Prediction and Learning with eventual almost sure guarantees

Narayana Santhanam (Univ of Hawaii, Manoa)

Joint work with
C Wu (Univ of Hawaii, Manoa)

Prediction and Learning with eventual almost sure guarantees

Narayana Santhanam (Univ of Hawaii, Manoa)

Joint work with
C Wu (Univ of Hawaii, Manoa)
M. Asadi, R. Paravi (UH), V. Anantharam and W.Szpankowski

From local to global information workshop
Feb. 5, 2020

University of Hawai'i Mānoa

Introduction

Theme: A different kind of statistical guarantee

Meta-question 1: Expanding uniform consistency to finitely many errors

Meta-question 2: If finitely many errors, stopping rule that anticipates the last error

Results

Characterization of

(i) model classes that admit predictors with finitely many errors, and
(ii) when there is a stopping rule that anticipates (with any given confidence) the point at which the last error is made

The first is a story of regularization (i.e. breaking the model class into smaller simpler classes appropriate for the amount of data on hand) and the second that of identifiability of the subclasses in the regularization

Cover (1973): is the bias of a coin rational?

Coin tosses: $X_1, X_2, \ldots \sim p$

After each toss: decide if p rational or not?

Finitely many errors?

Cover (1973): is the bias of a coin rational?

Coin tosses: $X_1, X_2, \ldots \sim p$

After each toss: decide if p rational or not?

Finitely many errors?

Seem impossible when rationals are dense in the real line

Cover (1973): is the bias of a coin rational?

Coin tosses: $X_1, X_2, \ldots \sim p$

After each toss: decide if p rational or not?

Finitely many errors?

Seem impossible when rationals are dense in the real line, but in fact, there is a scheme that makes only finitely many errors!

(for all rational, all irrationals except a Lebesgue measure 0 set)

Cover (1973): is the bias of a coin rational?

Coin tosses: $X_1, X_2, \ldots \sim p$

After each toss: decide if p rational or not?

Finitely many errors?

Seem impossible when rationals are dense in the real line, but in fact, there is a scheme that makes only finitely many errors!

(for all rational, all irrationals except a Lebesgue measure 0 set)

In anticipation of our results, we will take a regularization view

Regularization

Let $r_1, r_2, \ldots$ be an enumeration of rational numbers in $[0, 1]$

Build set $\mathcal{S}_n$ as follows:

0 1

Regularization

Let $r_1, r_2, \ldots$ be an enumeration of rational numbers in $[0, 1]$

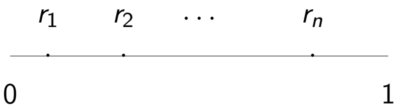
Build set $\mathcal{S}_n$ as follows:

0 only irrationals 1

Regularization

Let $r_1, r_2, \ldots$ be an enumeration of rational numbers in $[0, 1]$

Build set $\mathcal{S}_n$ as follows:

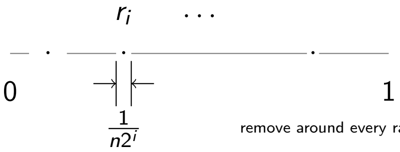


add back in first n rationals

Regularization

Let $r_1, r_2, \ldots$ be an enumeration of rational numbers in $[0, 1]$

Build set $\mathcal{S}_n$ as follows:

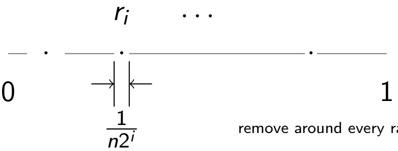


remove around every rational even if not added back

Regularization

Let $r_1, r_2, \ldots$ be an enumeration of rational numbers in $[0, 1]$

Build set $\mathcal{S}_n$ as follows:



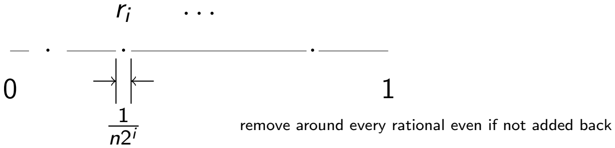
remove around every rational even if not added back

Note $\mathcal{S}_1 \subset \mathcal{S}_2 \subset \ldots$

Regularization

Let $r_1, r_2, \ldots$ be an enumeration of rational numbers in $[0, 1]$

Build set $\mathcal{S}_n$ as follows:



Note $\mathcal{S}_1 \subset \mathcal{S}_2 \subset \ldots$
In $\mathcal{S}_n$, total measure removed $\leq \frac{1}{n}$. If

$$\mathcal{S} = \bigcup_n \mathcal{S}_n,$$

$\mathcal{S}$ has measure 1 and contains every rational.

