Chapter 40

Entropy, Randomness, and Information

By Sariel Har-Peled, March 19, 2024[1]

> "If only once - only once - no matter where, no matter before what audience - I could better the record of the great Rastelli and juggle with thirteen balls, instead of my usual twelve, I would feel that I had truly accomplished something for my country. But I am not getting any younger, and although I am still at the peak of my powers there are moments - why deny it? - when I begin to doubt - and there is a time limit on all of us."
>
> Romain Gary, The talent scout

40.1. The entropy function

Definition 40.1.1. The ***entropy*** in bits of a discrete random variable X is given by

$$\mathbb{H}(X) = -\sum_x \mathbb{P}[X = x] \lg \mathbb{P}[X = x],$$

where $\lg x$ is the logarithm base 2 of x. Equivalently, $\mathbb{H}(X) = \mathbb{E}\left[\lg \frac{1}{\mathbb{P}[X]}\right]$.

The ***binary entropy*** function $\mathbb{H}(p)$ for a random binary variable that is 1 with probability p, is

$$\mathbb{H}(p) = -p \lg p - (1-p)\lg(1-p).$$

We define $\mathbb{H}(0) = \mathbb{H}(1) = 0$.

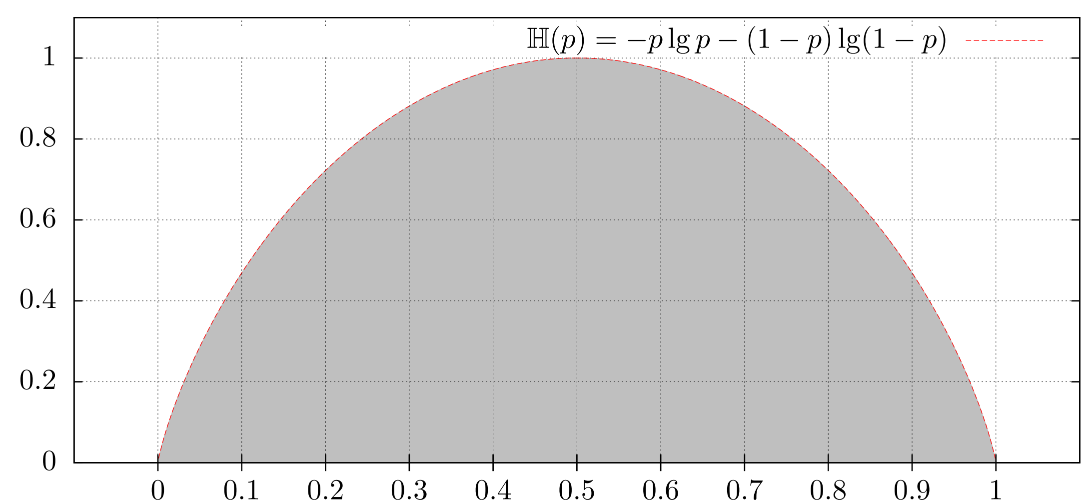


Figure 40.1: The binary entropy function.

The function $\mathbb{H}(p)$ is a concave symmetric around $1/2$ on the interval $[0, 1]$ and achieves its maximum at $1/2$. For a concrete example, consider $\mathbb{H}(3/4) \approx 0.8113$ and $\mathbb{H}(7/8) \approx 0.5436$. Namely, a coin that has $3/4$ probably to be heads have higher amount of "randomness" in it than a coin that has probability $7/8$ for heads.

Writing $\lg n = (\ln n)/\ln 2$, we have that

$$\mathbb{H}(p) = \frac{1}{\ln 2}(-p \ln p - (1-p)\ln(1-p))$$
$$\text{and} \quad \mathbb{H}'(p) = \frac{1}{\ln 2}\left(-\ln p - \frac{p}{p} - (-1)\ln(1-p) - \frac{1-p}{1-p}(-1)\right) = \lg \frac{1-p}{p}.$$

Deploying our amazing ability to compute derivative of simple functions once more, we get that

$$\mathbb{H}''(p) = \frac{1}{\ln 2}\frac{p}{1-p}\left(\frac{p(-1)-(1-p)}{p^2}\right) = -\frac{1}{p(1-p)\ln 2}.$$

Since $\ln 2 \approx 0.693$, we have that $\mathbb{H}''(p) \leq 0$, for all $p \in (0,1)$, and the $\mathbb{H}(\cdot)$ is concave in this range. Also, $\mathbb{H}'(1/2) = 0$, which implies that $\mathbb{H}(1/2) = 1$ is a maximum of the binary entropy. Namely, a balanced coin has the largest amount of randomness in it.

