AN INFORMATION-THEORETIC UPPER BOUND ON PRIME GAPS

Aidan Rocke
aidanrocke@gmail.com

October 21, 2021

ABSTRACT

The method of level sets allows us to define an equivalence relation over sequences of rare events with distinct rates of entropy production. This method clarifies the relation between the empirical density of primes and their source distribution which then allows us to address Cramér's conjecture, an open problem in probabilistic number theory. In consequence, this analysis places strong epistemic limits on the application of machine learning to analyse the distribution of primes.

1 Classifying rare events using level sets

Sequences of rare events that are not finite-state compressible are recurrent, of variable frequency, and unpredictable relative to a finite-state machine such as a machine learning model. For these reasons, they are of great scientific interest. But, how might we classify rare events that satisfy these criteria?

If we identify rare events using the indicator function $1_X : \mathbb{N} \to \{0, 1\}$, we may analyse the rate of entropy production of the computable binary sequence $X_n = \{x_i\}_{i=1}^n \in \{0, 1\}^n$. In order to qualify as a rare event, a randomly sampled element x_z of the sequence X_n must generally pass $\pi_x(n)$ tests. These tests $A_i, A_{j\neq i} \in \{A_i\}_{i=1}^{\pi_x(n)}$ are de-correlated for rare events as these are assumed to occur independently of each other:

$$\forall z \sim U([1, n]), P(z \in A_i \wedge z \in A_{j\neq i}) = P(z \in A_i) \cdot P(z \in A_{j\neq i}) \tag{1}$$

where the $P(z \in A_i)$ have a natural frequentist interpretation. We generally assume that the tests are consistent so $\{A_i\}_{i=1}^{\pi_x(n)} \subset \{A_i\}_{i=1}^{\pi_x(n+1)}$.

Furthermore, $\pi_x(n)$ counts the number of rare events in the discrete time interval $[1, n]$ so it is monotonically increasing. This may be deduced from the fact that each rare event that occurs in an interval defines one information-theoretic constraint A_i where $A_i \perp A_{j\neq i}$ due to pairwise independence. Therefore, if m rare events occur in an interval, a uniformly sampled element x_z must simultaneously satisfy m independent constraints in order to qualify as a rare event.

Given (1), we may define the average *probability density* using the inclusion-exclusion principle:

$$\forall z \sim U([1, n]), P(x_z = 1) = P(z \in \bigcap_{k\leq\pi_x(n)} A_k) = \prod_{k=1}^{\pi_x(n)} P(z \in A_k) \tag{2}$$

as well as the entropy of the uniformly distributed random variable:

$$\forall z \sim U([1, n]), H(X_n) = -\ln \prod_{k=1}^{\pi_x(n)} P(z \in A_k) \tag{3}$$

which is a measure of expected surprise, or the expected information gained from observing a rare event. Thus, all the information in X_n is contained in the location of rare events and the true-positive rate is a robust measure of finite-state incompressibility relative to a machine learning model as explained in section A of the Appendix.

Having defined the entropy (3), we may compare the entropies of distinct sequences using the typical probabilities $q_n \in (0,1)$, a parameter which allows us to define an equivalence relation over entropies:

$$\mathcal{L}_{q_n} = \{X_n \in \{0,1\}^\infty : H(X_n) \sim \ln\big(\frac{1}{q_n}\big)\} \tag{4}$$

This parameter is chosen because it corresponds to the multiplicative inverse of the exponential of entropy known hereafter as the *entropy rate* $\frac{1}{q_n} \approx e^{H(X_n)}$ which is a robust measure of the expected waiting time before rare events. As these parameters are both parameterisation invariant i.e. don't depend upon the choice of base for logarithms, they are an ideal choice for classifying rare events in terms of their rate of entropy production.

