RATE OF CONVERGENCE OF COMPUTABLE PREDICTIONS

KENSHI MIYABE

Abstract. We consider the problem of predicting the next bit in an infinite binary sequence sampled from the Cantor space with an unknown computable measure. We propose a new theoretical framework to investigate the properties of good computable predictions, focusing on such predictions' convergence rate.

Since no computable prediction can be the best, we first define a better prediction as one that dominates the other measure. We then prove that this is equivalent to the condition that the sum of the KL divergence errors of its predictions is smaller than that of the other prediction for more computable measures. We call that such a computable prediction is more general than the other.

We further show that the sum of any sufficiently general prediction errors is a finite left-c.e. Martin-Löf random real. This means the errors converge to zero more slowly than any computable function.

§1. Introduction. Machine learning has recently become one of the hottest topics, with many real-world applications transforming society. Since the Dartmouth Conference in 1956, there have been efforts to develop a deeper theoretical understanding of learning. Several frameworks, such as PAC learning and Gold-style limit learning, have been proposed to define learning, explain it, and explore its capabilities and limits.

This article explores the theoretical limits of learning based on Solomonoff's universal induction or algorithmic probability theory.

We consider the following problem. We predict the next bit in an infinite binary sequence. We know the infinite binary sequence is sampled from the Cantor space with an unknown computable probability measure.

In the standard setting of the theory of universal induction, the measure used for prediction is c.e., that is, it is computably approximable from below but not computable in general. The reason for considering this broader class of measures than that of computable measures is that there exists an optimal prediction for c.e., while no computable prediction is optimal. The theory of universal induction concerns the properties of optimal predictions. This theory is elegant from a theoretical standpoint and has succeeded in deepening our understanding of learning. However, optimal predictions cannot be implemented directly in a computer, and its claims about machine learning algorithms used in practice are quite limited.

Received May 20, 2025.

2020 *Mathematics Subject Classification.* Primary 03D32, 68Q30.

Key words and phrases. inductive inference, general computable prediction, KL divergence, Solovay reducibility.

0022-4812/00/0000-0000
DOI:10.1017/jsl.2025.10155

Even though there is no optimal computable prediction, can we prove any sufficiently good one that approximates the optimal one has specific properties? This article gives a positive answer to this question by introducing the concept of generality.

We call a measure more general than another measure if it dominates the other. We then prove the prediction induced from a more general measure performs well for sample points of more computable measures. In other words, a more general prediction can solve more tasks. More precisely, the prediction induced from a more general measure has smaller error sums when measured by KL divergence (Theorem 3.2).

Furthermore, if we fix a computable measure to take samples, the error sum of sufficiently general predictions is always a finite Martin-Löf random real (Theorem 4.1). This means the errors converge to zero more slowly than any monotone computable function. A sufficiently general prediction cannot converge quickly, and its convergence rate is uniquely determined up to a multiplicative constant (Theorem 4.2). While simple intuition suggests that good predictions should have small errors, general-purpose algorithms that can solve many tasks will converge slower than specialized algorithms.

As special cases, we analyse the convergence speed using the L^p-norm when the model measure μ is either a Dirac measure (Proposition 4.9) or a separated measure (Proposition 4.16).

This article is a sequel to [17]. While the notion of generality has already been defined in [17], we consider this notion more carefully in this article. In particular, we give a necessary and sufficient condition of domination in Theorem 3.2. Theorem 4.1 strengthens [17, Theorem 3.1] and Proposition 4.13 strengthens [17, Theorems 4.3 and 4.4].

§2. Preliminaries.

In this section, we fix the notation and review notions from some theories.

2.1. Notations.

The sets of all positive integers, rational numbers, and reals are denoted by $\mathbb{N} = \{1, 2, 3, \dots\}$, $\mathbb{Q}$, and $\mathbb{R}$, respectively.

The set of all finite binary strings is denoted by $\{0,1\}^*$. We denote finite binary strings using σ and τ. The length of a string σ is denoted by $|\sigma|$. For $\sigma, \tau \in \{0,1\}^*$, the concatenation of σ and τ is denoted by $\sigma\tau$.

The set of all infinite binary sequences is denoted by $\{0,1\}^{\mathbb{N}}$. We use X, Y, Z to denote infinite binary sequences. We write $X = X_1X_2X_3\dots$ and let $X_{<n} = X_1X_2\dots X_{n-1}$ and $X_{\leq n} = X_1X_2\dots X_n$ for $n \in \mathbb{N}$.

The Cantor space, also denoted by $\{0,1\}^{\mathbb{N}}$, is the space of all infinite binary sequences equipped with the topology generated from the cylinder sets $[\sigma] = \{X \in \{0,1\}^{\mathbb{N}} : \sigma \prec X\}$ for $\sigma \in \{0,1\}^*$ where $\prec$ is the prefix relation.

2.2. Computability theory.

We follow the standard notation and terminology in computability theory and computable analysis. For details, see, for instance, [6, 23, 28].

A partial function $f :\subseteq \{0,1\}^* \to \{0,1\}^*$ is a *partial computable function* if it can be computed using a Turing machine. A real $x \in \mathbb{R}$ is called *computable* if there

exists a computable sequence $(q_n)_{n\in\mathbb{N}}$ of rationals such that $|x - q_n| < 2^{-n}$ for all n. A real $x \in \mathbb{R}$ is called *left-c.e.* if there exists an increasing computable sequence $(q_n)_n$ converging to x. A real $x \in \mathbb{R}$ is called *right-c.e.* if $-x$ is left-c.e.

A function $f : \{0,1\}^* \to \mathbb{R}$ is called *computable* if $f(\sigma)$ is uniformly computable in $\sigma \in \{0,1\}^*$. A (probabilistic) measure μ on $\{0,1\}^{\mathbb{N}}$ is *computable* if the function $\sigma \mapsto \mu([\sigma]) =: \mu(\sigma)$ is computable. For details on computable measure theory, see, for instance, [3, 27, 29].

2.3. Theory of inductive inference. Now, we review the theory of inductive inference initiated by Solomonoff. The primary references for this are [13, 15]. For a more philosophical discussion, see [20].

We use μ to denote a computable measure on the Cantor space $\{0,1\}^{\mathbb{N}}$. This μ represents an unknown model. We call this measure μ a *model measure*.

Suppose an infinite binary sequence is sampled from the Cantor space with this μ. When given the first $n-1$ bits $X_{<n}$ of X, the next bit follows the *conditional model measure* on $\{0,1\}$ represented by

$$k \mapsto \mu(k|X_{<n}) = \frac{\mu(X_{<n}k)}{\mu(X_{<n})}. \tag{1}$$

Our ultimate goal is to construct a computable measure ξ such that the prediction $\xi(\cdot|X_{<n})$ is close to $\mu(\cdot|X_{<n})$. We call this measure ξ a *prediction measure* and call the measure $\xi(\cdot|\cdot)$ a *conditional prediction*.

Solomonoff's celebrated result states that every optimal prediction behaves rather well. A *semi-measure* is a function $\xi : \{0,1\}^* \to [0,1]$ such that $\xi(\varepsilon) \le 1$ and $\xi(\sigma) \ge \xi(\sigma 0) + \xi(\sigma 1)$ for every $\sigma \in \{0,1\}^*$ where ε is the empty string. A function $f : \{0,1\}^* \to \mathbb{R}$ is called *c.e.* or *lower semi-computable* if $f(\sigma)$ is left-c.e. uniformly in $\sigma \in \{0,1\}^*$.

Let μ, ξ be semi-measures on $\{0,1\}^{\mathbb{N}}$. We say that ξ (multiplicatively) *dominates* μ if, there exists $c \in \mathbb{N}$ such that $\mu(\sigma) \le c \cdot \xi(\sigma)$ for all $\sigma \in \{0,1\}^*$. A c.e. semi-measure ξ is called *optimal* if ξ dominates every c.e. semi-measure. An optimal c.e. semi-measure exists while no computable measure is optimal. The conditional prediction $\xi(\cdot|\cdot)$ induced by this optimal c.e. semi-measure is sometimes called *algorithmic probability*.

THEOREM 2.1 [24], see also [13, Theorem 3.19]. *Let μ be a computable measure on $\{0,1\}^{\mathbb{N}}$. Let ξ be an optimal c.e. semi-measure. Then, for both $k \in \{0,1\}$ we have*

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

as $n \to \infty$ almost surely when X follows μ.