Prediction in each subclass $\mathcal{S}_n$

In each $\mathcal{S}_n$: rational vs irrational with confidence $1 - 2^{-n}$?

Prediction in each subclass $\mathcal{S}_n$

In each $\mathcal{S}_n$: rational vs irrational with confidence $1 - 2^{-n}$?

Every rational in $\mathcal{S}_n$ is at least $\frac{1}{n2^n}$ away from an irrational

Prediction in each subclass $\mathcal{S}_n$

In each $\mathcal{S}_n$: rational vs irrational with confidence $1 - 2^{-n}$?

Every rational in $\mathcal{S}_n$ is at least $\frac{1}{n2^n}$ away from an irrational

For any confidence, in particular $1 - 2^{-n}$, there exists sample size b_n large enough that we can decide rationality of sources in $\mathcal{S}_n$

Prediction for $\mathcal{S}$

What about $\mathcal{S}$?

Prediction for $\mathcal{S}$

What about $\mathcal{S}$?

Break into phases
n'th phase: $b_n \leq$ sample size $< b_{n+1}$, use estimaxtor for $\mathcal{S}_n$

Prediction for $\mathcal{S}$

What about $\mathcal{S}$?

Break into phases
n'th phase: $b_n \leq$ sample size $< b_{n+1}$, use estimaxtor for $\mathcal{S}_n$

Every rational in $\mathcal{S}$ will eventually show up in $\mathcal{S}_m$ for some finite m, after which, the probability of error is $1/2^m$ in any phase

Prediction for $\mathcal{S}$

What about $\mathcal{S}$?

Break into phases
n'th phase: $b_n \leq$ sample size $< b_{n+1}$, use estimaxtor for $\mathcal{S}_n$

Every rational in $\mathcal{S}$ will eventually show up in $\mathcal{S}_m$ for some finite m, after which, the probability of error is $1/2^m$ in any phase
error only in finite number of phases (Borel-Cantelli)

Prediction for $\mathcal{S}$

What about $\mathcal{S}$?

Break into phases
n'th phase: $b_n \leq$ sample size $< b_{n+1}$, use estimaxtor for $\mathcal{S}_n$

Every rational in $\mathcal{S}$ will eventually show up in $\mathcal{S}_m$ for some finite m, after which, the probability of error is $1/2^m$ in any phase
error only in finite number of phases (Borel-Cantelli)

For any irrational in $\mathcal{S}$, error in each phase m is $1/2^m$,

Prediction for $\mathcal{S}$

What about $\mathcal{S}$?

Break into phases
n'th phase: $b_n \leq$ sample size $< b_{n+1}$, use estimaxtor for $\mathcal{S}_n$

Every rational in $\mathcal{S}$ will eventually show up in $\mathcal{S}_m$ for some finite m, after which, the probability of error is $1/2^m$ in any phase
error only in finite number of phases (Borel-Cantelli)

For any irrational in $\mathcal{S}$, error in each phase m is $1/2^m$,
again error only in finite number of phases (Borel-Cantelli)

Prediction for $\mathcal{S}$

What about $\mathcal{S}$?

Break into phases
n'th phase: $b_n \leq$ sample size $< b_{n+1}$, use estimaxtor for $\mathcal{S}_n$

Every rational in $\mathcal{S}$ will eventually show up in $\mathcal{S}_m$ for some finite m, after which, the probability of error is $1/2^m$ in any phase
error only in finite number of phases (Borel-Cantelli)

For any irrational in $\mathcal{S}$, error in each phase m is $1/2^m$,
again error only in finite number of phases (Borel-Cantelli)

Therefore, no matter what the source, only finite number of errors!

