The rate of convergence of computable predictions

Kenshi Miyabe

Meiji University

The 2020 annual meeting of the German Mathematical Society

15 Sep 2020

Table of Contents

Motivation

Question 1

Given a random data, how fast should a general prediction converge?

The motivation is based on the theory of inductive inference by Solomonoff, later developed by Hutter.

To overcome some weaknesses, we focus on computable predictions and study their properties.

Solovay reducibility plays an important role to do this.

Ref: Hutter, "Universal Artificial Intelligence", Springer, 2005.

Setting

2^ω : Cantor space

μ : an unknown computable measure on 2^ω

$X \in 2^\omega$: a (computably) random w.r.t. μ, given in succession

$$X = X_1 X_2 \cdots$$

$$X_{<n} = X_1 X_2 \cdots X_{n-1}$$

Task : Given $X_{<n}$, we want to compute a good approximation of

$$\frac{\mu(X_{<n}k)}{\mu(X_{<n})}$$

Some results in 60s-70s

Let ξ, ν be measures or semi-measures.

We say that ξ multiplicatively dominates ν if there exists $C \in \mathbb{N}$ such that $\nu(\sigma) \leq C\xi(\sigma)$ for all $\sigma \in 2^{<\omega}$.

There exists an optimal c.e. semi-measure ξ, that is, ξ dominates all c.e. semimeasures.

Theorem 2 (Solomonoff)

Let μ be a computable measure. For any optimal c.e. semi-measure ξ, the following holds for μ-a.s. X:

$$|\xi(k|X_{<n}) - \mu(k|X_{<n})| \to 0$$

Problems

For optimal ξ, $\xi(\cdot|\cdot)$ is Δ_2 but not computable.

The theory can talk about optimal (so non-computable) predictions,

but cannot talk about computable predictions.

Now we introduce a framework to study sufficiently general computable predictions.

Generality

We say that $P(n)$ holds for all sufficiently large n if there exists $N \in \mathbb{N}$ such that $P(n)$ holds for all $n \geq N$.

Definition 3

$P(\xi)$ holds for all sufficiently general computable measures ξ if there exists a computable measure ν such that $P(\xi)$ holds for all computable measures ξ m-dominating ν.

m-domination and generality

If ξ m-dominates ν, then ξ can converge for a larger class of μ than ν.

Thus, ξ learns more tasks than ν.

In this case, we say ξ is more general than ν.

A weak converge also holds.

Similar phenomena

I come up with this notion inspired by the following results:

1. For all sufficiently fast-growing t, time-bounded Kolmogorov complexity $K^t(n)$ is a Solovay function, that is, a good approximation of K.
2. If the convergence of a left-c.e. real α is sufficiently slow, then α is ML-random.

Table of Contents

The rate of convergence

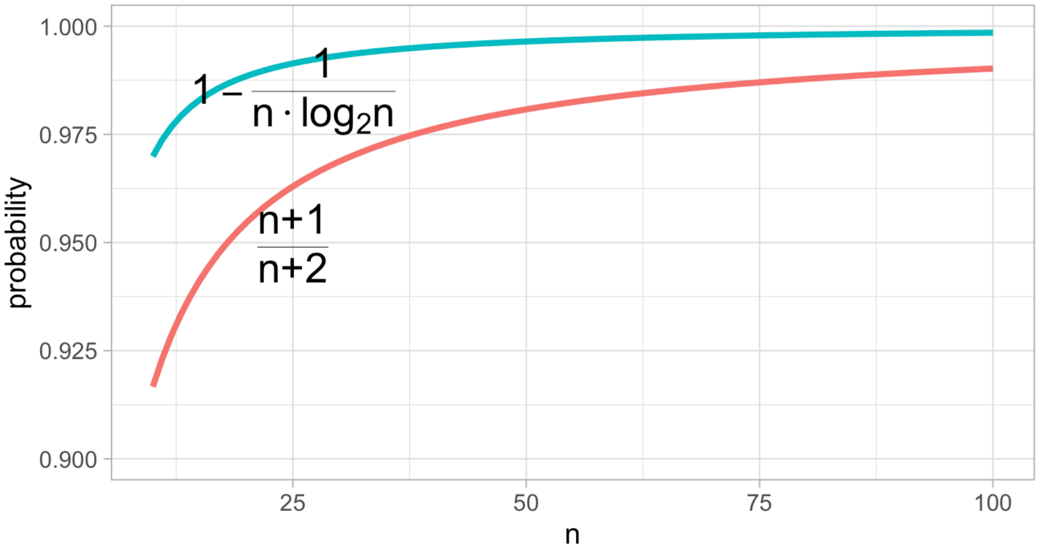
Theorem 4

Let $A \in 2^\omega$ be a computable sequence.