The prediction semi-measure ξ is arbitrary and lacks information about the model measures μ. The prediction by ξ investigates $X_{<n}$, which contains some information of μ, and predicts the next bit $X(n)$. The theorem above states that the conditional predictions $\xi(\cdot|X_{<n})$ are getting close to the true conditional model measures $\mu(\cdot|X_{<n})$ almost surely.

The rate of the convergence has been briefly discussed in [14] but has yet to be established.

§3. Generality. In this section, we introduce the concept of generality. Generality is a tool for comparing the well-behavedness of two measures. Just as optimality is defined by domination, generality is defined by domination. We expect that when one measure dominates another measure, the induced prediction also behaves better than the other. The question here is: what does it mean for one prediction to behave better than another? We answer this question by considering the sum of the prediction errors.

3.1. Definition of generality. Let ν, ξ be two measures on $\{0,1\}^{\mathbb{N}}$. We say that ξ is *more general* than ν if ξ dominates ν; that is, there exists $c \in \mathbb{N}$ such that $\nu(\sigma) \leq c \cdot \xi(\sigma)$ for all $\sigma \in \{0,1\}^*$.

The intuition is as follows. We are sequentially given a sequence $X \in \{0,1\}^{\mathbb{N}}$. The sequence $X \in \{0,1\}^{\mathbb{N}}$ may be a binary expansion of e or π, or a random sequence of $P(X_n = 0) = P(X_n = 1) = \frac{1}{2}$ independently. The task is to find such regularity and make a good prediction. The regularity is expressed as (or identified with) the measure μ such that X is random with respect to μ. The measure is a Dirac computable measure in the deterministic case, such as e or π. In general, the measure need not be deterministic; it can be an arbitrary computable measure.

Essentially, a prediction ξ is more general than another prediction ν if the prediction ξ behaves well for μ such that ν behaves well for μ. Thus, ξ performs better for a larger class of μ than ν. As we will see in Theorem 3.2, this relation is formalized by domination. This is the reason for using the terminology 'general' for domination.

We are interested in the property of sufficiently general computable predictions. We often say that a property P holds for all sufficiently large natural numbers if there exists N such that $P(n)$ holds for all natural numbers $n \geq N$. As an analogy, we say that a property P *holds for all sufficiently general computable prediction measures* if there exists a computable prediction measure ν such that the property $P(\xi)$ holds for all computable prediction measure ξ dominating ν. The author came up with the idea inspired by the study of Solovay functions, such as [2]. In particular, the computational complexity of computing such functions may be very low [12, Theorem 2].

In the inductive inference theory, we discuss the properties of an optimal c.e. semi-measure and its induced prediction. Similarly, we will see some properties of a sufficiently general computable measure and its induced prediction.

3.2. Domination and convergence. We claim that domination means better behavior by giving a necessary and sufficient condition for the convergence of the sum of the prediction errors. Here, the error is measured by Kullback–Leibler divergence.

The Kullback–Liebler divergence is the primary tool for discussing the convergence of the predictions. For details, see any standard text on information theory, such as [7].

Let μ, ξ be measures on the discrete space $\{0,1\}$. The KL divergence of μ with respect to ξ is defined by

$$d(\mu||\xi) = \sum_{k \in \{0,1\}} \mu(k) \ln \frac{\mu(k)}{\xi(k)},$$

where $0 \cdot \log \frac{0}{z} = 0$ for $z \geq 0$, $y \log \frac{y}{0} = \infty$ for $y > 0$, and ln is the natural logarithm.

Next, let μ, ξ be measures on the continuous space $\{0,1\}^{\mathbb{N}}$. We use the notation:

- $d_\sigma(\mu||\xi) = d(\mu(\cdot|\sigma)||\xi(\cdot|\sigma))$,
- $D_n(\mu||\xi) = \sum_{k=1}^{n} E_\mu[d_{X_{<k}}(\mu||\xi)]$,
- $D_\infty(\mu||\xi) = \lim_{n\to\infty} D_n(\mu||\xi)$,

where $\mu(\cdot|\sigma), \xi(\cdot|\sigma)$ are the measures on $\{0,1\}$ defined in (1). Thus, $d_\sigma(\mu||\xi)$ is the prediction error conditioning on σ, $D_n(\mu||\xi)$ is the expected sum of the prediction errors until the nth round when X follows μ, and D_∞ is its limit. Since KL divergence is non-negative, D_n is non-decreasing in n. Note that the finiteness of the sum of the prediction errors is a condition stronger than the convergence of the errors to 0.

REMARK 3.1. The chain rule for KL divergence states that

$$D_n(\mu||\xi) = E_\mu[\ln \frac{\mu(X_{\leq n})}{\xi(X_{\leq n})}].$$

See such as Hutter [13, (3.18)] and Cover and Thomas [7, Theorem 2.5.3].

THEOREM 3.2. *For two measures ξ, ν on $\{0,1\}^{\mathbb{N}}$, the following are equivalent.*

(i) *ξ dominates ν.*
(ii) *There exists a constant $c \in \mathbb{N}$ such that for every measure μ on $\{0,1\}^{\mathbb{N}}$, we have $D_\infty(\mu||\xi) \leq D_\infty(\mu||\nu) + c$.*

From this, domination means rapid convergence of a larger class of model measures. If ξ dominates ν and ν behaves well for μ (the error sum is finite), then ξ also behaves well for μ (the error sum is finite). Furthermore, the difference of the sums of the errors is, at most, a constant uniformly in μ. Thus, the error sum of ν is small, so is that of ξ.

Note that KL divergence can be infinity, and the finiteness of KL divergence is an essential aspect in the formulation of Theorem 3.2. Some other distances are discussed in [13, Section 3.2.5]. One example is the Hellinger distance, which plays a vital role in the proof of Theorem 2.1, but is bounded by 1. Thus, KL divergence seems helpful in the formulation.

PROOF. (i)⇒(ii). Suppose that

$$\nu \leq c\,\xi \tag{2}$$

for some $c \in \mathbb{N}$.

Suppose that there exists a string $\sigma \in \{0,1\}^*$ such that $\mu(\sigma) > 0$ and $\nu(\sigma) = 0$. Then, there exist a string $\tau \in \{0,1\}^*$ and a bit $k \in \{0,1\}$ such that $\mu(\tau 0) > 0$, $\mu(\tau 1) > 0$, $\nu(\tau) > 0$ and $\nu(\tau k) = 0$. For this τ, we have $d_\tau(\mu||\nu) = \infty$ and $D_\infty(\mu||\nu) = \infty$. Thus, the condition (ii) holds.

Now assume that

$$\mu(\sigma) > 0 \Rightarrow \nu(\sigma) > 0 \tag{3}$$

for all $\sigma \in \{0,1\}^*$. Fix an arbitrary $n \in \mathbb{N}$. For all $\sigma \in \{0,1\}^n$ such that $\mu(\sigma) > 0$, we have

$$\ln \frac{\mu(\sigma)}{\xi(\sigma)} \leq \ln \frac{\mu(\sigma)}{\nu(\sigma)} + \ln c \tag{4}$$

by (2). Here note that $\xi(\sigma) > 0$ by (3) and (2). By taking the integral of (4) with respect to μ, we have

$$D_n(\mu||\xi) \leq D_n(\mu||\nu) + \ln c$$

by Remark 3.1. Since both D_n are non-decreasing, this implies the condition (ii).

(ii)⇒(i). Let $\sigma \in \{0,1\}^*$ be an arbitrary string. We construct a measure μ such that the condition (ii) for this μ implies $\nu(\sigma) \leq e^c \xi(\sigma)$. We define the measure μ by

$$\mu(\tau) = \begin{cases} \nu(\tau)/\nu(\sigma), & \text{if } \sigma \preceq \tau, \\ 1, & \text{if } \tau \preceq \sigma, \\ 0, & \text{otherwise.} \end{cases}$$

In other words, μ is zero outside the cylinder $[\sigma]$ and is proportional to ν inside $[\sigma]$. Note that for any string $\rho \in \{0,1\}^*$ such that $|\rho| = |\sigma|$, the ratio $\mu(\rho\tau)/\nu(\rho\tau)$ is constant for all $\tau \in \{0,1\}^*$. Thus, $D_{|\sigma|}(\mu||\nu) = D_\infty(\mu||\nu)$. Hence,

$$c \geq D_\infty(\mu||\xi) - D_\infty(\mu||\nu) \geq D_{|\sigma|}(\mu||\xi) - D_{|\sigma|}(\mu||\nu) = \ln \frac{\mu(\sigma)}{\xi(\sigma)} - \ln \frac{\mu(\sigma)}{\nu(\sigma)},$$

where the last equality follows by Remark 3.1. Hence we have $\nu(\sigma) \leq e^c \xi(\sigma)$.