Example 40.1.2. A random variable X that has probability $1/n$ to be i, for $i = 1, \ldots, n$, has entropy $\mathbb{H}(X) = -\sum_{i=1}^{n} \frac{1}{n} \lg \frac{1}{n} = \lg n$.

Note, that the entropy is oblivious to the exact values that the random variable can have, and it is sensitive only to the probability distribution. Thus, a random variables that accepts $-1, +1$ with equal probability has the same entropy (i.e., 1) as a fair coin.

Lemma 40.1.3. *Let X and Y be two independent random variables, and let Z be the random variable (X, T). Then $\mathbb{H}(Z) = \mathbb{H}(X) + \mathbb{H}(Y)$.*

Proof: In the following, summation are over all possible values that the variables can have. By the independence of X and Y we have

$$\begin{aligned}
\mathbb{H}(Z) &= \sum_{x,y} \mathbb{P}[(X,Y) = (x,y)] \lg \frac{1}{\mathbb{P}[(X,Y) = (x,y)]} \\
&= \sum_{x,y} \mathbb{P}[X = x]\,\mathbb{P}[Y = y] \lg \frac{1}{\mathbb{P}[X = x]\,\mathbb{P}[Y = y]} \\
&= \sum_{x}\sum_{y} \mathbb{P}[X = x]\,\mathbb{P}[Y = y] \lg \frac{1}{\mathbb{P}[X = x]} \\
&\quad + \sum_{y}\sum_{x} \mathbb{P}[X = x]\,\mathbb{P}[Y = y] \lg \frac{1}{\mathbb{P}[Y = y]} \\
&= \sum_{x} \mathbb{P}[X = x] \lg \frac{1}{\mathbb{P}[X = x]} + \sum_{y} \mathbb{P}[Y = y] \lg \frac{1}{\mathbb{P}[Y = y]} = \mathbb{H}(X) + \mathbb{H}(Y).
\end{aligned}$$

■

Lemma 40.1.4. *Suppose that nq is integer in the range $[0, n]$. Then* $\dfrac{2^{n\mathbb{H}(q)}}{n+1} \leq \dbinom{n}{nq} \leq 2^{n\mathbb{H}(q)}$.

Proof: This trivially holds if $q = 0$ or $q = 1$, so assume $0 < q < 1$. We know that

$$\binom{n}{nq} q^{nq}(1-q)^{n-nq} \leq (q + (1-q))^n = 1$$
$$\Longrightarrow \quad \binom{n}{nq} \leq q^{-nq}(1-q)^{-n(1-q)} = 2^{n\left(-q\lg q - (1-q)\lg(1-q)\right)} = 2^{n\mathbb{H}(q)}.$$

As for the other direction, let

$$\mu(k) = \binom{n}{k} q^k (1-q)^{n-k}.$$

The claim is that $\mu(nq)$ is the largest term in $\sum_{k=0}^{n} \mu(k) = 1$, where $\mu(k) = \binom{n}{k} q^k (1-q)^{n-k}$. Indeed,

$$\Delta_k = \mu(k) - \mu(k+1) = \binom{n}{k} q^k (1-q)^{n-k} \left(1 - \frac{n-k}{k+1} \frac{q}{1-q}\right),$$

and the sign of this quantity is the sign of $(k+1)(1-q) - (n-k)q = k+1-kq-q-nq+kq = 1+k-q-nq$. Namely, $\Delta_k \geq 0$ when $k \geq nq + q - 1$, and $\Delta_k < 0$ otherwise. Namely, $\mu(k) < \mu(k+1)$, for $k < nq$, and $\mu(k) \geq \mu(k+1)$ for $k \geq nq$. Namely, $\mu(nq)$ is the largest term in $\sum_{k=0}^{n} \mu(k) = 1$, and as such it is larger than the average. We have $\mu(nq) = \binom{n}{nq} q^{nq} (1-q)^{n-nq} \geq \frac{1}{n+1}$, which implies

$$\binom{n}{nq} \geq \frac{1}{n+1} q^{-nq} (1-q)^{-(n-nq)} = \frac{1}{n+1} 2^{n\mathbb{H}(q)}. \qquad \blacksquare$$

Lemma 40.1.4 can be extended to handle non-integer values of q. This is straightforward, and we omit the easy details.