Regularization is also necessary!

We show the converse also holds: if any $\mathcal{S}$ admits a finite-error rationality estimator with only finite number of errors, then

$$\mathcal{S} = \bigcup_n \mathcal{S}_n$$

where $\mathcal{S}_1 \subset \mathcal{S}_2 \subset \cdots$ and each $\mathcal{S}_n$ satisfies

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

(Wu-Santhanam, arxiv)

Regularization is also necessary!

We show the converse also holds: if any $\mathcal{S}$ admits a finite-error rationality estimator with only finite number of errors, then

$$\mathcal{S} = \bigcup_n \mathcal{S}_n$$

where $\mathcal{S}_1 \subset \mathcal{S}_2 \subset \cdots$ and each $\mathcal{S}_n$ satisfies

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

(Wu-Santhanam, arxiv)
Namely, each $\mathcal{S}_n$ can be handled with arbitrary confidence with a finite sample size

If $\mathcal{S}$ admits a finite-error rationality predictor, then we can always find a regularization to tackle it

Rank Estimation

Let $\mathbf{X}$ be a $d \times d$ random matrix with entries $X_{i,j}$ to be independent Bernoulli random variables. Denote $p_{i,j} = \mathbb{E}[X_{i,j}]$ and $\mathbb{E}[X]$ be the matrix with entries $p_{i,j}$.

Rank Estimation

Let $\mathbf{X}$ be a $d \times d$ random matrix with entries $X_{i,j}$ to be independent Bernoulli random variables. Denote $p_{i,j} = \mathbb{E}[X_{i,j}]$ and $\mathbb{E}[X]$ be the matrix with entries $p_{i,j}$.

$\mathbf{X}_1, \cdots, \mathbf{X}_n$ are *i.i.d.* samples of $\mathbf{X}$, which are $d \times d$ binary matrices.

Rank Estimation

Let $\mathbf{X}$ be a $d \times d$ random matrix with entries $X_{i,j}$ to be independent Bernoulli random variables. Denote $p_{i,j} = \mathbb{E}[X_{i,j}]$ and $\mathbb{E}[X]$ be the matrix with entries $p_{i,j}$.

$\mathbf{X}_1, \cdots, \mathbf{X}_n$ are *i.i.d.* samples of $\mathbf{X}$, which are $d \times d$ binary matrices.

How could we *reasonably* estimate $\text{Rank}(\mathbb{E}[\mathbf{X}])$ by observing $\mathbf{X}_1, \cdots, \mathbf{X}_n$?

Rank Estimation

A naive way is to compute

$$\bar{\mathbf{X}}_n = \frac{1}{n}\sum_{k=1}^{n}\mathbf{X}_k,$$

and use $\text{Rank}(\bar{\mathbf{X}}_n)$ as an estimation of $\text{Rank}(\mathbb{E}[\mathbf{X}])$.

Rank Estimation

A naive way is to compute

$$\bar{\mathbf{X}}_n = \frac{1}{n} \sum_{k=1}^{n} \mathbf{X}_k,$$

and use $\mathrm{Rank}(\bar{\mathbf{X}}_n)$ as an estimation of $\mathrm{Rank}(\mathbb{E}[\mathbf{X}])$.

However, such an estimation is not *reasonable* since $\bar{\mathbf{X}}_n$ is full rank w.h.p. even for matrices $\mathbb{E}[\mathbf{X}]$ with same entries.

Rank Estimation

It seems one can't estimate the rank at all, since arbitrary small perturbation on $\mathbb{E}[\mathbf{X}]$ will significantly change the rank.

Rank Estimation

It seems one can't estimate the rank at all, since arbitrary small perturbation on $\mathbb{E}[\mathbf{X}]$ will significantly change the rank.