Since σ is arbitrary, the condition (i) holds. ⊣

3.3. Infinite chain rule for KL divergence. Here, with independent interest, we show that $D_\infty(\mu||\xi)$ is nothing but the usual KL divergence.

Let us recall the KL divergence on a non-discrete space. Let μ, ξ be measures on $\{0,1\}^{\mathbb{N}}$. Then, the KL divergence of μ with respect to ξ is defined by

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

where $0 \cdot \log 0 = 0$ and ln is the natural logarithm, and $\frac{d\mu}{d\xi}$ is the Radon–Nikodym derivative of μ with respect to ξ. If μ is the derivative $\frac{d\mu}{d\xi}$ does not exist, then let $D(\mu||\xi) = \infty$.

PROPOSITION 3.3. *Let ξ, μ be measures on $\{0,1\}^{\mathbb{N}}$. Then,*

$$D_\infty(\mu||\xi) = D(\mu||\xi).$$

This is an infinite version of the chain rule for KL divergence in Remark 3.1. The essential reason for this is that the Radon–Nikodym derivative $\frac{d\mu}{d\xi}$ can be approximated by $\frac{\mu(X_{\leq n})}{\xi(X_{\leq n})}$. For proof, we use the following facts.

LEMMA 3.4 (Theorem 5.3.3 in [11] in our terminology). *Suppose that $\xi(\sigma) = 0 \Rightarrow \mu(\sigma) = 0$ for all $\sigma \in \{0,1\}^*$. Let $f(X) = \limsup_n \frac{\mu(X_{\leq n})}{\xi(X_{\leq n})}$. Then,*

$$\mu(A) = \int_A f \, d\xi + \mu(A \cap \{f(X) = \infty\})$$

for all measurable sets A.

REMARK 3.5.

(i) The sequence $(\frac{\mu(X_{\leq n})}{\xi(X_{\leq n})})_n$ is a non-negative martingale with respect to ξ (see [11, Theorem 5.3.4]).
(ii) Hence, $\xi(\{f(X) = \infty\}) = 0$ by Doob's martingale maximal inequality.
(iii) If $\mu \ll \xi$, then $f = \lim_n \frac{\mu(X_{\leq n})}{\xi(X_{\leq n})} = \frac{d\mu}{d\xi}$, ξ-almost surely.

Proposition 3.3. *We divide the proof into four cases.*

Case 1. $\frac{d\mu}{d\xi}$ *exists and* $D(\mu||\xi) < \infty$.

We will show that $(\frac{\mu(X_{\leq n})}{\xi(X_{\leq n})} \ln \frac{\mu(X_{\leq n})}{\xi(X_{\leq n})})_n$ *is uniformly integrable with respect to* ξ. *For* $K \in \mathbb{N}$, *let*

$$U_n^K = \{X \in \{0,1\}^{\mathbb{N}} \ : \ \frac{\mu(X_{\leq n})}{\xi(X_{\leq n})} > K\}.$$

It suffices to show that

$$\sup_n \int_{U_n^K} \left| \frac{\mu(X_{\leq n})}{\xi(X_{\leq n})} \ln \frac{\mu(X_{\leq n})}{\xi(X_{\leq n})} \right| d\xi \to 0 \quad \text{as } K \to \infty.$$

Let $A_n^K = \{\sigma \in \{0,1\}^n \ : \ \mu(\sigma)/\xi(\sigma) > K\}$. *For* $K > 1$, *we have* $\ln(\mu(\sigma)/\xi(\sigma)) > \ln K > 0$. *Thus,*

$$\begin{aligned} \int_{U_n^K} \left| \frac{\mu(X_{\leq n})}{\xi(X_{\leq n})} \ln \frac{\mu(X_{\leq n})}{\xi(X_{\leq n})} \right| d\xi &= \sum_{\sigma \in A_n^K} \xi(\sigma) \frac{\mu(\sigma)}{\xi(\sigma)} \ln \frac{\mu(\sigma)}{\xi(\sigma)} \\ &\leq \sum_{\sigma \in A_n^K} \int_{[\sigma]} \frac{d\mu}{d\xi} \ln \frac{d\mu}{d\xi} d\xi = \int_{U_n^K} \frac{d\mu}{d\xi} \ln \frac{d\mu}{d\xi} d\xi \end{aligned} \tag{5}$$

Here, we used Jensen's inequality on $[\sigma]$ *with the convex function* $g(x) = x \ln x$:

$$g\left(\frac{1}{\xi(\sigma)} \int_{[\sigma]} \frac{d\mu}{d\xi} d\xi\right) \leq \frac{1}{\xi(\sigma)} \int_{[\sigma]} g\left(\frac{d\mu}{d\xi}\right) d\xi. \tag{6}$$

Since $\mu(X_{\leq n})/\xi(X_{\leq n})$ *is a non-negative martingale by Remark 3.5, we have* $\mu(U_n^K) < \frac{1}{K}$. *From the epsilon-delta type characterization of absolute continuity* (*see* [18, *Proposition* 15.5] *for a general measure space and* [5, *Theorem* 2.5.7] *for the Lebesgue integral*), *the supremum of the last term in* (5) *goes to* 0 *as* $K \to \infty$. *This shows uniform integrability.*

Finally, we use the Vitali convergence theorem to deduce

$$D_\infty(\mu||\xi) = \lim_n E\left[\frac{\mu(X_{\leq n})}{\xi(X_{\leq n})} \ln \frac{\mu(X_{\leq n})}{\xi(X_{\leq n})}\right] = E\left[\frac{d\mu}{d\xi} \ln \frac{d\mu}{d\xi}\right] = D(\mu||\xi)$$

by Remark 3.5(*iii*).

Case 2. $\frac{d\mu}{d\xi}$ *exists and* $D(\mu||\xi) = \infty$.

Then, $D_\infty(\mu||\xi) = \infty$ because, by the finite chain rule for KL divergence, we have

$$D_\infty(\mu||\xi) = \lim_n E_\mu[\ln \frac{\mu(X_{\leq n})}{\xi(X_{\leq n})}] \geq E_\mu[\ln \frac{d\mu}{d\xi}] = D(\mu||\xi),$$

where we have used Fatou's lemma in deducing the inequality.

Case 3. $\frac{d\mu}{d\xi}$ does not exist and $\xi(\sigma) = 0 \Rightarrow \mu(\sigma) = 0$ for all $\sigma \in \{0,1\}^*$.

By Lemma 3.4, $\mu(\{f(X) = \infty\}) = \varepsilon > 0$. Then, for each $K > 0$, we have $\mu(\{\lim_n \frac{\mu(X_{\leq n})}{\xi(X_{\leq n})} > K\}) \geq \varepsilon$, and thus, there exists $n \in \mathbb{N}$ such that $\mu(\{\frac{\mu(X_{\leq n})}{\xi(X_{\leq n})} > K\}) > \varepsilon/2$, which implies $D_n(\mu||\xi) \geq \frac{\varepsilon \ln K}{2}$. Since K is arbitrary, we have $D_\infty(\mu||\xi) = \infty$.

Case 4. $\xi(\sigma) = 0$ and $\mu(\sigma) > 0$ for some $\sigma \in \{0,1\}^*$.

In this case, we have $D_{|\sigma|}(\mu||\xi) \geq \mu(\sigma) \ln \frac{\mu(\sigma)}{\xi(\sigma)} = \infty$. Thus, $D_\infty(\mu||\xi) = \infty$. Since $\mu \not\ll \xi$, we also have $D(\mu||\xi) = \infty$.

§4. Rate of convergence. Let μ be a computable model measure on $\{0,1\}^{\mathbb{N}}$. Then, for any computable measure ξ that dominates μ, we have $D_\infty(\mu||\xi) < \infty$ by Theorem 3.2. Hence, any sufficiently general prediction converges to the conditional model measure, almost surely. In this section, we discuss its rate of convergence. The main result here is Martin-Löf randomness of the KL divergence, from which we show that the convergence rate is almost the same for any sufficiently general prediction.

4.1. Martin-Löf randomness of KL divergence. We review Martin-Löf random left-c.e. reals to analyze the convergence rate. For details, see such as [9, Chapter 9].