Given the definition (4), we have:

$$\forall X_n \in \mathcal{L}_{q_n}, \lim_{n\to\infty} -\frac{\ln q_n}{H(X_n)} = 1 \tag{5}$$

so the following holds *asymptotically almost surely*:

$$\frac{\partial}{\partial q_n} H(X_n) = \frac{\partial}{\partial q_n} - \ln q_n \implies \forall z \sim U([1,N]), P(x_z = 1) \sim q_N \tag{6}$$

and therefore for large $N \in \mathbb{N}$:

$$\pi_x(N) \sim N \cdot q_N \tag{7}$$

This leads us naturally to consider statistical generalisations of the Prime Number Theorem.

1.1 Statistical generalisations of the Prime Number Theorem

From statistical considerations, Gauss and Legendre inferred that the density of primes at $n \in \mathbb{N}$ is on the order of $\sim \frac{1}{\ln n}$. This led them to conjecture that the number of primes less than N is on the order of:

$$\pi(N) \sim \int_2^N \frac{1}{\ln x} dx \sim N \cdot \frac{1}{\ln N} \tag{8}$$

which is equivalent to the Prime Number Theorem.

However, given our assumptions we are more generally interested in the family of density functions that satisfy:

$$0 < c \ll N, \int_c^N \frac{1}{f(x)} dx \sim N \cdot \frac{1}{f(N)} \tag{9}$$

where f is analytic and $\lim_{N\to\infty} \frac{1}{f(N)} = 0$. Using integration by parts, this is equivalent to the criterion:

$$\int_c^N g(x)dx = N \cdot g(N) - c \cdot g(c) - \int_c^N x \cdot g'(x)dx \sim N \cdot g(N) \tag{10}$$

where $g(x) = \frac{1}{f(x)}$, and $g(x)$ dominates $x \cdot g'(x)$. So this family naturally includes density functions of the form:

$$\exists a > 0 \forall x \in \mathbb{R}_+, g(x) = (\ln x)^{-a} \tag{11}$$

and it is worth adding that (11) admits an interpretation as an average probability density which we may now make precise.

1.2 The typical probability as the average probability

Given that the density $g(x) = (\ln x)^{-a}$ is Riemann-Integrable on $[2, N]$ for $N < \infty$, by the Intermediate Value Theorem there exists a sequence $x_n \in [n, n+1]$ such that:

$$\int_2^N (\ln x)^{-a} dx = \sum_{n=2}^{N} \frac{1}{(\ln x_n)^a} \tag{12}$$

and since $\forall n \in \mathbb{N}, 1 \leq \left(\frac{\ln x_n}{\ln n}\right)^a \leq \left(\frac{\ln(n+1)}{\ln n}\right)^a$ where $\forall a > 0$, $\lim_{n\to\infty} \left(\frac{\ln(n+1)}{\ln n}\right)^a = 1$ so we have:

$$\int_2^N (\ln x)^{-a} dx \sim \sum_{n=2}^{N} \frac{1}{(\ln n)^a} \sim N \cdot \frac{1}{(\ln N)^a} \tag{13}$$

and therefore the typical probability is asymptotically equivalent to the average probability:

$$q_N \sim \frac{1}{N} \sum_{n=2}^{N} \frac{1}{(\ln n)^a} \tag{14}$$

In fact, for large n the average probability density in (2) may be expressed as:

$$\forall z \sim U([1, n]), P(x_z = 1) \sim \frac{1}{n} \sum_{k=1}^{n} P(x_k = 1) \tag{15}$$

1.3 A remark on applications of the level set method

In practice, the empirical data will consist of a binary sequence of rare event observations and the actual tests are generally unknown. These may be modelled as latent variables that are a function of discrete time(i.e. the integers) as well as hidden variables that provide us with initial conditions. Thus, we shall generally assume that empirical binary sequences are the output of a discrete dynamical system.

Assuming that deep neural networks may be used as universal data compressors, we may determine whether the data is finite-state incompressible using a machine-learning driven overfitting test provided in the appendix [12]. As a concrete demonstration of the analytical power of the level set method, we shall consider its application to an open problem in probabilistic number theory.

However, we must first carefully go over the information-theoretic derivation of the Prime Number Theorem [9].

