AN INFORMATION-THEORETIC UPPER BOUND ON PRIME GAPS

Aidan Rocke
aidanrocke@gmail.com

September 1, 2021

ABSTRACT

Within the setting of rare event modelling, the method of level sets allows us to define an equivalence relation over rare events with distinct rates of entropy production. As a measure of the efficacy of this method it is applied to Cramér's conjecture, an open problem in probabilistic number theory.

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 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 as rare events are assumed to be independent 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 $\{A_i\}_{i=1}^{\pi_x(n)} \subset \{A_i\}_{i=1}^{\pi_x(n+1)}$ where $\pi_x(n)$ counts the number of rare events in the time interval $[1,n]$ so it is monotonically increasing. 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:

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

which is a measure of expected surprise, or the expected information gained from observing a rare event.

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

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

Finally, (5) implies that:

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

so the entropy rates of distinct sequences $X_n, X'_n \in \mathcal{L}_{q_n}$ are asymptotically equivalent. From a data compression perspective, this tells us that in the asymptotic limit we need $\sim \ln\left(\frac{1}{q_n}\right)$ bits in order to compress a particular class of rare events.

One more interesting consequence is that 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 $N \in \mathbb{N}$:

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

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:

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

which we may call the log-zeta distribution in the sense that the un-normalised zeta distribution is given by:

$$\forall s > 1, \zeta(s) = \zeta_s \circ \mathbb{N} = \sum_{n\in\mathbb{N}} \frac{1}{n^s} \tag{12}$$

and if we log-transform the state-space $\mathbb{N}$ we have:

$$\forall s > 1, \zeta_s \circ \ln\left(\mathbb{N}\right) = \sum_{n\in\mathbb{N}} \frac{1}{(\ln n)^s} \tag{13}$$

where the logarithm transports us from the space of integers to the space of bits. It is worth adding that *prime encodings* are of great interest to rare-event modellers as they are not finite-state compressible. This follows from the fact that the prime numbers are empirically-distributed as if they were arranged uniformly, an immediate consequence of the information-theoretic derivation of the Prime Number Theorem [9].

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:

$$\int_2^N (\ln x)^{-a} dx = \sum_{n=2}^{N} \frac{1}{(\ln x_n)^a} \tag{14}$$

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:

$$\int_2^N (\ln x)^{-a} dx \sim \sum_{n=2}^{N} \frac{1}{(\ln n)^a} \sim N \cdot \frac{1}{(\ln N)^a} \tag{15}$$

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

$$q_N \sim \frac{1}{N} \sum_{n=2}^{N} \frac{1}{(\ln n)^a} \tag{16}$$

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

$$\forall z \sim U([1, n]), P(x_z = 1) \sim \frac{1}{n} \sum_{k=1}^{n} P(x_k = 1) \tag{17}$$

since $P(x_{n+1} = 1) \approx (n+1) \cdot q_{n+1} - n \cdot q_n = q_{n+1} + n \cdot (q_{n+1} - q_n) \sim q_{n+1}$.

1.3 The advantage of using the exponential of entropy

It is worth noting that the *entropy rate* $\frac{1}{q_n}$ corresponds to the exponential of entropy

$$\frac{1}{q_n} \approx e^{H(X_n)} \tag{18}$$

which has the advantage of being parametrisation invariant i.e. not depending on the choice of base for logarithms. This generally allows us to simplify calculations. Moreover, this is a robust measure for the expected number of observations that we need to collapse the entropy $H(X_n)$. Equivalently, it is a robust measure of the average distance between rare events.

On the other hand, the *typical probability* q_n corresponds to the density of rare events at the instant $n \in \mathbb{N}$. It may also be understood as the frequency with which we collapse the entropy $H(X_n)$ at the instant $n \in \mathbb{N}$. We may also observe that the typical probability q_n and entropy rate $\frac{1}{q_n}$ are both parametrisation invariant as they are multiplicative inverses of each other. Together they form the basis for the method of level sets which provides us with the necessary and sufficient apparatus for classifying sequences of rare events in terms of their rate of entropy production.

