Computable prediction

Kenshi Miyabe (宮部賢志) @ Meiji University

8 Aug 2019
AGI-19 @ Shenzhen, China

Motivation

Question

It is useful if one can answer:

Question 1. How do we construct a program which learns general regularity?

Let me ask:

Question 2. What properties should a general learning program have?

Possible? Yes, one can prove this in some setting.

Formalization

Formalizing learning by Solomonoff's setting:

- Underlying space : $2^\omega = \{0,1\}^{\mathbb{N}}$ (for simplicity).
- Sample : $X \in 2^\omega$ is sampled randomly from a computable measure μ.
- Learner : computable measure ξ.

The learner should be computable, otherwise cannot be implemented.
The underlying measure should also be computable, otherwise one cannot predict.

Computability

$f : \omega \to \omega$ is computable if it can be implemented by a Turing machine.
$\mathbb{Q}, 2^{<\omega}$ has a natural representation via ω.
A sequence $\{q_n\}_n$ of rationals is computable if $n \mapsto q_n$ is computable.
$x \in \mathbb{R}$ is computable if there exists a computable sequence $\{q_n\}_n$ of rationals such that $|x - q_n| \leq 2^{-n}$ for all n.
The measure μ on 2^ω is computable if there exists a computable function $f : \omega \times 2^{<\omega} \to \mathbb{Q}$ such that $|\mu([\sigma]) - f(n, \sigma)| \leq 2^{-n}$ for all n where $[\sigma] = \{X \in 2^\omega \ : \ \sigma \prec X\}$.

Reward

A martingale w.r.t. μ is $M : 2^{<\omega} \to \mathbb{R}^+$ such that

$$\mu(\sigma)M(\sigma) = \mu(\sigma 0)M(\sigma 0) + \mu(\sigma 1)M(\sigma 1).$$

This can be seen as a capital process.
Good predictions has a rapid capital grow.
Natural correspondence between a martingale M and a measure ξ by

$$\xi(\sigma) = \mu(\sigma)M(\sigma)$$

We do not know the underlying measure μ so we use a measure ξ in place of a martingale M.

Optimality

ξ behaves better than ν if $\xi(\sigma) \geq \nu(\sigma)$ for all σ.
No computable measure can behave best.

Theorem 3 (Classical)**.** *There exists an* optimal *c.e. semi-measure ξ, that is, for any c.e. semi-measure ν, there exists $C \in \omega$ such that*

$$\nu(\sigma) \leq C\xi(\sigma)$$

for all $\sigma \in 2^{<\omega}$.

Notice that the class of c.e. semi-measures is countable.

Generality

The optimal prediction may not behave well in short term, but behaves not badly compared to any other measure.
This generality prevents the overfitting problem.

Definition 4. A computable measure ξ multiplicatively dominates (or m-dominates) ν if there exists $C \in \omega$ such that

$$\nu(\sigma) \leq C\xi(\sigma)$$

for all $\sigma \in 2^{<\omega}$.

Roughly speaking, ξ is more general than ν.

Combination

$\mu_1, \mu_2, \cdots$ uniformly computable measures.
Let $\mu = \sum_n 2^{-n} \mu_n$.
Then, μ multiplicatively dominates μ_n for all n.
Roughly speaking, μ is more general than μ_n.

Proposition 5. *No computable measure is optimal.*

This is the reason that c.e. semi-measures have been studied extensively in the literature.

Question

Question 6. What properties a sufficiently general prediction should have?

$a_n > M$ for sufficiently large n if

$$(\exists N \in \omega)(\forall n \geq N) a_n > M$$

We say that a property P holds for all sufficiently general prediction if there exists a computable measure ν such that P holds for any ξ m-dominating ν.
In this case, P is witnessed by ν.

Results

Result 1

Theorem 7. *μ : comp. measure on 2^ω.*
ξ : comp. measure m-dominating μ.
$X \in 2^\omega$: μ-computably random.

$$\sum_{n=1}^{\infty} D(\mu(\cdot|X_{<n}) \,||\xi(\cdot|X_{<n})) < \infty.$$

In particular,

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

for both $k \in \{0, 1\}$.