A set $U \subseteq \mathbb{R}$ is a *c.e. open set* if there exists a computable sequence $(a_n, b_n)_{n\in\mathbb{N}}$ of open intervals with rational endpoints such that $U = \bigcup_n (a_n, b_n)$. Let λ be the Lebesgue measure on $\mathbb{R}$. A *ML-test* with respect to λ is a sequence $(U_n)_n$ of uniformly c.e. open sets with $\lambda(U_n) \leq 2^{-n}$ for all $n \in \mathbb{N}$. A real $\alpha \in \mathbb{R}$ is called *ML-random* if $\alpha \notin \bigcap_n U_n$ for every ML-test $(U_n)_n$.

An example of left-c.e. ML-random reals is the halting probability. The *halting probability* Ω_U of a prefix-free Turing machine U is defined by $\Omega_U = \sum_{\sigma\in \mathrm{dom(U)}} 2^{-|\sigma|}$. Then, Ω_U is a left-c.e. ML-random real for each universal prefix-free Turing machine U. This Ω_U is known as *Chaitin's omega*. Conversely, any left-c.e. ML-random real in $(0,1)$ is the halting probability of some universal machine (see [9, Theorems 9.2.2 and 9.2.3]).

THEOREM 4.1. *Let μ be a computable model measure on $\{0,1\}^{\mathbb{N}}$. Then, $D_\infty(\mu||\xi)$ is a finite left-c.e. ML-random real for all sufficiently general computable measures ξ.*

We can discuss the convergence rate from this Martin-Löf randomness. This is because all ML-random reals have almost the same rate of convergence, as follows:

THEOREM 4.2 [1], see also [16]. *Let α, β be left-c.e. reals with their increasing computable approximations $(\alpha_s), (\beta_s)$. If β is ML-random, then*

$$\lim_{s\to\infty} \frac{\alpha - \alpha_s}{\beta - \beta_s} \text{ exists}$$

and is independent from the approximation. Furthermore, the limit is zero if and only if α is not ML-random.

This theorem means that the convergence rate of ML-random left-c.e. reals is the same up to a multiplicative constant and much slower than that of non-ML-random left-c.e. reals.

Now we give a proof of Theorem 4.1. First we construct a computable measure ν such that $D_\infty(\mu||\nu)$ is ML-random. Then, we claim that if a computable measure ξ dominates ν, then $D_\infty(\mu||\xi) - D_\infty(\mu||\nu)$ is a left-c.e. real, which implies ML-randomness of $D_\infty(\mu||\xi)$ by a result of Solovay reducibility.

LEMMA 4.3. *Let μ be a computable measure. Then, there exists a computable measure ν such that*:

- *the Radon–Nikodym derivative $\frac{d\mu}{d\nu}$ exists,*
- *$\frac{d\mu}{d\nu}$ is a constant function on a μ-measure 1 set and 0 outside it,*
- *the constant value is a finite left-c.e. ML-random real.*

In particular, $D_\infty(\mu||\nu)$ is a finite left-c.e. ML-random real.

PROOF. First, we define the computable measure ν. Let $(z_n)_{n\in\mathbb{N}}$ be a sequence of uniformly computable positive reals such that $s = \sum_{n\in\mathbb{N}} z_n < 1$ is a ML-random real. Let $Z^\sigma \in \{0,1\}^{\mathbb{N}}$ be a computable sequence uniformly in σ such that $\sigma \prec Z^\sigma$ and $\mu(Z^\sigma) = 0$, whose existence will be shown in Lemma 4.4.

Define measures μ_n, ν by

$$\mu_n(\sigma) = \begin{cases} \mu(\sigma), & \text{if } |\sigma| \le n, \\ \mu(\tau), & \text{if } |\sigma| > n,\ \tau = \sigma_{\le n},\ \sigma \prec Z^\tau, \\ 0, & \text{if } |\sigma| > n,\ \tau = \sigma_{\le n},\ \sigma \not\prec Z^\tau, \end{cases}$$

for all $\sigma \in \{0,1\}^*$ and

$$\nu = \sum_n z_n \mu_n + (1-s)\mu. \tag{7}$$

The measure μ_n coincides with μ up to depth n, but beyond that point it collapses the distribution onto a single predetermined infinite path Z^τ extending each prefix τ of length n; in other words, all of the mass that μ assigns to τ is concentrated along one chosen branch, and every other continuation gets zero. The measure ν mixes the collapsed measures μ_n with weights z_n together with a portion of the original measure μ, so it combines μ with versions that eventually follow a single deterministic path.

Now, we claim that the measure ν is computable. This is because

$$\begin{aligned} \nu(\sigma) &= \sum_{n<|\sigma|} z_n \mu_n(\sigma) + \sum_{n\ge|\sigma|} z_n \mu_n(\sigma) + (1-s)\mu(\sigma) \\ &= \sum_{n<|\sigma|} z_n \mu_n(\sigma) + (1 - \sum_{n<|\sigma|} z_n)\mu(\sigma). \end{aligned}$$

Next we find $\frac{d\mu}{d\nu}$. Because $\mu \ll \nu$, by Remark 3.5(iii), $\frac{d\mu}{d\nu} = \lim_n \frac{\mu(X_{\le n})}{\nu(X_{\le n})}$ ν-almost surely.

Consider $X \in \{0,1\}^{\mathbb{N}}$ such that $\mu(X_{\leq n}) > 0$ for all n. Then, μ-almost such sequences satisfy $X \neq Z^\sigma$ for any $\sigma \in \{0,1\}^*$. For each n and sufficiently large k depending on n, we have $\mu_n(X_{\leq k}) = 0$. Thus, $\lim_k \frac{\mu(X_{\leq k})}{\nu(X_{\leq k})} = \frac{1}{1-s}$.

If $X = Z^\sigma$ for some $\sigma \in \{0,1\}^*$, then

$$\mu(X_{\leq n}) \to \mu(X) = \mu(Z^\sigma) = 0,$$
$$\nu(X_{\leq n}) \to \nu(X) = \sum\{z_n \mu_n(\sigma) \ : \ Z^\sigma = X\} > 0,$$

as $n \to \infty$. Hence, $\lim_k \frac{\mu(X_{\leq k})}{\nu(X_{\leq k})} = 0$.

We also observe that the set of X such that $\mu(X_{\leq n}) = 0$ for some n has μ-measure 0. Because s is a left-c.e. ML-random, so is $\frac{1}{1-s}$. Hence, the first half of the claim follows.

Finally,

$$D(\mu||\nu) = \int \ln \frac{d\mu}{d\nu} d\mu = \ln \frac{1}{1-s},$$

which is ML-random by Proposition 4.5. ⊣

LEMMA 4.4. *For each $\sigma \in \{0,1\}^*$, we can compute a sequence $Z^\sigma \in \{0,1\}^{\mathbb{N}}$ such that $\sigma \prec Z^\sigma$ and $\mu(Z^\sigma) = 0$. Furthermore, the construction is uniform in σ.*

We construct Z^σ as the limit of extending sequences $\sigma = \tau_0 \prec \tau_1 \prec \tau_2 \dots$.

One might attempt to define τ_{k+1} from τ_k with the following properties:

- $\tau_k \prec \tau_{k+1}$,
- $|\tau_{k+1}| = |\tau_k| + 1$,
- $\mu(\tau_{k+1}) < \frac{2}{3} \cdot \mu(\tau_k)$.

Roughly saying, one computes the conditional probability and takes the smaller one.

However, this simple idea does not work. Since $\mu(\sigma)$ may be 0 for some $\sigma \in \{0,1\}^*$, the conditional probability may not be computable.

To make the construction uniform, we need the following modified strategy to construct it.

PROOF. Let $p, q \in (0,1)$ be rational numbers such that

$$0 < p < q < 1, \quad pq > \frac{1}{2},$$

for example, $p = \frac{3}{4}$ and $q = \frac{4}{5}$.

Let $\tau_0 = \sigma$.

Suppose τ_k is already defined and satisfies

$$\mu(\tau_k) \leq q^k \max\left\{\mu(\sigma), p^k\right\}. \tag{8}$$

Notice that (8) holds for $k = 0$.