We show that there exist an estimator Φ such that

$$\Phi(\mathbf{X}_1, \cdots, \mathbf{X}_n) \to \text{Rank}[\mathbf{X}] \text{ w.p. } 1$$

as $n \to \infty$.

eas-predictable

$\mathcal{P}$: class of models over support $\mathbb{N}$
$X_1, X_2, \cdots \sim p \in \mathcal{P}$

eas-predictable

$\mathcal{P}$: class of models over support $\mathbb{N}$
$X_1, X_2, \cdots \sim p \in \mathcal{P}$

At step n: learner outputs $Y(X_1, \ldots, X_n)$ and is scored with a binary loss

$$\ell : \mathcal{P} \times \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$$

(Property we predict implictly defined by the set $\ell = 0$)

EAS-predictable

The pair $\mathcal{P}$, ℓ is *eventually almost surely* predictable if a learner Y achieves $\forall p \in \mathcal{P}$

$$p\left(\sum_{n=1}^{\infty} \ell\left(p, Y(X_1, \ldots, X_n), X_{n+1}\right) < \infty\right) = 1.$$

Main idea

As in Cover's case, we will connect eas-predictability to one that can be done with finite number of samples.

$\eta-$predictable

$\mathcal{Q}$ with loss ℓ is $\eta-$predictable if there exists a learner and number N_η such that $\forall p \in \mathcal{Q}$

$$p\left(\sum_{n=N_\eta}^{\infty} \ell(p, Y_n, X_{n+1}) > 0\right) \leq \eta$$

$\eta-$predictable

$\mathcal{Q}$ with loss ℓ is $\eta-$predictable if there exists a learner and number N_η such that $\forall p \in \mathcal{Q}$

$$p\left(\sum_{n=N_\eta}^{\infty} \ell(p, Y_n, X_{n+1}) > 0\right) \leq \eta$$

$\eta-$nesting For $\eta > 0$, $\mathcal{P}_1 \subset \mathcal{P}_2 \cdots$ with $\bigcup_n \mathcal{P}_n = \mathcal{P}$ is an $\eta-$nesting of $\mathcal{P}$ if each $\mathcal{P}_n$ is $\eta-$predictable

$\eta-$predictable

$\mathcal{Q}$ with loss ℓ is $\eta-$predictable if there exists a learner and number N_η such that $\forall p \in \mathcal{Q}$

$$p\left(\sum_{n=N_\eta}^{\infty} \ell(p, Y_n, X_{n+1}) > 0\right) \leq \eta$$

$\eta-$nesting For $\eta > 0$, $\mathcal{P}_1 \subset \mathcal{P}_2 \cdots$ with $\bigcup_n \mathcal{P}_n = \mathcal{P}$ is an $\eta-$nesting of $\mathcal{P}$ if each $\mathcal{P}_n$ is $\eta-$predictable

Universal nesting $\mathcal{P}_1 \subset \mathcal{P}_2 \cdots$ with $\bigcup_n \mathcal{P}_n = \mathcal{P}$ is an universal nesting of $\mathcal{P}$ if for all $\eta > 0$, each $\mathcal{P}_n$ is $\eta-$predictable

Characterization:

Theorem

If there is a universal nesting of $\mathcal{P}$, $(\mathcal{P}, \ell)$ is e.a.s.-predictable.
If $(\mathcal{P}, \ell)$ is e.a.s.-predictable then for each $\eta > 0$, there is an $\eta-$nesting of $\mathcal{P}$.

Characterization:

Theorem

If there is a universal nesting of $\mathcal{P}$, $(\mathcal{P}, \ell)$ is e.a.s.-predictable.
If $(\mathcal{P}, \ell)$ is e.a.s.-predictable then for each $\eta > 0$, there is an $\eta-$nesting of $\mathcal{P}$.

This base result can be strengthened in several ways as we will see. While the result above is intuitive, its usage in various contexts is what is interesting.

Applications

Diagonalization arguments often yield a matching converse

Applications

Diagonalization arguments often yield a matching converse

Insurance: Given $X_1, \ldots, X_n$ predict an upper bound on the next sample (loss $=0$ if prediction $\Phi(X_1, \ldots, X_n) > X_{n+1}$.

Applications

Diagonalization arguments often yield a matching converse

Insurance: Given $X_1, \ldots, X_n$ predict an upper bound on the next sample (loss =0 if prediction $\Phi(X_1, \ldots, X_n) > X_{n+1}$.