Corollary 40.1.5. *We have:*

(i) $q \in [0, 1/2] \Rightarrow \binom{n}{\lfloor nq \rfloor} \leq 2^{n\mathbb{H}(q)}$.

(ii) $q \in [1/2, 1] \Rightarrow \binom{n}{\lceil nq \rceil} \leq 2^{n\mathbb{H}(q)}$.

(iii) $q \in [1/2, 1] \Rightarrow \frac{2^{n\mathbb{H}(q)}}{n+1} \leq \binom{n}{\lfloor nq \rfloor}$.

(iv) $q \in [0, 1/2] \Rightarrow \frac{2^{n\mathbb{H}(q)}}{n+1} \leq \binom{n}{\lceil nq \rceil}$.

The bounds of Lemma 40.1.4 and Corollary 40.1.5 are loose but sufficient for our purposes. As a sanity check, consider the case when we generate a sequence of n bits using a coin with probability q for head, then by the Chernoff inequality, we will get roughly nq heads in this sequence. As such, the generated sequence Y belongs to $\binom{n}{nq} \approx 2^{n\mathbb{H}(q)}$ possible sequences that have similar probability. As such, $\mathbb{H}(Y) \approx \lg \binom{n}{nq} = n\mathbb{H}(q)$, by Example 40.1.2, this also readily follows from Lemma 40.1.3.

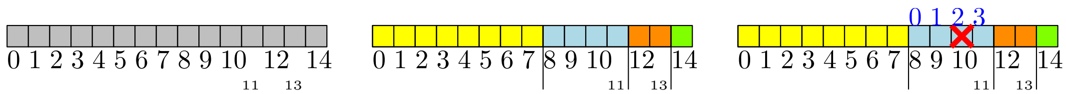
40.2. Extracting randomness

The problem. We are given a random variable X that is chosen uniformly at random from $[\![0 : m-1]\!] = \{0, \ldots, m-1\}$. Our purpose is built an algorithm that given X output a binary string, such that the bits in the binary string can be interpreted as the coin flips of a fair balanced coin. That is, the probability of the ith bit of the output (if it exists) to be 0 (or 1) is exactly half, and the different bits of the output are independent.

Idea. We break the $[\![0 : m-1]\!]$ into consecutive blocks that are powers of two. Given the value of X, we find which block contains it, and we output a binary representation of the location of X in the block containing it, where if a block is length 2^k, then we output k bits.

Entropy can be interpreted as the amount of unbiased random coin flips can be extracted from a random variable.

Definition 40.2.1. An ***extraction function*** `Ext` takes as input the value of a random variable X and outputs a sequence of bits y, such that $\mathbb{P}\left[\mathtt{Ext}(X) = y \,\middle|\, |y| = k\right] = 1/2^k$. whenever $\mathbb{P}[|y| = k] \geq 0$, where $|y|$ denotes the length of y.



(A) (B) (C)

Figure 40.2: (A) $m = 15$. (B) The block decomposition. (C) If $X = 10$, then the extraction output is 2 in base 2, using 2 bits – that is `10`.

As a concrete (easy) example, consider X to be a uniform random integer variable out of $0, \ldots, 7$. All that $\texttt{Ext}(x)$ has to do in this case, is just to compute the binary representation of x.

The definition of the extraction function has two subtleties:

(A) It requires that all extracted sequences of the same length (say k), have the same probability to be output (i.e., $1/2^k$).

(B) If the extraction function can output a sequence of length k, then it needs to be able to output *all* 2^k such binary sequences.

Thus, for X a uniform random integer variable in the range $0, \ldots, 11$, the function $\texttt{Ext}(x)$ can output the binary representation for x if $0 \leq x \leq 7$. However, what do we do if x is between 8 and 11? The idea is to output the binary representation of $x - 8$ as a two bit number. Clearly, Definition 40.2.1 holds for this extraction function, since $\mathbb{P}\Big[\texttt{Ext}(X) = 00 \;\Big|\; |\texttt{Ext}(X)| = 2\Big] = 1/4$. as required. This scheme can be of course extracted for any range.