1.4 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. 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:

$$X \sim U([1, N]) \tag{19}$$

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:

$$S_c = \frac{\log_2(2^N)}{\pi(N)} = \frac{N}{\pi(N)} \tag{20}$$

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 (19). So we would have:

$$S_c = \frac{N}{\pi(N)} \sim \ln N \tag{21}$$

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:

$$\sum_{k=1}^{N-1} \frac{1}{k} \cdot |(k, k+1]| = \sum_{k=1}^{N-1} \frac{1}{k} \approx \ln N \tag{22}$$

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:

$$\frac{\pi(N)}{N} \sim \frac{1}{\ln N} \tag{23}$$

which is in complete agreement with the Prime Number Theorem. Moreover, 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:

$$\mathbb{E}[K(X_N)] \sim \pi(N) \cdot \ln N \sim N \tag{24}$$

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.

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 which dates back to Von Neumann essentially states that the Von Neumann entropy is invariant to Unitary transformations.

3 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:

$$\forall z \sim U([1,n]), P(z \in \mathbb{P}) \sim \frac{1}{\ln n} \tag{25}$$

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:

$$\forall A_p \in \{A_{p_k}\}_{k=1}^{\pi(\sqrt{n})}, A_p = \{z \in [1,n] : \gcd(p,z) = 1\} \bigcup \{p\} \tag{26}$$

and we'll note that $\forall z \sim U([1,n]), P(z \in A_{p \geq \sqrt{n}}) = 1 - \frac{1}{n} + \frac{1}{n} = 1$ so we have:

$$\forall z \sim U([1,n]), H(X_n) = -\ln \prod_{k=1}^{\pi(n)} P(z \in A_{p_k}) = -\ln \prod_{k=1}^{\pi(\sqrt{n})} P(z \in A_{p_k}) \tag{27}$$

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:

$$\forall z \sim U([1,n]), P(z \in A_{p_k}) = \big(1 - \frac{1}{p_k}\big) + \frac{1}{n} \approx \big(1 - \frac{1}{p_k}\big) \tag{28}$$

which allows us to make (27) precise:

$$H(X_n) = -\ln \prod_{p \leq \sqrt{n}} \big(1 - \frac{1}{p}\big) \tag{29}$$

Using Mertens' third theorem we have:

$$\prod_{p \leq \sqrt{n}} \big(1 - \frac{1}{p}\big) \approx \frac{e^{-\gamma}}{\frac{1}{2} \cdot \ln n} \approx \frac{0.9}{\ln n} \tag{30}$$

so we may infer that:

$$H(X_n) \sim \ln \ln n \tag{31}$$

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

$$\mathcal{L}_{q_n} = \{X_n \in \{0,1\}^\infty : H(X_n) \sim \ln \ln n\} \tag{32}$$

From (32), we may deduce (25):

$$\forall z \sim U([1,n]), q_n = P(z \in \mathbb{P}) \sim \frac{1}{\ln n} \tag{33}$$

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:

$$\pi(n) \sim \int_2^n \frac{1}{\ln x} dx \sim \frac{n}{\ln n} \tag{34}$$

in complete agreement with the Prime Number Theorem.

4 Application to Cramér's random model, Part II

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

$$\sum_{k=1}^{G_n} P(x_{p_n+k} = 1 \wedge p_{n+1} - p_n = k) = 1 \tag{35}$$

where $G_n < p_n$ due to Bertrand's postulate and $P(p_{n+1} - p_n = k) = 1$ 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 (34) to consider its components:

$$\forall k \sim U([1, G_n]), P(p_{n+1} = p_n + k) = P(p_n = p_{n+1} - k) \sim \frac{1}{\ln p_n} \tag{36}$$

$$\forall k \sim U([1, G_n]), P(x_{p_n+k} = 1) \sim \frac{1}{\ln p_n} \tag{37}$$

where

$$P(x_{p_n+k} = 1 \wedge p_{n+1} - p_n = k) \geq P(x_{p_n+k} = 1) \cdot P(p_{n+1} = p_n + k) \tag{38}$$

Using (36),(37) and (38) we may derive the lower bound:

$$\forall k \sim U([1, G_n]), P(x_{p_n+k} = 1 \wedge p_{n+1} - p_n = k) \gtrsim \frac{1}{(\ln p_n)^2} \tag{39}$$

which implies:

$$\frac{1}{(\ln p_n)^2} \lesssim \frac{1}{G_n} \sum_{k=1}^{G_n} P(x_{p_n+k} = 1 \wedge p_{n+1} - p_n = k) = \frac{1}{G_n} \tag{40}$$

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

$$p_{n+1} - p_n = \mathcal{O}((\ln p_n)^2) \tag{41}$$

as conjectured by Harald Cramér in 1936 [1].