Finite errors iff $\mathcal{P} = \bigcup_n \mathcal{P}_n$, $\mathcal{P}_n$ tight (Wu, Santhanam 19)

Applications

Diagonalization arguments often yield a matching converse

Insurance: Given $X_1, \dots, X_n$ predict an upper bound on the next sample (loss =0 if prediction $\Phi(X_1, \dots, X_n) > X_{n+1}$.

Finite errors iff $\mathcal{P} = \bigcup_n \mathcal{P}_n$, $\mathcal{P}_n$ tight (Wu, Santhanam 19)

Classification: Given an instance space $\mathbb{R}^d$, a hypothesis space $\mathcal{H}$ and examples $X_i, h(X_i)$, $i = 1, \dots, n$, chosen from an arbitrary dist. μ, predict $h(X_{n+1})$.

Applications

Diagonalization arguments often yield a matching converse

Insurance: Given $X_1, \ldots, X_n$ predict an upper bound on the next sample (loss $=0$ if prediction $\Phi(X_1, \ldots, X_n) > X_{n+1}$.

Finite errors iff $\mathcal{P} = \bigcup_n \mathcal{P}_n$, $\mathcal{P}_n$ tight (Wu, Santhanam 19)

Classification: Given an instance space $\mathbb{R}^d$, a hypothesis space $\mathcal{H}$ and examples $X_i, h(X_i)$, $i = 1, \ldots, n$, chosen from an arbitrary dist. μ, predict $h(X_{n+1})$.

Finite errors iff $\mathcal{H} = \bigcup_n \mathcal{H}_n$, $\mathcal{H}_n$ effectively single hypothesis (WS, submitted)

Applications

Diagonalization arguments often yield a matching converse

Insurance: Given $X_1, \ldots, X_n$ predict an upper bound on the next sample (loss $=0$ if prediction $\Phi(X_1, \ldots, X_n) > X_{n+1}$.

Finite errors iff $\mathcal{P} = \bigcup_n \mathcal{P}_n$, $\mathcal{P}_n$ tight (Wu, Santhanam 19)

Classification: Given an instance space $\mathbb{R}^d$, a hypothesis space $\mathcal{H}$ and examples $X_i, h(X_i)$, $i = 1, \ldots, n$, chosen from an arbitrary dist. μ, predict $h(X_{n+1})$.

Finite errors iff $\mathcal{H} = \bigcup_n \mathcal{H}_n$, $\mathcal{H}_n$ effectively single hypothesis (WS, submitted)

Other formulations: entropy estimation (Wu-Santhanam, submitted), rank of matrices (Wu-Santhanam, arxiv), estimation of Markov chains $\cdots$

University of Hawai'i Mānoa

Applications

Diagonalization arguments often yield a matching converse

Insurance: Given $X_1, \ldots, X_n$ predict an upper bound on the next sample (loss =0 if prediction $\Phi(X_1, \ldots, X_n) > X_{n+1}$.

Finite errors iff $\mathcal{P} = \bigcup_n \mathcal{P}_n$, $\mathcal{P}_n$ tight (Wu, Santhanam 19)

Classification: Given an instance space $\mathbb{R}^d$, a hypothesis space $\mathcal{H}$ and examples $X_i, h(X_i)$, $i = 1, \ldots, n$, chosen from an arbitrary dist. μ, predict $h(X_{n+1})$.

Finite errors iff $\mathcal{H} = \bigcup_n \mathcal{H}_n$, $\mathcal{H}_n$ effectively single hypothesis (WS, submitted)

Other formulations: entropy estimation (Wu-Santhanam, submitted), rank of matrices (Wu-Santhanam, arxiv), estimation of Markov chains $\cdots$

Open Problem: Is universal nesting necessary in general?

Strengthening other results

Guiding technique here is finding appropriate decompositions

Doing so allows us to recover all the results in (Dembo-Peres, 94) and (Koplowitz et al., 97) with simple elementary proofs

Moreover, our approach provides stronger converse theorems than in (Dembo-Peres, 94)

Even though eas-predictable class have prediction rules that make only finitely many errors, we do not have any guarantee on when it will stop making errors...