2 Information-theoretic derivation of the Prime Number Theorem

If we know nothing about the distribution of primes, in the worst case we may assume that each prime less than or equal to N is drawn uniformly from $[1, N]$. So our source of primes is:

$$X \sim U([1, N]) \tag{16}$$

where $H(X) = \ln N$ is the Shannon entropy of the uniform distribution.

Now, we may define the prime encoding of $[1, N]$ as the binary sequence $X_N = \{x_n\}_{n=1}^{N}$ where $x_n = 1$ if n is prime and $x_n = 0$ otherwise. With no prior knowledge, given that each integer is either prime or not prime, we have 2^N possible prime encodings in $[1, N] \subset \mathbb{N}$.

If there are $\pi(N)$ primes less than or equal to N then the average number of bits per arrangement gives us the average amount of information gained from correctly identifying each prime in $[1, N]$ as:

$$S_c = \frac{\log_2(2^N)}{\pi(N)} = \frac{N}{\pi(N)} \tag{17}$$

and if we assume a maximum entropy distribution over the primes then we would expect that each prime is drawn from a uniform distribution as in (16). So we would have:

$$S_c = \frac{N}{\pi(N)} \sim \ln N \tag{18}$$

As for why the natural logarithm appears in (21), we may first note that the base of the logarithm in the Shannon Entropy may be freely chosen without changing its properties. Moreover, given the assumptions the expected information gained from observing each prime in $[1, N]$ is on the order of:

$$\sum_{k=1}^{N-1} \frac{1}{k} \cdot |(k, k+1]| = \sum_{k=1}^{N-1} \frac{1}{k} \approx \ln N \tag{19}$$

as there are k distinct ways to sample uniformly from $[1, k]$ and a frequency of $\frac{1}{k}$ associated with the event that $k \in \mathbb{P}$.

This implies that the average number of bits per prime number is given by $\frac{N}{\pi(N)} \sim \ln N$. Rearranging, we find:

$$\frac{\pi(N)}{N} \sim \frac{1}{\ln N} \tag{20}$$

which is in complete agreement with the Prime Number Theorem. More importantly, this derivation indicates that the prime numbers are empirically distributed as if they were arranged uniformly.

2.1 The Shannon source coding theorem and the algorithmic randomness of prime encodings

By the Shannon source coding theorem, we may infer that $\pi(N)$ primes can't be compressed into fewer than $\pi(N) \cdot \ln N$ bits. Furthermore, as the expected Kolmogorov Complexity equals the Shannon entropy for computable probability distributions:

$$\mathbb{E}[K(X_N)] \sim \pi(N) \cdot \ln N \sim N \tag{21}$$

where the identification of $\mathbb{E}[K(X_N)]$ with $\pi(N) \cdot \ln N$ is a direct consequence of the fact that $K(X_N)$ measures the information gained from observing X_N and $\pi(N) \cdot \ln N$ measures the expected information gained from observing X_N. From an information-theoretic perspective, this implies that prime encodings are finite-state incompressible as they have a maximum entropy distribution.

The only implicit assumption in this derivation is that all the information in the Universe is conserved as it is only in such Universes that Occam's razor is generally applicable. The Law of Conservation of information and its significance is discussed in more detail in section B of the Appendix.

3 Clarifying the relation between the empirical density of primes and their source distribution

In the worst case, rare events $x_k \in X_n$ are arranged uniformly in the discrete time interval $[1, n]$ so the expected information gained from observing all the events in X_n is on the order of:

$$I_N = \sum_{n=1}^{N} q_n \cdot \ln n \tag{22}$$

where each observation in X_n contributes at most one bit of information, so we have:

$$0 \leq I_N \leq N \tag{23}$$

Furthermore, in order to model rare events it is crucial to consider the sum of entropies (3):

$$S_N = \sum_{n=1}^{N} H(X_n) \tag{24}$$

subject to the constraint $H(X_n) \sim -\ln q_n$ so we may define the Lagrangian function:

$$\mathcal{L}(\lambda, q_n) = \sum_{n=1}^{N} q_n \cdot \ln n - \lambda\big(S_N + \sum_{n=1}^{N} \ln q_n\big) \tag{25}$$