1. *For every computable prediction ξ m-dominating $\mu = \mathbf{1}_A$, there exists a computable function $h : \mathbb{N} \to \mathbb{N}$ such that $K^h(n) \leq -\log(1 - \xi(A_n|A_{<n})) + O(1)$.*
2. *For every computable function $h : \mathbb{N} \to \mathbb{N}$, we have $-\log(1 - \xi(A_n|A_{<n})) \leq K^h(n) + O(1)$ for all sufficiently general computable prediction measure ξ.*
3. *The sum $\sum_n (1 - \xi(A_n|A_{<n}))$ is a finite left-c.e. ML-random real for all sufficiently general computable prediction measure ξ.*

The rate of convergence

As partially noted by Solomonoff, the convergence is slightly faster than the probability of the sunrise problem by Laplace.



1.000
0.975
0.950
0.925
0.900
probability
25
50
75
100
n
$1 - \frac{1}{n \cdot \log_2 n}$
$\frac{n+1}{n+2}$

The rate of convergence

We say that μ is separated (from 0 and 1) if $\inf_{\sigma\in 2^{<\omega}, k\in 2}\mu(k|\sigma) > 0$.

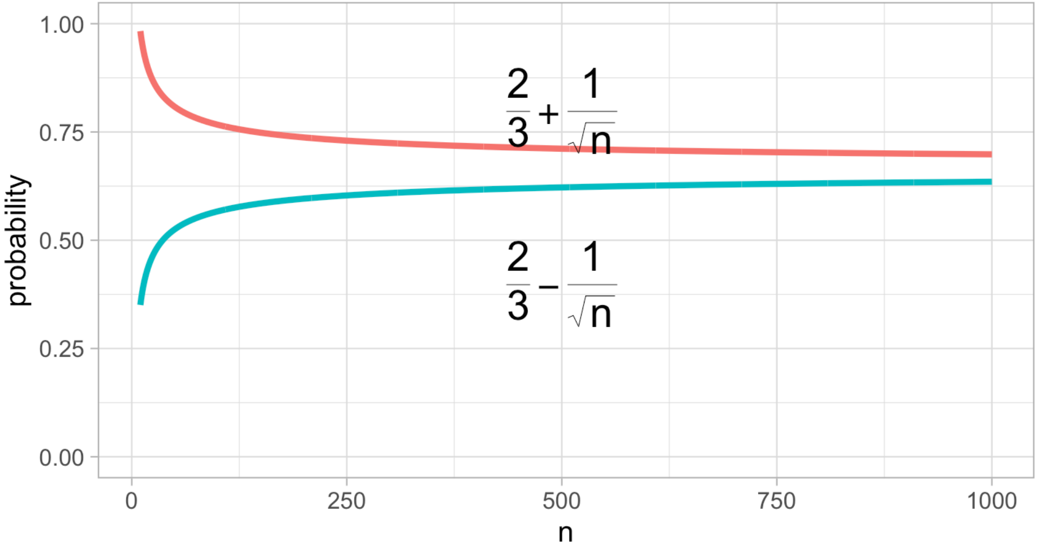
Theorem 5

For a separated computable measure μ, a μ-computably random seq. X, and a sufficiently general computable measure ξ, we have

$$\sum_{n=1}^{\infty} |\xi(X_n|X_{<n}) - \mu(X_n|X_{<n})|^2 < \infty$$

We cannot replace "2" in the exponent with a smaller positive one.

This corresponds to a claim by the central limit theorem.



1.00
0.75
0.50
0.25
0.00
probability
0
250
500
750
1000
n
$\frac{2}{3}+\frac{1}{\sqrt{n}}$
$\frac{2}{3}-\frac{1}{\sqrt{n}}$

Impossibility

There are some properties we cannot force.

Theorem 6

Let μ be a separated computable measure and X be μ-computably random. For any computable measure ν, there exists a computable measure ξ such that

1. ξ *m-dominates* ν,
2. $\xi(\cdot|X_{<n}) = \mu(\cdot|X_{<n})$

Thus, $\xi(\cdot|\cdot)$ can accidentally coincide with $\mu(\cdot|\cdot)$, and can accidentally be different from $\mu(\cdot|\cdot)$.

However, we can force to keep the distance from $\mu(\cdot|\cdot)$ in expectation.