Here, D is the KL-divergence.

Result 2

When μ is a Dirac measure, we can compute the speed of the convergence.

Theorem 8. $A \in 2^\omega$ *: computable*
Then, $\exists\nu$ *: comp. measure s.t.* $\forall\xi$ *m-dominating* ν

$$\sum_n (1 - \xi(A_n | A_{<n})) < \infty$$

Result 3

Theorem 9. $A \in 2^\omega$ *: comp.*
$(a_n)_n$ *: comp. seq. with* $\sum_n a_n < \infty$.
Then, $\exists \nu$ *: comp. meas. s.t.* $\forall \xi$ *m-dominating* ν
$\exists C \in \omega$ *s.t.*

$$\xi(\overline{A_n} | A_{<n}) \geq \frac{a_n}{C}$$

for all n.

Thus, all sufficiently general prediction converge at the same speed up to multiplicative constant!!

Laplace's result

The probability of the correct n-th bit via a general prediction is roughly

$$1 - \frac{C}{n(\log n)^{1+\epsilon}}$$

where C is a constant although the probability cannot be monotone.
Compare to the Laplace's result to the sunrise problem :

$$\frac{n+1}{n+2}$$

Overfitting problem

Another interpretation is possible :
If

$$\sum_n (1 - \xi(A_n | A_{<n})) = \infty,$$

then the convergence is slower than the ”correct” one and the learner fails to find the regularity of A.

If the convergence of

$$\sum_n (1 - \xi(A_n | A_{<n})) < \infty$$

is too fast, then the convergence is faster than the ”correct” one and the learner overfits A.

Hypothesis

Check again the hypothesis which makes this argument possible :

(i) The learner is computable, so the class is countable.
(ii) The definition of generality uses how fast the capital grows up to multiplicative constant.
(iii) The underlying measure is computable, really weak restriction.

If one considers only i.i.d., then the result may change. The underlying space may be generalized using computable analysis.

Proof

Randomness

μ : comp. meas. on 2^ω.

Definition 10 (Rute 2016)**.** A μ-martingale is a partial function $M :\subseteq 2^{<\omega} \to \mathbb{R}^+$ s.t.

(i) (Impossibility condition) If $M(\sigma)$ is undefined, then $\mu(\sigma) = 0$.

(ii) (Fairness condition) For all $\sigma \in 2^{<\omega}$, we have

$$M(\sigma 0)\mu(\sigma 0) + M(\sigma 1)\mu(\sigma 1) = M(\sigma)\mu(\sigma)$$

where undefined $\cdot\, 0 = 0$ and $\mathbb{R}^+$ is the set of all non-negative reals.

Existence of the limit

Theorem 11. *$X \in 2^\omega$ is μ-computably random if and only if $\lim_{n\to\infty} M(X_{\leq n})$ exists for all a.e. computable μ-martingales M.*

For Martin-Löf randomness, we do not know any characterization via such existence of the limit.

Convergence

μ : comp. meas. on 2^ω.
Let $C \in \omega$ s.t. $\mu(\sigma) \le C\xi(\sigma)$ and

$$D(\sigma) = D(\mu(\cdot|\sigma) \mid\mid \xi(\cdot|\sigma))$$

Then,

$$M(\sigma) = \ln C - \ln \frac{\mu(\sigma)}{\xi(\sigma)} + \sum_{t=1}^{|\sigma|} D(\sigma_{<t})$$

is a μ-martingale.
Its convergence implies the convergence of D to 0.