Thus, we find that:

$$\frac{\partial \mathcal{L}}{\partial q_n} = \ln n - \frac{\lambda}{q_n} = 0 \implies q_n = \frac{\lambda}{\ln n} \tag{26}$$

where $\lambda = 1$ maximises the expected information gained (23):

$$q_n = \frac{1}{\ln n} \tag{27}$$

and therefore among the rare-event densities, the empirical density of primes(which we may observe) is a natural model for a uniform source distribution which is not directly observable.

4 Application to Cramér's random model, Part I

Using the method of level sets, we may demonstrate that the typical probability that an integer in the interval $[1, n] \subset \mathbb{N}$ is prime is given by:

$$\forall z \sim U([1, n]), P(z \in \mathbb{P}) \sim \frac{1}{\ln n} \tag{28}$$

If $X_n = \{x_i\}_{i=1}^n \in \{0, 1\}^n$ defines a prime encoding i.e. a sequence where $x_k = 1$ if $k \in \mathbb{P}$ and $x_k = 0$ otherwise, then we may define $\pi(\sqrt{n})$ primality tests as any composite integer in $[1, n]$ has at most $\pi(\sqrt{n})$ distinct prime factors:

$$\forall A_p \in \{A_{p_k}\}_{k=1}^{\pi(\sqrt{n})}, A_p = \{z \in [1, n] : \text{gcd}(p, z) = 1\} \bigcup \{p\} \tag{29}$$

and therefore:

$$\forall z \sim U([1, n]), H(X_n) = -\ln \prod_{k=1}^{\pi(\sqrt{n})} P(z \in A_{p_k}) \tag{30}$$

In order to define $P(z \in A_{p_k})$ we shall implicitly use the approximation that for $p \leq \sqrt{n}$, $\frac{1}{p} - \frac{1}{n} \approx \frac{1}{p}$ so we have:

$$\forall z \sim U([1, n]), P(z \in A_{p_k}) = \big(1 - \frac{1}{p_k}\big) + \frac{1}{n} \approx \big(1 - \frac{1}{p_k}\big) \tag{31}$$

which allows us to make (30) precise:

$$H(X_n) \sim -\ln \prod_{p \leq \sqrt{n}} \big(1 - \frac{1}{p}\big) \tag{32}$$

Using Mertens' third theorem we have:

$$\prod_{p \leq \sqrt{n}} \big(1 - \frac{1}{p}\big) \approx \frac{e^{-\gamma}}{\frac{1}{2} \cdot \ln n} \approx \frac{0.9}{\ln n} \tag{33}$$

so we may infer that:

$$H(X_n) \sim \ln \ln n \tag{34}$$

and therefore $X_n \in \mathcal{L}_{q_n}$ where:

$$\mathcal{L}_{q_n} = \{X_n \in \{0, 1\}^\infty : H(X_n) \sim \ln \ln n\} \tag{35}$$

From (35), we may deduce (28):

$$\forall z \sim U([1, n]), q_n = P(z \in \mathbb{P}) \sim \frac{1}{\ln n} \tag{36}$$

which means that the density of the primes at $x \in \mathbb{N}$ is on the order of $\sim \frac{1}{\ln x}$ and therefore the expected number of primes less than n is given by:

$$\pi(n) \sim \int_2^n \frac{1}{\ln x} dx \sim \frac{n}{\ln n} \tag{37}$$

in complete agreement with the Prime Number Theorem.

