flexibility changes the model in a substantial way. In this sense, the regularization stabilizes, and is a guarantee that the regularization *finds* a specific model. The later case, where regularization stabilizes is intimately connected with the finitely many error paradigm, and the dissertation establishes the nuances of these connections.

We also studied several concrete problem setups of this paradigm, specifically in the hypothesis testing, online learning and risk prediction scenarios. In these settings, tight characterizations were obtained.

While our main focus of this dissertation is on the theory and conceptual foundations of learning, we intend this to be a base to understand large scale problems that we have little prior knowledge in. As mentioned repeatedly in the dissertation in different settings, our framework also provides a different angle for regularization, asking when regularization stabilizes in an online setting, in addition to asking when and how regularization can provide online guarantees.

Informally, regularization is a process that adds constraints in order to resolve learning problems that do not admit uniform consistency. However, the selection of regularization is often quite empirical, and it is usually not clear what information we should add in order to resolve a problem we have on hand. The *e.a.s.*-prediction paradigm provides us a criterion on the selection of regularization. More precisely, for any model class $\mathcal{P}$, a regularization should be able to decompose the class $\mathcal{P}$ into nested subclasses $\{\mathcal{P}_i, i \geq 1\}$ such that each of the $\mathcal{P}_i$ admits uniform consistency and the union $\bigcup \mathcal{P}_i$ recovers the whole class $\mathcal{P}$. Suppose the sample size we have on hand is n, we should select a model class $\mathcal{P}_k$ with *maximum* k such that the sample complexity of $\mathcal{P}_k$ that achieves confidence $1 - \eta_k$ is less than n, where η_k is an arbitrary sequence such that $\sum_k \eta_k < \infty$. Now, the implication of such a selection rule is that we will be able to settle the learning problem with finite sample size almost surely, no matter what underlying model is from $\mathcal{P}$. For many natural settings, if no such decomposition exists, then any regularization will be meaningless. In the sense that for any learning rule there must be some model in the class $\mathcal{P}$ such that the rule fails infinitely often no matter how large sample size we have.

8.1 Future directions

We recall several specific open problems throughout this dissertation, including: Problem 1 in Chapter 4.5, Problem 2 in Chapter 5.3, Conjecture 1 in Chapter 5.2, Conjecture 2 in Chapter 6.4 and Conjecture 3 in Chapter 6.5. Besides these, there are broader angles that would be of interest.

The control setting and interactive sampling process: All the setups in this dissertation consider a sampling process that is independent of the predictions made by the learner. It would be interesting to investigate the situation when the learner could interactively control the sampling process. A simple example is as follows. Considering the online learning of the linear threshold functions over $[0, 1]$, we know that such a task is not $e.a.s.$-predictable if the samples are *i.i.d.*. However, if we allow the learner to *choose* some sample to label at each time step instead of revealing the label of the testing sample, it is easy to see that we will be able to make finitely many errors almost surely. More generally, while the current disseration does not handle the control setting specifically, it will be useful to compare and adapt our setting to control settings.

Studying optimality of prediction rules: In this dissertation, we investigated the $e.a.s.$-predictability from the consistency point of view, i.e., whether there exists a prediction rule or not. However, it would also be interesting to study the optimality of the prediction rules. A possible way of formulating the optimality is through competitive analysis. The idea is as follows. Suppose we have some *canonical* way of regularizing the class into $\mathcal{P} = \bigcup_{i\in\mathbb{N}} \mathcal{P}_i$, such that each $\mathcal{P}_i$ is uniformly predictable. We could study how the universal $e.a.s.$-prediction rule performs on $\mathcal{P}_i$ competitively to the optimal prediction rule that is restricted only on $\mathcal{P}_i$.

Finding optimal stopping rules: The construction of stopping rules for the $e.a.s.$-learning framework as we outlined in Section 4.4 is quite abstract. However, for many problems one often will be able to find more natural stopping rules. It would be interesting to study the optimal stopping rules for specific real applications.

Appendix A

Omitted Proofs

A.1 Omitted proofs in Example 4

Lemma 21. *Let $\mathcal{B}_k$ be the set of all Bernoulli processes with parameters in*

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

where $B(r_i, \frac{1}{k2^i})$ is the open balls centered at r_i with radius $\frac{1}{k2^i}$, $r_1, r_2, \cdots$ is an arbitrary enumeration of rational numbers in $[0,1]$. Then $\mathcal{B}_k$ is η-predictable with irrationality loss for any $k \geq 1$ and $\eta > 0$.