Now we define τ_{k+1} so that $\tau_k \prec \tau_{k+1}$, $|\tau_{k+1}| = |\tau_k| + 1$,

$$\mu(\tau_{k+1}) < q^{k+1} \max\left\{\mu(\sigma), p^{k+1}\right\}. \tag{9}$$

We claim that τ_{k+1} computationally can be found. If neither of the strings extending τ_k satisfies (9), then

$$\mu(\tau_k) \geq 2q^{k+1} \max\left\{\mu(\sigma), p^{k+1}\right\} > q^k \max\left\{\mu(\sigma), p^k\right\},$$

which contradicts (8). Hence, one of the two strings extending τ_k satisfies (9), which can be found computably.

Finally, the claim follows by letting k tend to infinity in (8). ⊣

PROPOSITION 4.5. *Let I be an open interval in the real line and $f : I \to \mathbb{R}$ be a computable function in C^1. If $z \in I$ is ML-random and $f'(z) \neq 0$, then $f(z)$ is ML-random. Here f' is the derivative of f.*

This fact follows from the more advanced fact called randomness preservation or conservation of randomness [4, Theorem 3.2]. However, we give a direct proof here.

PROOF. Without loss of generality, we can assume $f'(z) > 0$. Because f' is continuous, there exists a closed interval $[a, b]$ with rational endpoints such that $z \in [a, b] \subseteq I$ and $f'(x) > 0$ for every $x \in [a, b]$. Because f' is continuous and $[a, b]$ is a bounded closed set, by the extreme value theorem, we have a positive rational $m < \inf_{x\in[a,b]} f'(x)$.

Suppose $f(z)$ is not ML-random. Then there exists a ML-test $(U_n)_n$ such that $f(z) \in \bigcap_n U_n$. Let $V_n = \{x : f(x) \in U_n\} \cap [a, b]$. Then, $(V_n)_n$ is a sequence of uniformly c.e. open sets. We also have $z \in \bigcap_n V_n$ because $f(z) \in U_n$ for all n.

We claim that $\mu(V_n) \leq 2^{-n}/m$ for all n. When some interval $(c, d) \subseteq [f(a), f(b)]$ is enumerated into U_n, the corresponding interval $(f^{-1}(c), f^{-1}(d)) \subseteq [a, b]$ is enumerated into V_n. By the mean-value theorem, there exists $w \in (f^{-1}(c), f^{-1}(d))$ such that

$$(d - c) = f'(w)(f^{-1}(d) - f^{-1}(c)) \geq m(f^{-1}(d) - f^{-1}(c)).$$

Hence, the claim follows. ⊣

The last piece for the proof is the following result on Solovay reducibility. For a proof, see [9, Theorem 9.1.4] or [19, Proposition 3.2.27].

PROPOSITION 4.6. *The sum of a left-c.e. ML-random real and a left-c.e. real is ML-random.*

Theorem 4.1. *Let ν be the measure constructed in Lemma 4.3. Let ξ be a measure dominating ν. Then,*

$$D(\mu||\xi) = \int \ln \frac{d\mu}{d\xi} d\mu = \int \ln \frac{d\mu}{d\nu} d\mu + \int \frac{d\mu}{d\nu} \ln \frac{d\nu}{d\xi} d\nu = D(\mu||\nu) + \alpha D(\nu||\xi),$$

where α is the left-c.e. real such that $\frac{d\mu}{d\nu} = \alpha$ μ-a.s. Here, $D(\mu||\nu)$ is ML-random by Lemma 4.3 and $D(\mu||\nu)$ and $D(\nu||\xi)$ are left-c.e., as in Proposition 3.3. Thus, by Proposition 4.6, $D_\infty(\mu||\xi)$ is ML-random.

4.2. L^p-norm of measures. We begin by introducing distances between measures on the finite alphabet $\{0, 1\}$. These distances will later be applied to conditional distributions arising from measures on the infinite sequence space $\{0, 1\}^{\mathbb{N}}$.

Let μ, ξ be measures on the discrete space $\{0,1\}$. For $p \geq 1$, the distance between μ and ξ by the L^p-norm is

$$||\mu - \xi||_p = (\sum_{k\in\{0,1\}} |\mu(k) - \xi(k)|^p)^{1/p}.$$

Let

$$\ell_p(\mu, \xi) = ||\mu - \xi||_p^p.$$

Some closely related distances are:

- $\ell_1(\mu, \xi) = ||\mu - \xi||_1$ is the Manhattan distance.
- $\ell_2(\mu, \xi) = ||\mu - \xi||_2^2$ is the squared Euclidian distance.
- $\frac{1}{2}\ell_1(\mu, \xi) = \frac{1}{2}||\mu - \xi||_1$ is the total variation distance.

We now extend these notions to measures on the sequence space $\{0,1\}^{\mathbb{N}}$, in the same way as was previously done for the KL divergence. For measures μ, ξ on $\{0,1\}^{\mathbb{N}}$, we write:

- $\ell_{p,\sigma}(\mu, \xi) = \ell_p(\mu(\cdot|\sigma), \xi(\cdot|\sigma))$,
- $L_{p,n}(\mu, \xi) = \sum_{k=1}^{n} E_{X\sim\mu}[\ell_{p,X_{<k}}(\mu, \xi)]$,
- $L_{p,\infty}(\mu, \xi) = \lim_{n\to\infty} L_{p,n}(\mu, \xi)$.

If μ, ξ, and p are computable and $L_{p,\infty}(\mu, \xi)$ is finite, then $L_{p,\infty}(\mu, \xi)$ is left-c.e.

Let μ be a computable measure on $\{0,1\}^{\mathbb{N}}$. We ask at which p the left-c.e. reals $D_\infty(\mu, \xi)$ and $L_{p,\infty}(\mu, \xi)$ have the same rate of convergence, which mainly depends on μ.

In the theory of algorithmic randomness, Solovay reducibility measures the convergence rate of left-c.e. reals. Instead of the original definition by Solovay, we use the following characterization by Downey, Hirschfeldt, and Nies [10] (see also [9, Theorem 9.1.8]). For two left-c.e. reals α, β, we say that α is Solovay reducible to β, denoted by $\alpha \leq_S \beta$, if there exists a constant $c \in \mathbb{N}$ and a left-c.e. real γ such that $c\beta = \alpha + \gamma$. Roughly saying, $\alpha \leq_S \beta$ means that the convergence rate of β is not faster than α. The induced equivalence relation, denoted by $\equiv_S$, is defined by $\alpha \equiv_S \beta \iff (\alpha \leq_S \beta \text{ and } \beta \leq_S \alpha)$. If α is ML-random and $\alpha \leq_S \beta$, then β is ML-random by Proposition 4.6.

DEFINITION 4.7. We define $R(\mu)$ to be the set of positive computable reals p such that $L_{p,\infty}(\mu, \xi) < \infty$ and $D_\infty(\mu, \xi) \equiv_S L_{p,\infty}(\mu, \xi)$ for all computable measures ξ dominating μ.

In what follows, we determine $R(\mu)$ for Dirac measures μ and separated measures μ. If $R(\mu)$ is a single point set, we write $R(\mu) = p$ for $R(\mu) = \{p\}$.

The rough rate of convergence of left-c.e. reals can be represented by the effective Hausdorff dimension. Let K be the prefix-free Kolmogorov complexity, that is, $K(\sigma) = \min\{|\tau| : U(\tau) = \sigma\}$ where U is a fixed universal prefix-free Turing machine. The Levin–Schnorr theorem states that $X \in \{0,1\}^{\mathbb{N}}$ is ML-random if and only if $K(X \upharpoonright n) > n - O(1)$ where we identify a real in the unit interval with its binary expansion. The effective Hausdorff dimension of $X \in \{0,1\}^{\mathbb{N}}$ is defined by

$$\dim(X) = \liminf_n \frac{K(X \upharpoonright n)}{n}.$$

In particular, $\dim(X) = 1$ for each ML-random sequence X. See [9, Chapter 13] for details.

THEOREM 4.8 (Theorem 3.2 in [25]). *Let $(a_n)_n$ be a sequence of uniformly computable positive reals such that $\sum_n a_n$ is finite and is ML-random. Then, the following holds*:

(i) $\dim(\sum_n (a_n)^p) = 1/p$ *for each computable* $p \geq 1$.
(ii) $\sum_n (a_n)^p = \infty$ *for each* $p \in (0, 1)$.

The original statement by Tadaki is about the halting probability but the statement also holds for any sequence of uniformly computable positive reals whose sum is finite and ML-random by almost the same proof.

4.3. Case of Dirac measures. From now on, we discuss the rate of convergence more concretely. First, we consider the case in which the model measure μ is a Dirac measure, which means that the model is deterministic.