5 Application to Cramér's random model, Part II

Given the prime encoding X_n with uniform source distribution, let's define the nth prime gap $G_n = p_{n+1} - p_n$ so we may consider the cumulative probability:

$$\sum_{k=1}^{G_n} P(x_{p_n+k} = 1 \wedge p_{n+1} - p_n = k) = 1 \tag{38}$$

and its associated entropy:

$$H_n = \sum_{k=1}^{G_n} -P(x_{p_n+k} = 1 \wedge p_{n+1} - p_n = k) \cdot \ln P(x_{p_n+k} = 1 \wedge p_{n+1} - p_n = k) \tag{39}$$

where $G_n < p_n$ due to Bertrand's postulate and we'll note that

$$\exists k, P(p_{n+1} - p_n = k) = 1 \implies H_n = 0 \tag{40}$$

so the entropy H_n collapses if and only if we simultaneously measure the values of both p_n and p_{n+1}.

As the subsequence $\{x_{p_n+k}\}_{k=1}^{G_n}$ halts at $k \in [1, G_n]$ where $x_{p_n+k} = 1$, there are at most G_n possible ways for this sequence to halt. Furthermore, in order to bound the *halting probability* $P(x_{p_n+k} = 1 \wedge p_{n+1} - p_n = k)$ we may use the density formula (37) to consider its components:

$$\forall k \sim U([1, G_n]), P(p_{n+1} = p_n + k) = P(p_n = p_{n+1} - k) \sim \frac{1}{\ln p_n} \tag{41}$$

$$\forall k \sim U([1, G_n]), P(x_{p_n+k} = 1) \sim \frac{1}{\ln p_n} \tag{42}$$

where

$$P(x_{p_n+k} = 1 \wedge p_{n+1} - p_n = k) \geq P(x_{p_n+k} = 1) \cdot P(p_{n+1} = p_n + k) \tag{43}$$

Using (41),(42) and (43) we may then derive the lower bound:

$$\forall k \sim U([1, G_n]), P(x_{p_n+k} = 1 \wedge p_{n+1} - p_n = k) \gtrsim \frac{1}{(\ln p_n)^2} \tag{44}$$

which implies:

$$\frac{1}{(\ln p_n)^2} \lesssim \frac{1}{G_n} \sum_{k=1}^{G_n} P(x_{p_n+k} = 1 \wedge p_{n+1} - p_n = k) = \frac{1}{G_n} \tag{45}$$

and since $G_n = p_{n+1} - p_n$, we may conclude:

$$p_{n+1} - p_n = \mathcal{O}((\ln p_n)^2) \tag{46}$$

as conjectured by Harald Cramér in 1936 [1].

6 Details of the $\frac{1}{p}-\frac{1}{n}\approx\frac{1}{p}$ approximation in Cramér's model

If we define,

$$\Lambda_k := \text{set of } \binom{\pi(\sqrt{n})}{k} \text{ distinct subsets of } \{p_k\}_{k=1}^{\pi(\sqrt{n})} \tag{47}$$

then $|\Lambda_k| = \binom{\pi(\sqrt{n})}{k}$ and we may derive the absolute error:

$$\Big|\prod_{p\leq\sqrt{n}}(1-\frac{1}{p}+\frac{1}{n})-\prod_{p\leq\sqrt{n}}(1-\frac{1}{p})\Big| \leq \sum_{k=1}^{\pi(\sqrt{n})-1}\frac{1}{n^k}\sum_{\lambda\in\Lambda_k}\prod_{p\in\lambda}(1-\frac{1}{p})+\frac{1}{n^{\pi(\sqrt{n})}} \tag{48}$$

and given that:

$$\pi(\sqrt{n})<\sqrt{n} \implies \frac{1}{n^k}\cdot\binom{\pi(\sqrt{n})}{k}\leq\frac{1}{n^{k/2}} \tag{49}$$

the absolute error satisfies the following inequality:

$$\Delta_n = \Big|\prod_{p\leq\sqrt{n}}(1-\frac{1}{p}+\frac{1}{n})-\prod_{p\leq\sqrt{n}}(1-\frac{1}{p})\Big| \leq \frac{1}{\sqrt{n}}+\frac{1}{n} \tag{50}$$

so the relative error converges to zero:

$$\lim_{n\to\infty}\frac{\Delta_n}{\prod_{p\leq\sqrt{n}}(1-\frac{1}{p})} \leq \lim_{n\to\infty}(\frac{\ln n}{\sqrt{n}}+\frac{\ln n}{n}) = 0 \tag{51}$$

which provides us with the necessary justifications in our derivation of Cramér's random model, specifically formulas (31),(32) and (33).