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

If we define,

$$\Lambda_k := \text{set of } \binom{\pi(\sqrt{n})}{k} \text{ distinct subsets of } \{p_k\}_{k=1}^{\pi(\sqrt{n})} \tag{42}$$

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

$$\Big| \prod_{p \leq \sqrt{n}} (1 - \frac{1}{p} + \frac{1}{n}) - \prod_{p \leq \sqrt{n}} (1 - \frac{1}{p}) \Big| \leq \sum_{k=1}^{\pi(\sqrt{n})-1} \frac{1}{n^k} \sum_{\lambda \in \Lambda_k} \prod_{p \in \lambda} (1 - \frac{1}{p}) + \frac{1}{n^{\pi(\sqrt{n})}} \tag{43}$$

and given that:

$$\pi(\sqrt{n}) < \sqrt{n} \implies \frac{1}{n^k} \cdot \binom{\pi(\sqrt{n})}{k} \leq \frac{1}{n^{k/2}} \tag{44}$$

the absolute error satisfies the following inequality:

$$\Delta_n = \Big| \prod_{p \leq \sqrt{n}} (1 - \frac{1}{p} + \frac{1}{n}) - \prod_{p \leq \sqrt{n}} (1 - \frac{1}{p}) \Big| \leq \frac{1}{\sqrt{n}} + \frac{1}{n} \tag{45}$$

so the relative error converges to zero:

$$\lim_{n \to \infty} \frac{\Delta_n}{\prod_{p \leq \sqrt{n}} (1 - \frac{1}{p})} \leq \lim_{n \to \infty} (\frac{\ln n}{\sqrt{n}} + \frac{\ln n}{n}) = 0 \tag{46}$$

which provides us with the necessary justifications in our derivation of Cramér's random model, specifically formulas (29),(30) and (31).

References

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

A An invariance theorem for algorithmically random data

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

$$x_n \in \{0,1\}, x_{n+1} = \varphi \circ x_{1:n} \tag{47}$$

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 $\hat{f} \in F_\theta$:

$$\exists k \in [1, n-1], x_{n+1} = \hat{f} \circ x_{n-k:n} \Rightarrow \delta_{\hat{f}(x_{n-k:n}),x_{n+1}} = 1 \tag{48}$$

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

$$\hat{f} = \max_{f \in F_\theta} \frac{1}{N-k} \sum_{n=k+1}^{N} \delta_{f(x_{n-k:n}),x_{n+1}} \tag{49}$$

has an expected performance:

$$\frac{1}{N-k} \sum_{n=N+k+1}^{2N-1} \delta_{\hat{f}(x_{n-k:n}),x_{n+1}} \leq \frac{1}{2} \tag{50}$$

Furthermore, if the dataset is imbalanced i.e. $\frac{1}{N}\sum_{i=1}^{N} x_i \neq \frac{1}{2}$ then we may generalise this result by introducing the auxiliary definitions:

$$y_n = x_{n+1} \tag{51}$$

$$\hat{y_n} = \hat{f} \circ x_{n-k:n} \tag{52}$$

$$\beta_n = \delta_{y_n,\hat{y_n}} \tag{53}$$

$$N_1 = \sum_{n=N+k+1}^{2N} \delta_{y_n,1} \tag{54}$$

$$N_0 = \sum_{n=N+k+1}^{2N} \delta_{y_n,0} \tag{55}$$

and so for large N, we have:

$$\mathcal{L}_N[\hat{f}] = \min\Big[\frac{1}{N_0} \sum_{n=N+k+1}^{2N-1} \delta_{y_n,0} \cdot \beta_n, \frac{1}{N_1} \sum_{n=N+k+1}^{2N-1} \delta_{y_n,1} \cdot \beta_n\Big] \leq \frac{1}{2} \tag{56}$$

and therefore:

$$\forall \hat{f} \in F_\theta, \lim_{N\to\infty} P(\mathcal{L}_N[\hat{f}] > \frac{1}{2}) = 0 \tag{57}$$

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.