Proof. We show that for any $k \in \mathbb{N}$ and $\eta > 0$, there exists b_η such that $\mathcal{B}_k$ is η-predictable with sample size b_η. Let $X_1, \cdots, X_n$ be an *i.i.d.*sample from some $p \in \mathcal{B}_k$ with $\mathbb{E}[X_i] = \mu$ and $\bar{X} = \frac{X_1 + \cdots + X_n}{n}$. We have $\mathrm{Var}[X_i] \leq 1$. Chebyshev's inequality then shows that

$$p\left(|\bar{X} - \mu| \geq \epsilon\right) \leq \frac{1}{n\epsilon^2}.$$

Fix $\epsilon = \frac{1}{k2^{k+1}}$. Let b_η be a number large enough so that $\frac{1}{b_\eta \epsilon^2} < \eta$. Therefore for $n > b_\eta$, $p\left(|\bar{X} - \mu| \geq \epsilon\right)$ is less than η. Thus, we can conclude that $\mathcal{B}_k$ is η-predictable by simply predicting the irrationality of element in $\mathcal{S}_k$ that is closest to $\bar{X}$ at step b_η, retaining the prediction perpetually thereafter. □

Lemma 22. *Let $\mathcal{B}$ be a class of Bernoulli processes with parameters in $\mathcal{S}$, if $\mathcal{B}$ is η-predictable w.r.t. the irrationality loss for some $0 < \eta < \frac{1}{2}$, then*

$$\inf\{|x - r| : x, r \in \mathcal{S} \text{ and } r \in \mathbb{Q},\ x \in [0,1]\backslash\mathbb{Q}\} > 0. \tag{A.1}$$

Proof. By definition of η-predictability, there exists a number N_η and prediction rule Φ_η such that Φ_η makes no errors after step N_η with probability at least $1 - \eta$ for all $p \in \mathcal{B}$. Suppose, otherwise, that the infimum in equation (A.1) is 0. We now select two sources p_0, p_1 from $\mathcal{B}$ with parameters b_0, b_1 respectively, such that b_0 is rational and b_1 is irrational and $|b_0 - b_1| < \frac{1-2\eta}{2N_\eta}$.

We now have $||p_0^{N_\eta} - p_1^{N_\eta}||_{TV} < 1 - 2\eta$, where $p_i^{N_\eta}$ is the distribution of p_i restricted to the first N_η samples—using the fact that $||p^N - q^N||_{TV} \leq N||p - q||_{TV}$ for any distributions p, q with N-fold *i.i.d.* distributions p^N, q^N. Now, by Lemma 4, any prediction rule (in particular Φ_η) will make an error at time step $N_\eta + 1$ with probability $> \eta$ for either p_0 or p_1. This contradicts η predictability of Φ_η on $\mathcal{B}$. □

Lemma 23. *Let $\mathcal{S}_1 \subset \mathcal{S}_2 \subset \cdots \subset \mathcal{S}_k \subset \cdots \subset [0,1]$ be countably many sets, such that*

$$\forall k,\ \inf\{|x - r| : x, r \in \mathcal{S}_k \text{ and } r \in \mathbb{Q},\ x \in [0,1]\backslash\mathbb{Q}\} > 0. \tag{A.2}$$

If $\bigcup_{k\in\mathbb{N}} \mathcal{S}_k$ contains all rational numbers in $[0,1]$, then the irrational numbers in $\mathcal{S}_k$ are nowhere dense in $[0,1]$ for all k.

Proof. Suppose otherwise, the set of irrational numbers $\mathcal{I}_k$ in $\mathcal{S}_k$ is not nowhere dense. By definition, there exists an interval $[a, b] \subset \text{col}(\mathcal{I}_k)$, where col denotes for closure. Since the rational numbers in $[0, 1]$ are dense, there exists some rational number $r \in [a, b]$, and therefore $r \in \text{col}(\mathcal{I}_k)$. Since $r \in \bigcup_{k\in\mathbb{N}^+} \mathcal{S}_k$, there exist some $k' \geq k$ such that $r \in \mathcal{S}_{k'}$. However, we also have $\mathcal{S}_k \subset \mathcal{S}_{k'}$. Which implies that r is the limit point of irrational numbers in $\mathcal{S}_{k'}$, contradicting the assumption (A.2) . □

References

Sushant Agarwal, Nivasini Ananthakrishnan, Shai Ben-David, Tosca Lechner, and Ruth Urner. On learnability wih computable learners. In *Algorithmic Learning Theory (ALT)*, pages 48–60. PMLR, 2020.