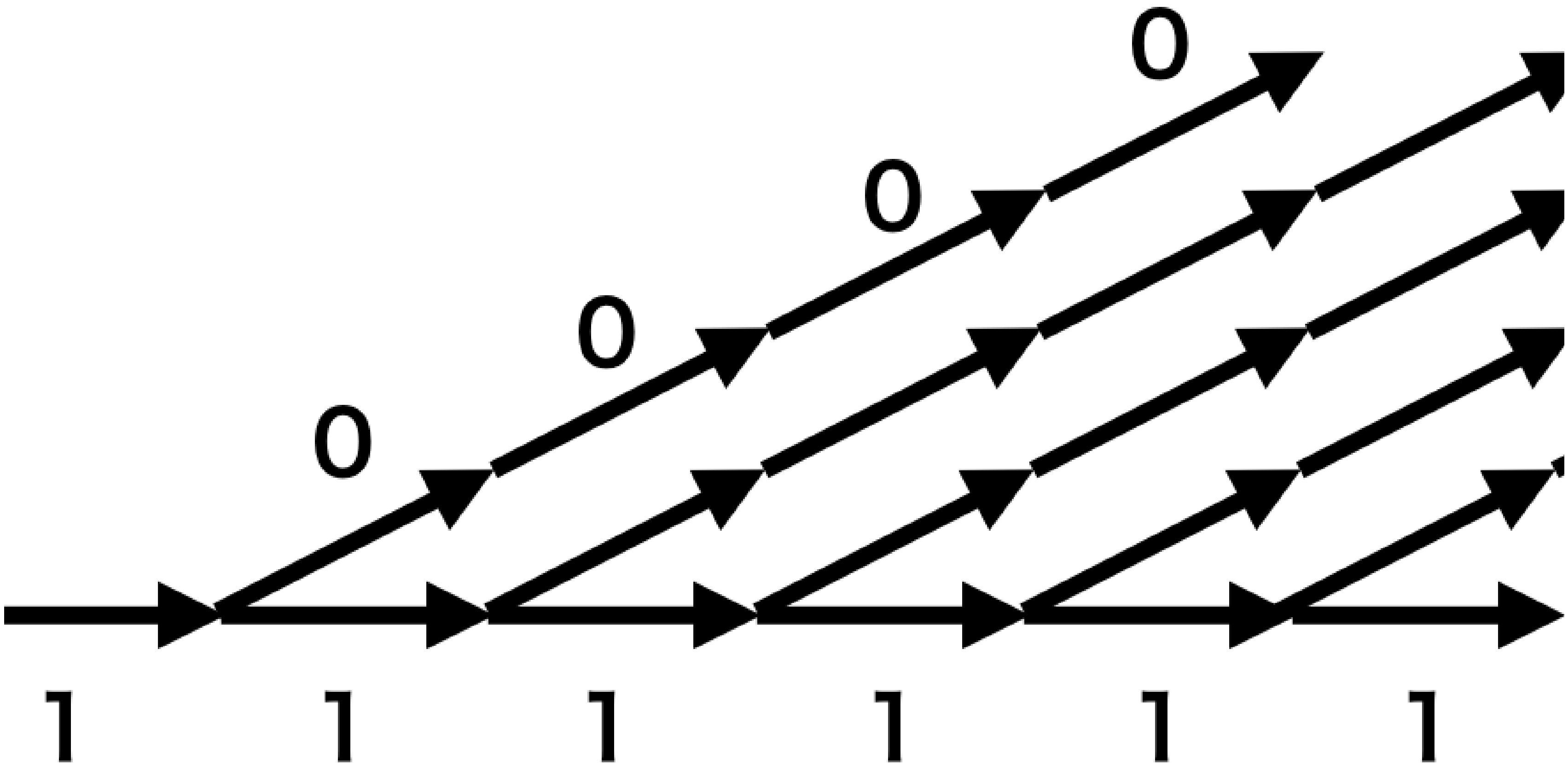
Non-convergence

For simplicity, consider $A = 1^\omega$.
Every general prediction should cover the case

$$A_n = 1^n 0^\omega$$

for all n.
For simple n, the weight of A_n should be large.
This prevents a general prediction bets 1 too much.

Picture



Summary

Summary

(i) We propose a framework to give the correct convergence speed to the correct measure.

(ii) It is based on Solomonoff's framework.

(iii) We propose the definition of generality inspired by Solomonoff's result.

(iv) We use computable randomness rather than Martin-Löf randomness.

(v) The correct probability is determined only up to multiplicative constant in the limit.

(vi) Sufficiently good approximation by functions computable in polynomial time.

Future work

(i) The underlying space is too restricted. Can it be generalized?
(ii) Compare with the existing framework.
(iii) Study the computational and/or descriptive complexity more.

End

Thank you for listening.

MEIJI UNIVERSITY