Let μ be a computable Dirac measure; that is, $\mu = \mathbf{1}_A$ for some $A \in \{0,1\}^{\mathbb{N}}$. Because A is an atom of the computable measure μ, the sequence A is computable (see, for example, [9, Lemma 6.12.7]). The goal is to evaluate the error of ξ

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

for each $n \in \mathbb{N}$ for general computable prediction measures ξ.

PROPOSITION 4.9. *Let $A \in \{0,1\}^{\mathbb{N}}$ be a computable sequence and $\mu = \mathbf{1}_A$. Then, $R(\mu) = 1$. In particular, $L_{1,\infty}(\mu, \xi)$ is finite and is a left-c.e. ML-random real for all sufficiently general computable prediction measures ξ.*

LEMMA 4.10. *Let $A \in \{0,1\}^{\mathbb{N}}$ be a computable sequence and $\mu = \mathbf{1}_A$. Let ξ be a computable measure dominating μ. Then,*

$$L_{1,\infty}(\mu, \xi) = 2 \sum_{n=1}^{\infty} (1 - \xi(A_n | A_{<n})).$$

PROOF. For each $\sigma \in \{0,1\}^*$, we have

$$\ell_{1,\sigma} = |\mu(0|\sigma) - \xi(0|\sigma)| + |\mu(1|\sigma) - \xi(1|\sigma)|.$$

Since $\mu = \mathbf{1}_A$, we have

$$E_{X \sim \mu}[\ell_{1,X_{<n}}(\mu, \xi)] = |\mu(0|A_{<n}) - \xi(0|A_{<n})| + |\mu(1|A_{<n}) - \xi(1|A_{<n})|$$

for each $n \in \mathbb{N}$. Since $\mu(A_n | A_{<n}) = 1$ and $\mu(\overline{A_n} | A_{<n}) = 0$ where $\overline{k} = 1 - k$, we have

$$L_{1,\infty}(\mu, \xi) = \sum_{n=1}^{\infty} E_{X \sim \mu}[\ell_{1,X_{<n}}(\mu, \xi)] = \sum_{n=1}^{\infty} (1 - \xi(A_n | A_{<n}) + \xi(\overline{A_n} | A_{<n})).$$

Finally, notice that $\xi(\overline{A_n} | A_{<n}) = 1 - \xi(A_n | A_{<n})$. Hence, the claim follows. ⊣

LEMMA 4.11. *Let $A \in \{0,1\}^{\mathbb{N}}$ be a computable sequence and $\mu = \mathbf{1}_A$. Then, $1 \in R(\mu)$.*

PROOF. Let ξ be a computable measure dominating μ.

First, we demonstrate that $L_{1,\infty}(\mu,\xi) < \infty$. By the inequality

$$\ln(1-x) \leq_- x$$

for all $x \in \mathbb{R}$, we have

$$1 - \xi(A_n|A_{<n}) \leq_- \ln \xi(A_n|A_{<n}) = d_{A_{<n}}(\mu||\xi). \tag{10}$$

From this and by Lemma 4.10, we have

$$L_{1,\infty}(\mu,\xi) \leq 2D_\infty(\mu||\xi) < \infty,$$

where the last inequality follows from Theorem 3.2.

Let $f(n)$ be a computable function from $\mathbb{N}$ to $\mathbb{R}$ such that

$$\ell_{1,A_{<n}}(\mu,\xi) + f(n) = 2d_{A_{<n}}(\mu||\xi).$$

Then, $f(n) \geq 0$ for all n by (10). Hence,

$$L_{1,\infty}(\mu,\xi) + \sum_n f(n) = 2D_\infty(\mu||\xi),$$

which implies $L_{1,\infty}(\mu,\xi) \leq_S D_\infty(\mu||\xi)$.

Next, we prove the converse relation. For sufficiently large n, we have

$$\ell_{1,A_{<n}}(\mu||\xi) > 2(\ln 2)(1-\xi(A_n|A_{<n})) \geq_- \ln \xi(A_n|A_{<n}) = d_{A_{<n}}(\mu||\xi),$$

where we used $0 < \ln 2 < 1$ for the first inequality and $\ln(1-x) \geq_- 2(\ln 2)x$ for all $x \in [0, 1/2]$ for the second inequality. Also note that, since $L_{1,\infty}(\mu,\xi) < \infty$ by above, we have $1 - \xi(A_n|A_{<n}) \to 0$ as $n \to \infty$. Thus, there exists a left-c.e. real α such that $L_{1,\infty}(\mu,\xi) = D_\infty(\mu||\xi) + \alpha$. Hence, $D_\infty(\mu||\xi) \leq_S L_{1,\infty}(\mu,\xi)$. ⊣

LEMMA 4.12. *Let $A \in \{0,1\}^{\mathbb{N}}$ be a computable sequence and $\mu = \mathbf{1}_A$. Then, $p \notin R(\mu)$ for each positive computable real $p \neq 1$.*

PROOF. Let ξ be a computable measure on $\{0,1\}^{\mathbb{N}}$ dominating ν constructed in Lemma 4.3. Then, $L_{1,\infty}(\mu,\xi)$ is ML-random by Lemma 4.11. We also have

$$\begin{aligned} L_{p,\infty}(\mu,\xi) &= \sum_{n=1}^{\infty} \ell_{p,A_{<n}}(\mu,\xi) = \sum_{n=1}^{\infty} \sum_{a\in\{0,1\}} |\mu(a|A_{<n}) - \xi(a|A_{<n})|^p \\ &= 2\sum_{n=1}^{\infty} |\mu(A_n|A_{<n}) - \xi(A_n|A_{<n})|^p. \end{aligned}$$

Now, by Theorem 4.8(ii), $L_{p,\infty}(\mu,\xi) = \infty$ for each computable $p \in (0,1)$. Similarly, by Theorem 4.8(i), $L_{p,\infty}(\mu,\xi) < \infty$ is not ML-random for each computable $p > 1$, which is not Solovay equivalent to a left-c.e. ML-random real $D_\infty(\mu,\xi)$. Hence, $p \notin R(\mu)$ for each positive computable real $p \neq 1$. ⊣

PROOF OF PROPOSITION 4.9. The claim $R(\mu) = 1$ follows from Lemmas 4.11 and 4.12. Since $1 \in R(\mu)$, we have $L_{1,\infty}(\mu,\xi) < \infty$ and $D_\infty(\mu||\xi) \equiv_S L_{1,\infty}(\mu,\xi)$ for all computable measures ξ dominating μ. By Theorem 4.1, there exists a computable measure ν such that $D_\infty(\mu||\xi)$ is a left-c.e. ML-random real for

all computable measures ξ dominating ν. Thus, $L_{1,\infty}(\mu,\xi)$ is ML-random for all computable measures ξ dominating μ and ν. ⊣

When the model measure is a Dirac measure, the rate of convergence can be expressed more concretely by time-bounded Kolmogorov complexity. Let $h:\mathbb{N}\to\mathbb{N}$ be a computable function, and let $M :\subseteq \{0,1\}^*\to\mathbb{N}$ be a prefix-free machine. The Kolmogorov complexity relative to M with time bound h is

$$K^h_M(\sigma)=\min\{|\tau| \ : \ M(\tau)=\sigma \text{ in at most } h(|\sigma|) \text{ steps }\}.$$

Here, $h:\mathbb{N}\to\mathbb{N}$ is a total computable function. We write $K^h(\sigma)$ as the mean $K^h_U(\sigma)$ for a fixed universal prefix-free machine U.

PROPOSITION 4.13. *Let $A\in\{0,1\}^{\mathbb{N}}$ be a computable sequence.*

(i) *For every total computable prediction ξ dominating $\mu=\mathbf{1}_A$, there exists a computable function $h:\mathbb{N}\to\mathbb{N}$ such that*

$$K^h(n)\leq^{+} -\log(1-\xi(A_n|A_{<n}))+O(1).$$

(ii) *For every total 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.

Here, log *is the logarithm with base* 2.

From this theorem, we know that the error $1-\xi(A_n|A_{<n})$ is essentially the same as $2^{-K^h(n)}$ up to a multiplicative constant. We use this formulation because of the non-optimality of the time-bounded Kolmogorov complexity.