András Antos and Gábor Lugosi. Strong minimax lower bounds for learning. *Machine learning*, 30(1):31–56, 1998.

Sanjeev Arora and Boaz Barak. *Computational complexity: a modern approach*. Cambridge University Press, 2009.

Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of Computing*, 8(1):121–164, 2012.

Krishna B Athreya and Soumendra N Lahiri. *Measure theory and probability theory*. Springer Science & Business Media, 2006.

Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516, 2008.

Andrew Barron, Jorma Rissanen, and Bin Yu. The minimum description length principle in coding and modeling. *IEEE transactions on information theory*, 44(6):2743–2760, 1998.

Andrew R Barron. Complexity regularization with application to artificial neural networks. In *Nonparametric functional estimation and related topics*, pages 561–576. Springer, 1991.

Andrew R Barron and Thomas M Cover. Minimum complexity density estimation. *IEEE transactions on information theory*, 37(4):1034–1054, 1991.

Jānis Martynovich Barzdiņš and RV Freivald. On the prediction of general recursive functions. *Doklady Akademii Nauk*, 206(3):521–524, 1972.

Shai Ben-David, Dávid Pál, and Shai Shalev-Shwartz. Agnostic online learning. In *Conference on Learning Theory (COLT)*, volume 3, page 1, 2009.

Shai Ben-David, Pavel Hrubeš, Shay Moran, Amir Shpilka, and Amir Yehudayoff. Learnability can be undecidable. *Nature Machine Intelligence*, 1(1):44–48, 2019.

Shai Ben-David, Pavel Hrubes, Shay Moran, Amir Shpilka, and Amir Yehudayoff. Learnability can be independent of set theory. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pages 11–11, 2021.

Gyora M Benedek and Alon Itai. Nonuniform learnability. *Journal of Computer and System Sciences*, 48(2):311–323, 1994.

Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, 2019.

Lenore Blum and Manuel Blum. Toward a mathematical theory of inductive inference. *Information and control*, 28(2):125–155, 1975.

Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Learnability and the Vapnik–Chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

Olivier Bousquet, Steve Hanneke, Shay Moran, Ramon van Handel, and Amir Yehudayoff. A theory of universal learning. *arXiv:2011.04483*, 2020a.

Olivier Bousquet, Steve Hanneke, Shay Moran, and Nikita Zhivotovskiy. Proper learning, Helly number, and an optimal SVM bound. In *Conference on Learning Theory (COLT)*, pages 582–609. PMLR, 2020b.

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

Nicolo Cesa-Bianchi and Gábor Lugosi. On prediction of individual sequences. *The Annals of Statistics*, 27(6):1865–1895, 1999.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

Siu On Chan, Qinghua Ding, and Sing Hei Li. Learning and testing irreducible Markov chains via the k-cover time. In *Algorithmic Learning Theory (ALT)*, pages 458–480. PMLR, 2021.

Doron Cohen, Aryeh Kontorovich, and Geoffrey Wolfer. Learning discrete distributions with infinite support. *arXiv preprint arXiv:2004.12680*, 2020.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3): 273–297, 1995.

Robert D Cousins. The Jeffreys–Lindley paradox and discovery criteria in high energy physics. *Synthese*, 194(2):395–432, 2017.

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

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley and sons., 1991.

Lee D Davisson. Universal noiseless coding. *IEEE Transactions on Information Theory*, 19 (6):783–795, 1973.

Lee D Davisson. Minimax noiseless universal coding for markov sources. *IEEE Transactions on Information Theory*, 29(2):211–215, 1983.

Lee D Davisson and Alberto Leon-Garcia. A source matching approach to finding minimax codes. *IEEE Transactions on Information Theory*, 26(2):166–174, 1980.

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

Frank Den Hollander. Probability theory: The coupling method. *Lecture notes available online (http://websites. math. leidenuniv. nl/probability/lecturenotes/CouplingLectures. pdf)*, 2012.

Paul Erdős and Alfréd Rényi. On cantor's series with convergent. *qn Ann. Univ. Eötvös de Budapest Sect. Math*, 2:93–109, 1959.

Meir Feder and Neri Merhav. Hierarchical universal coding. *IEEE transactions on information theory*, 42(5):1354–1364, 1996.

Boris M Fitingof. The compression of discrete information. *Problemy Peredachi Informatsii*, 3(3):28–36, 1967.

Robert G Gallager. Source coding with side information and universal coding. *Unplished manuscript*, 1979.