Even though eas-predictable class have prediction rules that make only finitely many errors, we do not have any guarantee on when it will stop making errors...

"This is a characterization of the problem and is not a fault of the test" – (Cover, 1973)

Even though eas-predictable class have prediction rules that make only finitely many errors, we do not have any guarantee on when it will stop making errors...

"This is a characterization of the problem and is not a fault of the test" – (Cover, 1973)

However, from practical consideration, one may still hope a stopping rule that specifies when the mistakes will stop.

e.a.s.-learnable

Suppose $(\mathcal{P}, \ell)$ is *e.a.s.*-predictable.

If for any $\eta > 0$ there is a stopping rule τ_η that predicts with confidence $1 - \eta$ when we have made the last error, then $(\mathcal{P}, \ell)$ is *e.a.s.*-learnable.

Identifiability

Let $\mathcal{U}$ be a collection of *i.i.d.* processes over sequences of naturals and $\mathcal{Q} \subset \mathcal{U}$.

$\mathcal{Q}$ is identifiable in $\mathcal{U}$ if $1(p \in \mathcal{Q})$ is *e.a.s.*-learnable.

For example, $\mathcal{Q}$ is identifiable in $\mathcal{U}$ iff the single letter marginals of $\mathcal{Q}$ are relatively open in $\mathcal{U}$ with respect to ℓ_1 metric.

More involved definition for non *i.i.d.* collections in terms of universal nesting of $\mathcal{Q}$ for the property $1(p \in \mathcal{Q})$.

Characterization of eas-learnable

Theorem

A class $\mathcal{P}$ with a loss ℓ is eas-learnable, if there is a nesting $\{\mathcal{P}_n\}_{n\in\mathbb{N}}$ of $\mathcal{P}$ such that

1. *For all $n \in \mathbb{N}$, $(\mathcal{P}_n, \ell)$ is uniformly predictable;*
2. *For all $n \in \mathbb{N}$, $\mathcal{P}_n$ is identifiable in $\mathcal{P}$.*

Again, the converse holds in several problems as we will see

Applications

Matching converses again. All problems only require with high confidence.

Applications

Matching converses again. All problems only require with high confidence.

Insurance: Given $X_1, \ldots, X_n$ predict an upper bound on the next sample (loss $=0$ if prediction $\Phi(X_1, \ldots, X_n) > X_{n+1}$)

Learnable iff $\mathcal{P} = \bigcup_n \mathcal{P}_n$, $\mathcal{P}_n$ tight, relatively open

(Santhanam, Anantharam 16)

Applications

Compression: Given *i.i.d.* samples from some $p \in \mathcal{P}$, find universal compressor q and a stopping time such that per-symbol codelength difference falls and remains $\leq \delta$ (Santhanam, Anantharam, Szpankowski) Tomorrow afternoon?

Countable collection of "compressible" classes

Applications

Markov estimation: Samples from a binary Markov source with arbitrary memory (and arbitrarily slow mixing), given accuracy ϵ, estimate conditional and stationary probabilities associated with arbitrary strings. Stopping rule (Asadi-Paravi-Santhanam 14-17, Wu-Santhanam, arxiv)

Coupling from the past, continuity condition

Clustering algorithms (Paravi-Santhanam 18)

Conclusion

Our framework provides a way of resolving estimation and prediction problems that involve (really) large model class.

The construction of eas-prediction rules will often result in a natural regularization on the model classes

The eas-learning framework could be used as an alternative for uniform consistency in very rich settings

Other things we are thinking about

Bayesian priors: brittle vs. not brittle

Learning: (when) can you uniformly sample from the space of all labelings? (Wu, Santhanam 20)

Feedforward neural networks with threshold activations

Ad-hoc: Use predictions on eigenvalue-related properties during training?

Conclusion

Several extensions may be considered for further research:

1. Consider restricted prediction rules, e.g. computational bounded predictors (partial results in (Wu-Santhanam, submitted) ;
2. Consider interactive sampling process, i.e. the prediction will affect the sampling
3. Bounds on the stopping time, e.g. optimal expectation of the stopping time

Thank you!