HAUSDORFF DIMENSION OF THE SET OF NUMBERS WITH A GIVEN FREQUENCY OF DIGITS

MANJUNATH KRISHNAPUR

For $x \in [0,1]$, let x_i be the digits in its binary expansion, that is, $x = \sum_{k=1}^{\infty} \frac{x_k}{2^k}$. Then, for $p \in [0,1]$, define

$$E_p = \left\{ x \in [0,1] \ : \ \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} x_j = p \right\}.$$

If x is chosen from Lebesgue measure on $[0,1]$, then x_k are i.i.d. Bernoulli(1/2) random variables. Hence,

$$\text{Leb}(E_p) = \begin{cases} 1 & \text{if } p = \frac{1}{2}. \\ 0 & \text{otherwise.} \end{cases}$$

Thus Lebesgue measure does not distinguish between E_p for $p \neq \frac{1}{2}$. As we shall see, Hausdorff dimension does! Below, we write q for $1-p$.

Proposition 1. *$dim_H(E_p) = H(p)$, where $H(p) = -p\log_2 p - q\log_2 q$.*

Then key reason for the appearance of $H(p)$ is the following.

Lemma 2. *For $p \le \frac{1}{2}$ and integer $n \ge 1$, define*

$$T_n^p = \{(z_1,\dots,z_n) \in \{0,1\}^n \ : \ z_1 + \dots + z_n = np\}.$$

$$S_n^p = \{(z_1,\dots,z_n) \in \{0,1\}^n \ : \ z_1 + \dots + z_n \le np\}.$$

Then, as $n \to \infty$,

(1) $\frac{1}{n}\log_2 \#T_n(p) \to H(p)$.
(2) $\frac{1}{n}\log_2 \#S_n(p) \to H(p)$.

Proof. To prove the first claim, note that $\#T_n = \binom{n}{np}$ and by Stirling's approximation,

$$\binom{n}{np} = \frac{n!}{(np)!(nq)!} \sim \frac{n^{n+\frac{1}{2}}e^{-n}\sqrt{2\pi}}{(np)^{np+\frac{1}{2}}e^{-np}\sqrt{2\pi}(nq)^{nq+\frac{1}{2}}e^{-nq}\sqrt{2\pi}} = \frac{1}{\sqrt{2\pi npq}}\, p^{-np}q^{-nq}$$

where the "$\sim$" in the middle means that the ratio of the left and right hand sides converges to 1 as $n \to \infty$. Therefore, taking logarithms and dividing by n, we get

$$\frac{\log_2 \#T_n}{n} = \frac{-np\log_2 p - nq\log_2 q + O(\log n)}{n} \to H(p).$$

Now the second claim follows by observing that $\#T_n = \sum_{k=0}^{np} \binom{n}{k}$ and hence

$$\#T_n \leq \#S_n \leq np\,\#T_n$$

where the first inequality holds because $T_n \subset S_n$ and the second hold because $\binom{n}{k}$ increases in k as k increases from 0 up to $\frac{n}{2}$ (and we have assumed that $np \leq \frac{n}{2}$). ☐

Proof of Proposition 1. **Upper bound** For $p < \frac{1}{2}$, let $\tilde{E}_p = \{x \ : \ \limsup_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} x_j < p\}$ so that $\tilde{E}_p = \cup_{N\geq 1} \tilde{E}_p(N)$ where

$$\tilde{E}_p(N) = \left\{ x \ : \ \frac{1}{n} \sum_{j=1}^{n} x_j < p \text{ for all } n \geq N \right\}.$$

Then we claim that $\dim_H(E_p(N)) \leq H(p)$. Once we show this, it follows that $\dim_H(\tilde{E}_p) = \sup_N \dim_H(\tilde{E}_p(N)) \leq H(p)$ and since for any $p' > p$, $E_p \subset \tilde{E}_{p'}$, we get $\dim_H(E_p) \leq H(p')$ for any $p' \geq p$. As $p \to H(p)$ is continuous, letting p' decrease to p we get $\dim_H(E_p) \leq H(p)$ as required to show.

Now fix $N \geq 1$, $p \leq \frac{1}{2}$, and consider $\tilde{E}_p(N)$. Take any $\delta > 0$, and let n be such that $2^{-n} \leq \delta < 2^{-n+1}$. Then for each $\mathbf{z} \in S_n^p$, define the set $A_{\mathbf{z}} = \{x \ : \ x_j = z_j \text{ for } j \leq n\}$. Then, each $|A_{\mathbf{z}}| = 2^{-n} \leq \delta$, and further $\cup_{\mathbf{z}\in S_n^p} A_z$ covers $\tilde{E}_p(N)$ provided $n \geq N$ (and the latter holds if δ is small enough).

Thus for small enough δ, we have $N_\delta(\tilde{E}_p(N)) \leq \#S_n^p$, and therefore

$$\dim_H(\tilde{E}_p(N)) \leq \underline{\dim}_M(\tilde{E}_p(N)) \leq \lim_{n\to\infty} \frac{\log \#S_n^p}{n \log 2} = H(p)$$

the last inequality because of Lemma 2.