PROOF. (i) By Proposition 4.9, we have

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

By the KC-theorem [9, Theorem 3.6.1], there exists a prefix-free machine $M :\subseteq \{0,1\}^*\to\mathbb{N}$ and a computable sequence $(\sigma_n)_n$ of strings such that

$$M(\sigma_n)=n,\ \ |\sigma_n|\leq^{+} -\log(1-\xi(A_n|A_{<n}))+O(1).$$

Let $\tau\in\{0,1\}^*$ be a string such that $U(\tau\sigma)\simeq M(\sigma)$ for all $\sigma\in\{0,1\}^*$. Then, the function $n\mapsto U(\tau\sigma_n)$ is a total computable function. Therefore, there exists a total computable function $h:\mathbb{N}\to\mathbb{N}$ such that, for every $n\in\mathbb{N}$, the computation of $U(\tau\sigma_n)$ halts within at most $h(n)$ steps. By this definition of h, we obtain

$$K^h(n)\leq |\tau|+|\sigma_n|\,.$$

(ii) We define a computable prediction measure ν by

$$\nu=\sum_n 2^{-K^h(n)}\mathbf{1}_{A_{<n}\overline{A_n}0^{\mathbb{N}}}+(1-s)\mathbf{1}_A,$$

where $s=\sum_n 2^{-K^h(n)}<1$ and $\overline{k}=1-k$ for $k\in\{0,1\}$.

We claim that this measure ν is computable. We show that $\nu(\sigma)$ is computable uniformly in $\sigma \in \{0,1\}^*$. If $\sigma \prec A$, then

$$\nu(\sigma) = \sum_{n>|\sigma|} 2^{-K^h(n)} + (1-s) = 1 - \sum_{n\leq|\sigma|} 2^{-K^h(n)}.$$

If $\sigma = A_{<k}\overline{A_k}0^i$ for some $k, i \in \mathbb{N}$, then

$$\nu(\sigma) = 2^{-K^h(k)}.$$

If $\sigma = A_{<k}\overline{A_k}0^i 1\tau$ for some $k, i \in \mathbb{N}$ and $\tau \in \{0,1\}^*$, then

$$\nu(\sigma) = 0.$$

In any case, $\nu(\sigma)$ is computable from n. Furthermore, these relations are decidable.

Let ξ be a computable measure dominating ν. Then, there exists $c \in \mathbb{N}$ such that $\nu(\sigma) \leq c\xi(\sigma)$ for all $\sigma \in \{0,1\}^*$. Then,

$$1 - \xi(A_n|A_{<n}) = 1 - \frac{\xi(A_{\leq n})}{\xi(A_{<n})} = \frac{\xi(A_{<n}\overline{A_n})}{\xi(A_{<n})} \geq \frac{\nu(A_{<n}\overline{A_n})}{c} = \frac{2^{-K^h(n)}}{c}. \qquad \dashv$$

4.4. Case of separated measures. Now, we discuss the convergence rate of general computable predictions when the computable model measure is separated. In this case, the convergence rate is much slower than that for the Dirac measures.

We call a measure to be separated if the conditional probabilities are far away from 0 and 1. A formal definition is as follows.

DEFINITION 4.14 (See before Theorem 196 in [22]). A measure μ on $\{0,1\}^{\mathbb{N}}$ is called *separated* (from 0 to 1), if

$$\inf_{\sigma\in\{0,1\}^*,\, k\in\{0,1\}} \mu(k|\sigma) > 0.$$

REMARK 4.15. Li–Vitányi's book called this notion "conditionally bounded away from zero" [15, Definition 5.2.3].

PROPOSITION 4.16. *Let μ be a computable separated measure. Then, $R(\mu) = 2$. In particular, $L_{2,\infty}(\mu,\xi) < \infty$ and is a left-c.e. ML-random real for all sufficiently general computable prediction measure ξ.*

LEMMA 4.17. *Let μ be a computable separated measure. Then, $2 \in R(\mu)$.*

In the following proof, we use a version of Pinsker's inequality and a reverse Pinsker inequality. A Pinsker inequality bounds the squared total variation from above by the KL divergence (see, for example, Verdú [26, (51)]). A reverse inequality does not hold in general, but it does under separation assumptions (see, for instance, [8, Lemma 6.3]). For a more comprehensive survey, see the work of Sason [21].

PROOF. Let ξ be a computable measure dominating μ. By Pinsker's inequality and a reverse Pinsker inequality, there are $a, b \in \mathbb{N}$ such that

$$(\ell_{1,\sigma}(\mu,\xi))^2 \leq a \cdot d_\sigma(\mu||\xi) \leq b \cdot (\ell_{1,\sigma}(\mu,\xi))^2.$$

Now we look at the relation between $(\ell_{1,\sigma}(\mu,\xi))^2$ and $\ell_{2,\sigma}(\mu,\xi)$. We use the inequalities

$$x^2 + y^2 \leq (x + y)^2 \leq 2(x^2 + y^2)$$

for $x, y \geq 0$ to deduce

$$\ell_{2,\sigma}(\mu, \xi) \leq a \cdot d_\sigma(\mu||\xi) \leq 2b \cdot \ell_{2,\sigma}(\mu, \xi). \tag{11}$$

The first inequality implies

$$L_{2,\infty}(\mu, \xi) \leq aD_\infty(\mu||\xi) < \infty$$

by Theorem 4.1, and thus $2 \in R(\mu)$. The first inequality in (11) also implies the existence of a computable function $f : \{0,1\}^* \to \mathbb{R}$ such that

$$\ell_{2,\sigma}(\mu, \xi) + f(\sigma) = ad_\sigma(\mu||\xi),$$

and thus the existence of a left-c.e. real γ such that

$$L_{2,\sigma}(\mu, \xi) + \gamma = aD_\infty(\mu||\xi).$$

Hence, $L_{2,\infty}(\mu, \xi) \leq_S D_\infty(\mu, \xi)$. Similarly, the second inequality in (11) implies $D_\infty(\mu, \xi) \leq_S L_{2,\infty}(\mu, \xi)$. Hence, we have $L_{2,\infty}(\mu, \xi) \equiv_S D_\infty(\mu, \xi)$. ⊣

LEMMA 4.18. *Let μ be a computable separated measure. Then, $p \notin R(\mu)$ for each positive computable real $p \neq 2$.*

PROOF. By Theorem 4.1, there exists a computable ξ such that ξ dominates μ and $D_\infty(\mu||\xi)$ is a finite left-c.e. ML-random real. By Lemma 4.17, $D_\infty(\mu||\xi) \equiv_S L_{2,\infty}(\mu, \xi)$, which implies $L_{2,\infty}(\mu, \xi)$ is a finite left-c.e. ML-random real by Proposition 4.6. By (ii) of Theorem 4.8, we have $L_{p,\infty}(\mu, \xi) = \infty$ for each $p \in (0, 2)$. In particular, $p \notin R(\mu)$ for each $p \in (0, 2)$.

Let $p > 2$ be a computable real. We construct a computable measure ν such that:

(i) ν dominates μ,
(ii) $\dim(L_{2,\infty}(\mu, \nu)) = \frac{1}{2}$,
(iii) $\dim(L_{p,\infty}(\mu, \nu)) = \frac{1}{p}$.

Suppose such a measure ν exists and $p \in R(\mu)$. By $2 \in R(\mu)$ and (i), we have $D_\infty(\mu, \nu) \equiv_S L_{2,\infty}(\mu, \nu)$. By $p \in R(\mu)$ and (i), we have $D_\infty(\mu, \nu) \equiv_S L_{p,\infty}(\mu, \nu)$. Since Solovay equivalence implies the same effective Hausdorff dimension, we have $\dim(L_{2,\infty}(\mu, \nu)) = \dim(L_{p,\infty}(\mu, \nu))$, which contradicts with (ii) and (iii). Thus, $p \notin R(\mu)$.