7 The incompressibility of prime encodings as a fundamental law of physics?

Information is physical.-Rolf Landauer

Let's suppose we get the world's best theoretical physicists to design a machine learning system whose aim is to predict the location of the Nth prime given the locations of the first $N-1$ primes. What is the best case scenario we can hope for?

Considering that the prime numbers have a maximum entropy distribution the true positive rate of any such system can't exceed 50%. This state of affairs should hold true at all times regardless of technological progress, given what is known about Turing Machines without access to Oracles. So this fundamental limit might as well be a Physical Law.

This Law might even have Cosmological implications. In fact, let's suppose that the Universe emerged from a Singularity which would be consistent with Big Bang Cosmology. At this precise instant, when the Universe and all of its mathematical structure came into existence there would have been no meaningful prior knowledge concerning the distribution of primes. Therefore, at the moment of the Singularity, the most reasonable machine learning system would have to assume a maximum entropy distribution over the location of the primes. This would be equivalent to the application of Occam's razor, a principle that is generally applicable in Universes where Quantum Information is conserved.

Thus, at this Singularity the Minimum Description Length of prime encodings would have been defined relative to a Universal Turing Machine as follows:

$$\mathbb{E}[K(X_N)] \sim \pi(N) \cdot \ln N \sim N \tag{52}$$

which is consistent with present-day observations i.e. the information-theoretic analyses on pages 4 and 5. How might we explain this mysterious coincidence?

In accordance with Big Bang Cosmology and the principle that all the Quantum Information in the Universe is conserved we are led to a natural Cosmological hypothesis concerning the distribution of primes. The maximum entropy distribution of prime encodings may be understood as a mathematical signature that time travelled from the earliest moments of the Big Bang.

The author concedes that such a Hypothesis would appear heretical to scientists who haven't considered the possibility that the Universe may be simulated by a Universal Quantum Turing Machine.

8 Conclusion

The finite-state incompressibility of prime encodings is of general scientific importance as it implies that it is not possible for any finite state machine such as a machine learning model to infer the definition of prime numbers(i.e. Unique Factorisation Theorem) from a prime encoding X_n of any length n. Equivalently, theoretical physicists can't design a machine learning system which has favourable odds of locating the next prime number i.e. a true-positive rate greater than 50%. Therefore, we may conclude that machine learning does not confer an advantage to a mathematician investigating the distribution of primes.

Having said this, it is worth clarifying that this analysis does not preclude the application of machine learning to derive number-theoretic insights. In fact, Yang-Hui He recently found machine learning to be particularly effective at identifying interesting structures in the setting of arithmetic geometry and he suspects that it is for the following reason [11]:

> At the most basic level, every computation in algebraic geometry, be it a spectral sequence or a Gröbner basis, reduces to finding kernels and cokernels of sets of matrices (over $\mathbb{Z}$ or even over $\mathbb{C}$), albeit of quickly forbidding dimensions. Matrix/tensor manipulation is the heart of any neural network. Number theory, on the other hand, ultimately involves patterns of prime numbers which, as is well known, remain elusive.

which is both consistent with our findings and gives us reason for measured optimism.