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

Geoffrey Grimmett and David Stirzaker. *Probability and random processes*. Oxford university press, 2020.

Peter Grunwald. A tutorial introduction to the minimum description length principle. *arXiv preprint math/0406077*, 2004.

Peter Grünwald and Teemu Roos. Minimum description length revisited. *International journal of mathematics for industry*, 11(01):1930001, 2019.

Peter D Grünwald. *The minimum description length principle*. MIT press, 2007.

Steve Hanneke and Aryeh Kontorovich. Stable sample compression schemes: New applications and an optimal SVM margin bound. In *Algorithmic Learning Theory (ALT)*, pages 697–721. PMLR, 2021.

David H. Haussler, Nick Littlestone, and Manfred K. Warmuth. Predicting {0, 1}-functions on randomly drawn points. *Information and Computation*, 115(2):248–292, 1994.

Ralf Herbrich and Robert Williamson. Algorithmic luckiness. *Journal of Machine Learning Research*, 3:175–212, 2002.

Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

Wassily Hoeffding. Lower bounds for the expected sample size and the average risk of a sequential procedure. *The Annals of Mathematical Statistics*, pages 352–368, 1960.

David Hunter. Asymptotic tools. *Lecture notes available online (http://personal.psu.edu/drh20/asymp/fall2006/lectures/updateFA06/chapter02.pdf)*, 2006.

Jack Kiefer and Lionel Weiss. Some properties of generalized sequential probability ratio tests. *The Annals of Mathematical Statistics*, pages 57–74, 1957.

Jack Kiefer and Jacob Wolfowitz. On the deviations of the empiric distribution function of vector chance variables. *Transactions of the American Mathematical Society*, 87(1): 173–186, 1958.

Jack Kieffer. A unified approach to weak universal source coding. *IEEE Transactions on Information Theory*, 24(6):674–682, 1978.

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

Lauwerens Kuipers and Harald Niederreiter. *Uniform distribution of sequences*. Courier Corporation, 2012.

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

Lucien Le Cam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1(1):38–53, 1973.

Amir Leshem. Cover's test of rationality revisited: Computability aspects of hypothesis testing. In *2006 IEEE 24th Convention of Electrical & Electronics Engineers in Israel*, pages 213–216. IEEE, 2006.

Dennis V Lindley. A statistical paradox. *Biometrika*, 44(1/2):187–192, 1957.

Nathan Linial, Yishay Mansour, and Ronald L Rivest. Results on learnability and the Vapnik-Chervonenkis dimension. *Information and Computation*, 90(1):33–49, 1991.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2(4):285–318, 1988.

Nick Littlestone and Manfred K Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.

Pascal Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, pages 1269–1283, 1990.

Rahul Mazumder and Trevor Hastie. Exact covariance thresholding into connected components for large-scale graphical lasso. *The Journal of Machine Learning Research*, 13(1):781–794, 2012.

Neri Merhav and Meir Feder. Universal prediction. *IEEE Transactions on Information Theory*, 44(6):2124–2147, 1998.

Sofia Mosci, Lorenzo Rosasco, Matteo Santoro, Alessandro Verri, and Silvia Villa. Solving structured sparsity regularization with proximal methods. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 418–433. Springer, 2010.

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

Michael Naaman. On the tight constant in the multivariate Dvoretzky–Kiefer–Wolfowitz inequality. *Statistics & Probability Letters*, 173:109088, 2021.

D Neuhoff, R Gray, and L Davisson. Fixed rate universal block source coding with a fidelity criterion. *IEEE Transactions on Information Theory*, 21(5):511–523, 1975.

Kalyanapuram Rangachari Parthasarathy. *Probability measures on metric spaces*, volume 352. American Mathematical Soc., 2005.

Jan Poland and Marcus Hutter. Asymptotics of discrete mdl for online prediction. *IEEE Transactions on Information Theory*, 51(11):3780–3795, 2005.

Hugo M Proença and Matthijs van Leeuwen. Interpretable multiclass classification by mdl-based rule lists. *Information Sciences*, 512:1372–1393, 2020.

J Ross Quinlan and Ronald L Rivest. Inferring decision trees using the minimum description lenght principle. *Information and computation*, 80(3):227–248, 1989.

Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

Jorma Rissanen. A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, pages 416–431, 1983.

Jorma Rissanen. Universal coding, information, prediction, and estimation. *IEEE Transactions on Information theory*, 30(4):629–636, 1984.