Lower bound To get a lower bound we fix a large integer M and consider the set T_M defined in Lemma 2. Then, let $Y_1, Y_2, \ldots$ be i.i.d. T_M-valued random variables with $P[Y_1 = \mathbf{z}] = \frac{1}{\#T_M}$ for each $\mathbf{z} \in T_M$. Write $Y_j = (z_{j,1}, \ldots, z_{j,M})$ and set

$$X = \sum_{j=1}^{\infty} \frac{1}{2^{(j-1)M}} \sum_{k=1}^{M} \frac{z_{j,k}}{2^k}$$

be the number with binary digits $z_{1,1}, \ldots, z_{1,M}, z_{2,1}, \ldots, z_{2,M}, \ldots$. Let μ be the law of X. It is easy to see that $\mu(E_p) = 1$ (why?). Then, for $\alpha \geq 0$, we have

$$\mathcal{E}_\alpha(\mu) = \mathbf{E}[|X - X'|^{-\alpha}]$$

where X, X' are i.i.d with distribution μ. Let X be made up from $Y_j = (z_{j,1}, \ldots, z_{j,M})$ and X' be made up from $Y'_j = (z'_{j,1}, \ldots, z'_{j,M})$. Define $L = \min\{k \ : \ Y_k \neq Y'_k\}$. Then (at least) the first $(L-1)M$ digits of X and X' coincide, and hence, $|X - X'| \leq 2^{-LM+M}$. We shall get a lower bound for $|X - X'|$ of the same order as follows.

Let $y_* = \sum_{j=Mq+1}^{M} \frac{1}{2^j}$ and let $y^* = \sum_{j=1}^{Mp} \frac{1}{2^j}$. Then, it is easy to see that for any $\mathbf{z} \in T_M$, we have

$$y_* \leq \sum_{k=1}^{M} \frac{z_k}{2^k} \leq y^*.$$

Returning to $|X - X'|$, without loss of generality suppose that $X > X'$. Write

$$X - X' = \sum_{j=L}^{\infty} \frac{1}{2^{(j-1)L}} \sum_{k=1}^{M} \frac{z_{j,k} - z'_{j,k}}{2^k}.$$

Note that the terms with $j < L$ cancel. Since we assume that $X > X'$, we have

$$\frac{z_{L,k} - z'_{L,k}}{2^k} \geq \frac{1}{2^M}.$$

On the other hand, for $j \geq L + 1$, we have

$$\sum_{k=1}^{M} \frac{z_{j,k}}{2^k} \geq y_* \text{ and } \sum_{k=1}^{M} \frac{z'_{j,k}}{2^k} \leq y^*$$

which implies that

$$\sum_{k=1}^{M} \frac{z_{j,k} - z'_{j,k}}{2^k} \geq -(y^* - y_*).$$

Putting everything together, we get

$$X - X' \geq \frac{1}{2^{LM}} - \frac{y^* - y_*}{2^{LM}} \sum_{j=1}^{\infty} \frac{1}{2^{jM}} \geq \frac{1}{2^{LM}} \left(1 - (y^* - y_*)\right) = C 2^{-LM}$$

where $C = 1 - (y^* - y_*) > 0$. Thus

$$\mathbf{E}[|X - X'|^{-\alpha}] \leq C\, \mathbf{E}[2^{\alpha LM}].$$

Now, $\mathbf{P}(Y_1 = Y_1') = \frac{1}{\#T_M}$, and hence $\mathbf{P}[L = \ell] = \frac{1 - \frac{1}{\#T_M}}{(\#T_M)^{\ell - 1}}$. From this it is immediate that $\mathbf{E}[2^{\alpha LM}]$ is finite whenever $\alpha < \frac{1}{M} \log_2 \#T_M$.

Thus, for any M, we see that $\dim_{\mathcal{E}}(E_p) \geq \frac{1}{M} \log_2 \#T_M$. Let $M \to \infty$ and apply Lemma 2 to conclude that $\dim_H(E_p) \geq \dim_{\mathcal{E}}(E_p) \geq H(p)$. □

Remark 3. (1) There is nothing special about base 2. For instance, if for any p_1, p_2, p_3 with $p_1 + p_2 + p_3 = 1$, we set

$$E_{p_0,p_1,p_2} = \left\{ x \in [0,1] \ : \ \lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} \delta_{x_j - k} = p_k \text{ for } k = 0, 1, 2 \right\}$$

to be the set of all $x \in [0,1]$ whose base three expansion has a proportion p_0 of zeros, a proportion p_1 of ones and a proportion p_2 of twos, then the same proof method shows that $\dim_H(E_{p_0,p_1,p_2}) = -\sum_{k=0}^{2} p_k \log_3 p_k$.

(2) This is the simplest example of what is called 'mutifractal decomposition'. The interval $[0,1]$ is divided into sets parameterized by p, and for different p, the resulting sets have different Hausdorff dimensions.

(3) $H(p)$ is what is called the entropy of the Bernulli(p) measure. More generally if μ is a discrete measure with $\mu\{x_i\} = p_i$ with $\sum p_i = 1$ and x_is are distinct, then the entropy of μ is defined to be $-\sum p_i \log_2 p_i$.