A An invariance theorem for algorithmically random data

Let's suppose we have a natural signal described by the process X:

$$x_n \in \{0,1\}, x_{n+1} = \varphi \circ x_{1:n} \tag{53}$$

If we should use machine learning to approximate φ given the datasets $X_N^{\text{train}} = \{x_i\}_{i=1}^{N}, X_N^{\text{test}} = \{x_i\}_{i=N+1}^{2N}$ such that for any $\hat{f} \in F_\theta$:

$$\exists k \in [1, n-1], x_{n+1} = \hat{f} \circ x_{n-k:n} \Rightarrow \delta_{\hat{f}(x_{n-k:n}),x_{n+1}} = 1 \tag{54}$$

then X_N is asymptotically incompressible if for large N any solution to the empirical risk minimisation problem:

$$\hat{f} = \max_{f \in F_\theta} \frac{1}{N-k} \sum_{n=k+1}^{N} \delta_{f(x_{n-k:n}),x_{n+1}} \tag{55}$$

has an expected performance:

$$\frac{1}{N-k} \sum_{n=N+k+1}^{2N-1} \delta_{\hat{f}(x_{n-k:n}),x_{n+1}} \leq \frac{1}{2} \tag{56}$$

Furthermore, if the dataset is imbalanced i.e. $\frac{1}{N}\sum_{i=1}^{N} x_i \neq \frac{1}{2}$ and all the information in the sequence x_n is contained in the location of unitary values then we may generalise this result by introducing the auxiliary definitions:

$$y_n = x_{n+1} \tag{57}$$

$$\hat{y_n} = \hat{f} \circ x_{n-k:n} \tag{58}$$

$$\beta_n = \delta_{y_n,\hat{y_n}} \tag{59}$$

$$N_1 = \sum_{n=N+k+1}^{2N} \delta_{y_n,1} \tag{60}$$

$$N_0 = \sum_{n=N+k+1}^{2N} \delta_{y_n,0} \tag{61}$$

and so for large N, we have:

$$\mathcal{L}_N[\hat{f}] = \min\Big[\frac{1}{N_0} \sum_{n=N+k+1}^{2N-1} \delta_{y_n,0} \cdot \beta_n, \frac{1}{N_1} \sum_{n=N+k+1}^{2N-1} \delta_{y_n,1} \cdot \beta_n\Big] \leq \frac{1}{2} \tag{62}$$

and therefore:

$$\forall \hat{f} \in F_\theta, \lim_{N\to\infty} P(\mathcal{L}_N[\hat{f}] > \frac{1}{2}) = 0 \tag{63}$$

Finally, as these results are invariant to transformations that preserve the phase-space dimension of X, this theorem may be used as an overfitting test for binary sequences that are finite-state incompressible.

B Revisiting the unreasonable effectiveness of mathematics

62 years since Eugene Wigner's highly influential essay on the unreasonable effectiveness of mathematics in the natural sciences, it may be time for a re-appraisal. On balance, with important theoretical advances in algorithmic information theory and Quantum Computation it appears that the remarkable effectiveness of mathematics in the natural sciences is quite reasonable.

By effectiveness, I am specifically referring to Wigner's observation that mathematical laws have remarkable generalisation power.

B.1 An information-theoretic perspective

An acute observer will note that the same mathematical laws with remarkable generalisation power in the natural sciences are also constrained by Occam's razor. Given two computable theories, Einstein explicitly stated that a physicist ought to choose the simplest theory that yields negligible experimental error:

> It can be scarcely denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience.-Einstein(1933)

In fact, from an information-theoretic perspective the remarkable generalisation power of mathematical laws in the natural sciences is a direct consequence of the effectiveness of Occam's razor.

B.2 The Law of Conservation of Information

From an information-theoretic perspective, a Universe where Occam's razor is generally applicable is one where information is generally conserved. This law of conservation of information which dates back to von Neumann essentially states that the von Neumann entropy is invariant to Unitary transformations. This is meaningful within the framework of Everettian Quantum Mechanics as a density matrix may be assigned to the state of the Universe. This way information is conserved as we run a simulation of the Universe forward in time.

Moreover, given that Occam's razor has an appropriate formulation within the context of algorithmic information theory as the Minimum Description Length principle, this information-theoretic perspective generally presumes that the Universe itself may be simulated by a Universal Turing Machine.