Jorma Rissanen. Stochastic complexity and fisher information'. *IEEE Trans. Inf. Theory*, 1994.

Jorma Rissanen. Strong optimality of the normalized ml models as universal codes and information in data. *IEEE Transactions on Information Theory*, 47(5):1712–1717, 2001.

Christian P Robert. On the Jeffreys-Lindley paradox. *Philosophy of Science*, 81(2):216–232, 2014.

Adam J Rothman, Peter J Bickel, Elizaveta Levina, and Ji Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

Walter Rudin. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1964.

Walter Rudin. Real and complex analysis. *McGraw Hill Inc.,*, 1974.

Walter Rudin. *Functional Analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill, 2006. ISBN 9780070619883.

Boris Yakovlevich Ryabko. Twice-universal coding. *Problems of information transmission*, 20(3):173–177, 1984.

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.

Narayana Santhanam, Venkat Anantharam, Aleksandar Kavcic, and Wojciech Szpankowski. Data driven consistency (working title). *arXiv preprint arXiv:1411.4407*, 2014.

Dale Schuurmans. Characterizing rational versus exponential learning curves. *Journal of computer and system sciences*, 55(1):140–160, 1997.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and trends in Machine Learning*, 4(2):107–194, 2011.

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

John Shawe-Taylor, Peter L Bartlett, Robert C Williamson, and Martin Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE transactions on Information Theory*, 44(5):1926–1940, 1998.

Yurii Mikhailovich Shtar'kov. Universal sequential coding of single messages. *Problemy Peredachi Informatsii*, 23(3):3–17, 1987.

Andrew C Singer and Meir Feder. Universal linear prediction by model order weighting. *IEEE Transactions on Signal Processing*, 47(10):2685–2699, 1999.

Ray J Solomonoff. A formal theory of inductive inference. Part I. *Information and control*, 7(1):1–22, 1964a.

Ray J Solomonoff. A formal theory of inductive inference. Part II. *Information and control*, 7(2):224–254, 1964b.

David Soloveichik. Statistical learning of arbitrary computable classifiers. *arXiv preprint arXiv:0806.3537*, 2008.

Steven Squires, Adam Prügel-Bennett, and Mahesan Niranjan. Rank selection in nonnegative matrix factorization using minimum description length. *Neural computation*, 29(8):2164–2176, 2017.

Sashi Mohan Srivastava. *A course on Borel sets*, volume 180. Springer Science & Business Media, 2008.

Gaisi Takeuti and Wilson M Zaring. *Introduction to Axiomatic Set Theory*, volume 1 of *Graduate Texts in Mathematics*. Springer-Verlag New York, 1982.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Andrey Nikolayevich Tikhonov. On the stability of inverse problems. *Doklady Akademii Nauk SSSR*, 39(5):195—-198, 1943.

Onno Van Gaans. Probability measures on metric spaces. *Lecture notes*, 2003.

Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

Vladimir N. Vapnik and Alexey Y. Chervonenkis. Theory of pattern recognition. 1974.

Bengt von Bahr and Carl-Gustav Esseen. Inequalities for the r th absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, 36(1):299–303, 1965.

Volodimir G Vovk. Aggregating strategies. *Proc. of Computational Learning Theory (COLT), 1990*, 1990.

Volodya Vovk. Competitive on-line linear regression. *Advances in Neural Information Processing Systems*, pages 364–370, 1998.

Volodya Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001.

Abraham Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.

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

Changlong Wu and Narayana Santhanam. Entropy property testing with finitely many errors. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2568–2573. IEEE, 2020.

Changlong Wu and Narayana Santhanam. Non-uniform consistency of online learning with random sampling. In *Algorithmic Learning Theory (ALT)*, pages 1265–1285. PMLR, 2021a.

Changlong Wu and Narayana Santhanam. Prediction with finitely many errors almost surely. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 3223–3231. PMLR, 2021b.

Richard S Zemel. *A minimum description length framework for unsupervised learning.* Citeseer, 1994.

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

Chicheng Zhang and Kamalika Chaudhuri. The extended Littlestone's dimension for learning with mistakes and abstentions. In *Conference on Learning Theory (COLT)*, pages 1584–1616, 2016.

Jacob Ziv. Coding of sources with unknown statistics–i: Probability of encoding error. *IEEE Transactions on Information Theory*, 18(3):384–389, 1972a.

Jacob Ziv. Coding of sources with unknown statistics–ii: Distortion relative to a fidelity criterion. *IEEE Transactions on Information Theory*, 18(3):389–394, 1972b.