Table of Contents

KL-divergence

We define the KL-divergence of μ w.r.t. ξ by

$$D(\mu||\xi) = \int \frac{d\mu}{d\xi} \log \frac{d\mu}{d\xi} d\xi = \int \log \frac{d\mu}{d\xi} d\mu$$

where $0 \cdot \log 0 = 0$.

If ξ m-dominates μ, then the Radon-Nikodym derivative exists:

$$\frac{d\mu}{d\xi}(X) = \lim_n \frac{\mu(X_{\leq n})}{\xi(X_{\leq n})}.$$

Key theorem

Theorem 7

Let μ be a computable model measure on $\{0,1\}^{\mathbb{N}}$. Then, the predictability of μ w.r.t. ξ

$$D(\mu||\xi) = E_\mu(\sum_{n=1}^{\infty} D_n^X(\mu||\xi))$$

is a finite left-c.e. ML-random real for all sufficiently general computable measures ξ.

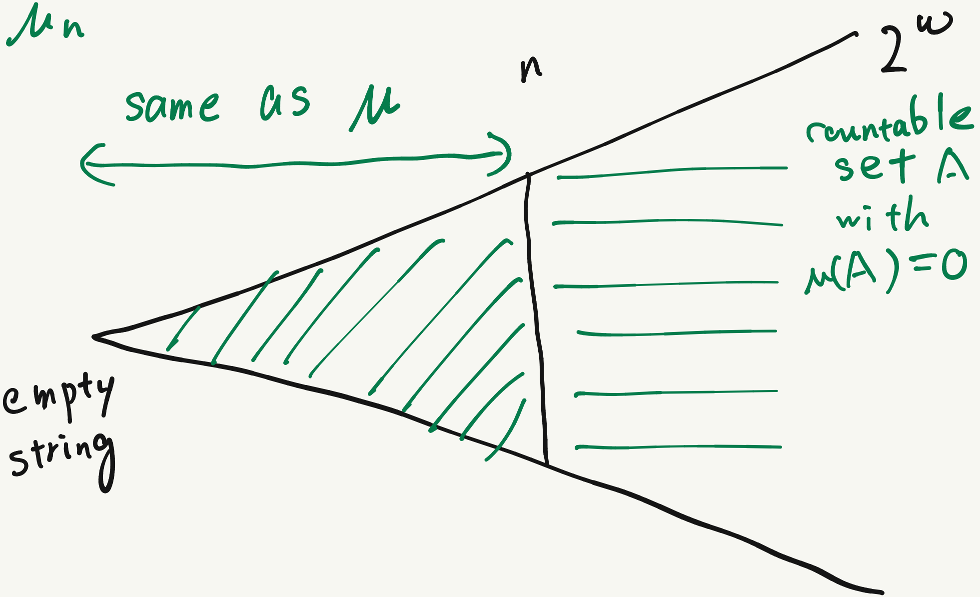
The chain rule for KL-divergence can be proved by the martingale convergence theorem.

Key theorem

The goal is to construct a computable measure ν such that

1. $D(\mu||\nu)$ is ML-random,
2. if ξ m-dominates ν, then $D(\mu||\nu)$ is Solovay reducible to $D(\mu||\xi)$.

Here, ν can depend on μ.


μ_n
2^ω
n
same as μ
countable set A with $\mu(A)=0$
empty string

Key theorem

We construct a computable measure

$$\nu = \sum_n z_n \mu_n + (1 - s)\mu$$

where $s = \sum_n z_n$ is ML-random. Then,

$$D(\mu || \nu) = \log \frac{1}{1 - s}$$

is also ML-random.

Key theorem

We construct a computable function g such that

$$E_\mu\Big(\sum_{t>g(n)} D_t^X(\mu||\nu)\Big) \leq \log\left(1 + \sum_{t>n} z_t\right) + z_{n+1} \leq c_0 \sum_{t>n} z_n$$

and

$$E_\mu\Big(\sum_{t>n} D_t^X(\mu||\xi)\Big) \geq \frac{1}{c_1} \sum_{t>n} z_t.$$

The former holds because the effect of ν up to n decreases exponentially.

The latter comes from $\nu \leq c'\xi$.

Summary

1. We proposed a framework to study the properties of general computable predictions.
2. In this framework, we gave the rate of convergence for deterministic measures and separated measures.
3. The strong connection between computable randomness and KL-divergence.
4. Solovay reducibility seems a very useful tool in this study.

Thank you for listening.

明治大学
MEIJI UNIVERSITY