B.3 The Physical Church-Turing thesis

The research of David Deutsch(and others) on the Physical Church-Turing thesis explains how a Universal Quantum computer may simulate the laws of physics. This is consistent with the general belief that Quantum Mechanics may be used to simulate all of physics so the most important contributions to the Physical Church-Turing thesis have been via theories of quantum computation.

More importantly, the Physical Church-Turing thesis provides us with a credible explanation for the remarkable effectiveness of mathematics in the natural sciences.

B.4 What is truly remarkable

If we view the scientific method as an algorithmic search procedure then there is no reason, a priori, to suspect that a particular inductive bias should be particularly powerful. This much was established by David Wolpert in his No Free Lunch theorems [22].

On the other hand, the history of the natural sciences indicates that Occam's razor is remarkably effective. The effectiveness of this inductive bias has recently been used to explain the generalisation power of deep neural networks when applied to data that is finite-state compressible [23].

References

[1] Cramér, H. "On the Order of Magnitude of the Difference Between Consecutive Prime Numbers." Acta Arith. 2, 23-46, 1936.

[2] Hardy, G. H.; Ramanujan, S. (1917), "The normal number of prime factors of a number n", Quarterly Journal of Mathematics

[3] Turán, Pál (1934), "On a theorem of Hardy and Ramanujan", Journal of the London Mathematical Society

[4] Yufei Zhao. The Probabilistic Method in Combinatorics. 2019.

[5] F. Mertens. J. reine angew. Math. 78 (1874)

[6] Olivier Rioul. This is IT: A Primer on Shannon's Entropy and Information. Séminaire Poincaré. 2018.

[7] E.T. Jaynes. Information Theory and Statistical Mechanics. The Physical Review. 1957.

[8] Lance Fortnow. Kolmogorov Complexity. 2000.

[9] Aidan Rocke (https://mathoverflow.net/users/56328/aidan-rocke), information-theoretic derivation of the prime number theorem, URL (version: 2021-04-08): https://mathoverflow.net/q/384109

[10] Aidan Rocke (https://mathoverflow.net/users/56328/aidan-rocke), Egyptian number theory, URL (version: 2021-06-22): https://mathoverflow.net/q/395939

[11] Yang-Hui He. Deep-Learning the Landscape. Arxiv. 2018.

[12] Aidan Rocke (https://cstheory.stackexchange.com/users/47594/aidan-rocke), An invariance theorem for algorithmically random data in statistical learning, URL (version: 2021-02-22): https://cstheory.stackexchange.com/q/48452

[13] Eugene Wigner. The Unreasonable Effectiveness of Mathematics in the Natural Sciences. 1960.

[14] David Deutsch. Quantum theory, the Church–Turing principle and the universal quantum computer. 1985.

[15] Peter D. Grünwald. The Minimum Description Length Principle . MIT Press. 2007.

[16] A. N. Kolmogorov Three approaches to the quantitative definition of information. Problems of Information and Transmission, 1(1):1–7, 1965

[17] G. J. Chaitin On the length of programs for computing finite binary sequences: Statistical considerations. Journal of the ACM, 16(1):145–159, 1969.

[18] R. J. Solomonoff A formal theory of inductive inference: Parts 1 and 2. Information and Control, 7:1–22 and 224–254, 1964.

[19] Michael Nielsen. Interesting problems: The Church-Turing-Deutsch Principle. 2004. https://michaelnielsen.org/blog/interesting-problems-the-church-turing-deutsch-principle/

[20] Marcus Hutter et al. (2007) Algorithmic probability. Scholarpedia, 2(8):2572.

[21] The Evolution of Physics, Albert Einstein and Leopold Infeld, 1938, Edited by C.P. Snow, Cambridge University Press.

[22] Wolpert, D.H., Macready, W.G. (1997), "No Free Lunch Theorems for Optimization", IEEE Transactions on Evolutionary Computation 1, 67.

[23] Guillermo Valle Pérez, Chico Camargo, Ard Louis. Deep Learning generalizes because the parameter-function map is biased towards simple functions. 2019.