The construction of ν is as follows. Let α be a rational such that $0 < \alpha < \inf\{\mu(a|\sigma) : a \in \{0,1\}, \sigma \in \{0,1\}^*\}$. Since μ is separated, such α exists. Let $(z_n)_n$ be a computable sequence of positive rationals such that $z_n < \frac{\alpha}{2}$ and $\sum_{n=0}^{\infty} z_n$ is a finite left-c.e. ML-random real. Fix a sufficiently small rational $\varepsilon > 0$. Consider a computable function $\sigma \in \{0,1\}^* \mapsto a_\sigma \in \{0,1\}$ such that $\mu(a_\sigma|\sigma) > \frac{1}{2} - \varepsilon$. We define a computable measure ν as follows:

$$\nu(a|\sigma) = \begin{cases} \mu(a|\sigma) - z_{|\sigma|}, & \text{if } a = a_\sigma, \\ \mu(a|\sigma) + z_{|\sigma|}, & \text{if } a \neq a_\sigma. \end{cases}$$

(i). First we evaluate $\nu(a|\sigma)/\mu(a|\sigma)$. If $a = a_\sigma$, then

$$\frac{\nu(a|\sigma)}{\mu(a|\sigma)} = 1 - \frac{z_{|\sigma|}}{\mu(a_\sigma|\sigma)} \geq 1 - \frac{z_{|\sigma|}}{1/2 - \varepsilon}.$$

If $a \neq \sigma$, then

$$\frac{\nu(a|\sigma)}{\mu(a|\sigma)} = 1 + \frac{z_{|\sigma|}}{\mu(a|\sigma)} \geq 1.$$

Thus, we have

$$\nu(\sigma) = \prod_{n=1}^{|\sigma|} \nu(\sigma_n|\sigma_{<n}) \geq \prod_{n=1}^{|\sigma|} \mu(\sigma_n|\sigma_{<n})(1 - \frac{z_{n-1}}{1/2 - \varepsilon}) \geq \frac{\mu(\sigma)}{c}$$

for some constant $c \in \mathbb{N}$.

(ii)(iii). Notice that

$$L_{1,\infty}(\mu, \nu) = \sum_{n=0}^{\infty} z_n$$

is a finite left-c.e. ML-random real, and that

$$L_{q,\infty}(\mu, \nu) = \sum_{n=0}^{\infty} z_n^q$$

for any $q \geq 1$. Thus, the claims follow by Theorem 4.8. ⊣

Proof of Proposition 4.16. The claim $R(\mu) = 2$ follows by Lemmas 4.17 and 4.18. Since $2 \in R(\mu)$, we have $L_{2,\infty}(\mu, \xi) < \infty$ and $D_\infty(\mu||\xi) \equiv_S L_{2,\infty}(\mu, \xi)$ for all computable measures ξ dominating μ. By Theorem 4.1, there exists a computable measure ν such that $D_\infty(\mu||\xi)$ is a left-c.e. ML-random real for all computable measures ξ dominating ν. Thus, $L_{2,\infty}(\mu, \xi)$ is ML-random for all computable measures ξ dominating μ and ν. ⊣

Acknowledgments. The author appreciates the anonymous reviewers' efforts and helpful feedback. In particular, the proof of Theorem 3.2 was shortened by one of the reviewers.

Funding. The author is supported by Research Project Grant (B) by Institute of Science and Technology Meiji University, and JSPS KAKENHI (Grant Numbers 22K03408, 21K18585, 21K03340, and 21H03392). This work was also supported by the Research Institute for Mathematical Sciences, an International Joint Usage/Research Center located in Kyoto University.

REFERENCES

[1] G. Barmpalias and A. Lewis-Pye, *Differences of halting probabilities*. ***Journal of Computer and System Sciences***, vol. 89 (2017), pp. 349–360.

[2] L. Bienvenu, R. Downey, A. Nies, and W. Merkle, *Solovay functions and their applications in algorithmic randomness*. ***Journal of Computer and System Sciences***, vol. 81 (2015), no. 8, pp. 1575–1591.

[3] L. Bienvenu, P. Gács, M. Hoyrup, C. Rojas, and A. Shen, *Algorithmic tests and randomness with respect to a class of measures*, ***Proceedings of the Steklov Institute of Mathematics***, volume 274, Springer, Berlin, Heidelberg, 2011, pp. 41–102.

[4] L. Bienvenu and C. Porter, *Strong reductions in effective randomness*. ***Theoretical Computer Science***, vol. 459 (2012), pp. 55–68.

[5] V. Bogachev, ***Measure Theory***, Springer, Berlin, Heidelberg, 2007.

[6] V. Brattka, P. Hertling, and K. Weihrauch, *A tutorial on computable analysis*, ***New Computational Paradigms*** (S. B. Cooper, B. Löwe, and A. Sorbi, editors), Springer, New York, 2008, pp. 425–491.

[7] T. M. Cover and J. A. Thomas, ***Elements of Information Theory***, second ed., John Wiley & Sons, Hoboken, NJ, 2006.

[8] I. Csiszar and Z. Talata. *Context tree estimation for not necessarily finite memory processes, via bic and mdl*, ***Proceedings of International Symposium on Information Theory, 2005 (ISIT 2005)***, 2005, pp. 755–759.

[9] R. G. Downey and D. R. Hirschfeldt, ***Algorithmic Randomness and Complexity***, Theory and Applications of Computability, Springer, New York, 2010.

[10] R. G. Downey, D. R. Hirschfeldt, and A. Nies, *Randomness, computability, and density*. ***SIAM Journal on Computing***, vol. 31 (2002), no. 4, pp. 1169–1183.

[11] R. Durrett. ***Probability: Theory and Examples***, fourth ed., Cambridge University Press, Cambridge, 2010.

[12] R. Hölzl, T. Kräling, and W. Merkle, *Time-bounded Kolmogorov complexity and Solovay functions*. ***Theory of Computing Systems***, vol. 52 (2013), no. 1, pp. 80–94.

[13] M. Hutter, ***Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability***, Springer, Berlin, Heidelberg, 2005.

[14] M. Hutter and A. Muchnik, *On semimeasures predicting Martin-Löf random sequences*. ***Theoretical Computer Science***, vol. 382 (2007), pp. 247–261.

[15] M. Li and P. Vitányi, ***An Introduction to Kolmogorov Complexity and its Applications***, fourth ed., Texts in Computer Science, Springer, Cham, New York, 2019.

[16] J. S. Miller, *On work of Barmpalias and Lewis-Pye: A derivation on the D.C.E. reals*, ***Computability and Complexity - Essays Dedicated to Rodney G. Downey on the Occasion of His 60th Birthday***, volume 10010 of Lecture Notes in Computer Science (A. R. Day, M. R. Fellows, N. Greenberg, B. Khoussainov, A. G. Melnikov, and F. A. Rosamond, editors), Springer, Cham, 2017, pp. 644–659.

[17] K. Miyabe. *Computable prediction*, ***Artificial General Intelligence. AGI 2019***, volume 11654 of Lecture Notes in Computer Science (P. Hammer, P. Agrawal, B. Goertzel, and M. Iklé, editors), Springer, Cham, 2019, pp. 137–147.

[18] O. A. Nielsen, ***An Introduction to Integration and Measure Theory***, John Wiley & Sons, New York, NY, 1997.

[19] A. Nies, ***Computability and Randomness***, vol. 51, Oxford University Press, Oxford, 2009.

[20] S. Rathmanner and M. Hutter, *A philosophical treatise of universal induction*. ***Entoropy***, vol. 13 (2011), pp. 1076–1136.

[21] I. Sason, *On reverse Pinsker inequalities*, preprint, 2015, arXiv:1503.07118, submitted 24 March 2015; revised 13 April 2015.

[22] A. Shen, V. A. Uspensky, and N. Vereshchagin, ***Kolmogorov Complexity and Algorithmic Randomness***, American Mathematical Society, Providence, RI, 2017.

[23] R. I. Soare, ***Turing Computability***, Theory and Applications of Computability, Springer, Berlin, Heidelberg, 2016.

[24] R. J. Solomonoff, *Complexity-based induction systems: Comparisons and convergence theorems*. ***IEEE Transaction on Information Theory***, vol. IT-24 (1978), pp. 422–432.

[25] K. Tadaki, *A generalization of Chaitin's halting probability Ω and halting self-similar sets*. ***Hokkaido Mathematical Journal***, vol. 31 (2002), no. 1, pp. 219–253.

[26] S. Verdú, *Total variation distance and the distribution of relative information*, ***2014 Information Theory and Applications Workshop (ITA)***, San Diego, CA, 2014, pp. 1–3.

[27] K. Weihrauch, *Computability on the probability measures on the Borel sets of the unit interval*. ***Theoretical Computer Science***, vol. 219 (1999), nos. 1–2, pp. 421–437.

[28] ———, ***Computable Analysis: An Introduction***, Springer, Berlin, 2000.

[29] ———, *Computability on measurable functions*. ***Computability***, vol. 6 (2017), no. 1, pp. 79–104.

DEPARTMENT OF MATHEMATICS
MEIJI UNIVERSITY
JAPAN
E-mail: research@kenshi.miyabe.name