Tedium 40.2.2. For $x \leq y$ positive integers, and any positive integer Δ, we have that

$$\frac{x}{y} \leq \frac{x + \Delta}{y + \Delta} \iff x(y + \Delta) \leq y(x + \Delta) \iff x\Delta \leq y\Delta \iff x \leq y.$$

Theorem 40.2.3. *Suppose that the value of a random variable X is chosen uniformly at random from the integers $\{0, \ldots, m - 1\}$. Then there is an extraction function for X that outputs on average (i.e., in expectation) at least $\lfloor \lg m \rfloor - 1 = \lfloor \mathbb{H}(X) \rfloor - 1$ independent and unbiased bits.*

Proof: We represent m as a sum of unique powers of 2, namely $m = \sum_i a_i 2^i$, where $a_i \in \{0, 1\}$. Thus, we decomposed $\{0, \ldots, m - 1\}$ into a disjoint union of blocks that have sizes which are distinct powers of 2. If a number falls inside such a block, we output its relative location in the block, using binary representation of the appropriate length (i.e., k if the block is of size 2^k). It is not difficult to verify that this function fulfills the conditions of Definition 40.2.1, and it is thus an extraction function.

Now, observe that the claim holds if m is a power of two, by Example 40.1.2 (i.e., if $m = 2^k$, then $\mathbb{H}(X) = k$). Thus, if m is not a power of 2, then in the decomposition if there is a block of size 2^k, and the X falls inside this block, then the entropy is k.

The remainder of the proof is by induction – assume the claim holds if the range used by the random variable is strictly smaller than m. In particular, let $K = 2^k$ be the largest power of 2 that is smaller than m, and let $U = 2^u$ be the largest power of two such that $U \leq m - K \leq 2U$.

If the random number $X \in [\![0 : K - 1]\!]$, then the scheme outputs k bits. Otherwise, we can think about the extraction function as being recursive and extracting randomness from a random variable $X' = X - K$ that is uniformly distributed in $[\![0 : m - K]\!]$.

By Tedium 40.2.2, we have that

$$\frac{m - K}{m} \leq \frac{m - K + (2U + K - m)}{m + (2U + K - m)} = \frac{2U}{2U + K}$$

Let Y be the random variable which is the number of random bits extracted. We have that

$$\begin{aligned}\mathbb{E}[Y] &\geq \frac{K}{m}k + \frac{m-K}{m}(\lfloor \lg(m-K) \rfloor - 1) = k - \frac{m-K}{m}k + \frac{m-K}{m}(u-1) = k + \frac{m-K}{m}\overbrace{(u-k-1)}^{<0} \\ &\geq k - \frac{2U}{2U+K}(u-k-1) = k - \frac{2U}{2U+K}(1+k-u).\end{aligned}$$

If $u = k-1$, then $\mathbb{H}(X) \geq k - \frac{1}{2} \cdot 2 = k-1$, as required. If $u = k-2$ then $\mathbb{H}(X) \geq k - \frac{1}{3} \cdot 3 = k-1$. Finally, if $u < k-2$ then

$$\mathbb{E}[Y] \geq k - \frac{2U}{2U+K}(1+k-u) \geq k - \frac{2U}{K}(1+k-u) = k - \frac{k-u+1}{2^{(k-u+1)-2}} \geq k-1,$$

since $k - u + 1 \geq 4$ and $i/2^{i-2} \leq 1$ for $i \geq 4$. ■

Theorem 40.2.4. *Consider a coin that comes up heads with probability $p > 1/2$. For any constant $\delta > 0$ and for n sufficiently large:*

(A) One can extract, from an input of a sequence of n flips, an output sequence of $(1-\delta)n\mathbb{H}(p)$ (unbiased) independent random bits.

(B) One can not extract more than $n\mathbb{H}(p)$ bits from such a sequence.

Proof: There are $\binom{n}{j}$ input sequences with exactly j heads, and each has probability $p^j(1-p)^{n-j}$. We map this sequence to the corresponding number in the set $\left\{0, \ldots, \binom{n}{j} - 1\right\}$. Note, that this, conditional distribution on j, is uniform on this set, and we can apply the extraction algorithm of Theorem 40.2.3. Let Z be the random variables which is the number of heads in the input, and let B be the number of random bits extracted. We have

$$\mathbb{E}[B] = \sum_{k=0}^{n} \mathbb{P}[Z = k]\, \mathbb{E}\Big[B \,\Big|\, Z = k\Big],$$

and by Theorem 40.2.3, we have $\mathbb{E}\Big[B \,\Big|\, Z = k\Big] \geq \left\lfloor \lg \binom{n}{k} \right\rfloor - 1$. Let $\varepsilon < p - 1/2$ be a constant to be determined shortly. For $n(p-\varepsilon) \leq k \leq n(p+\varepsilon)$, we have

$$\binom{n}{k} \geq \binom{n}{\lfloor n(p+\varepsilon) \rfloor} \geq \frac{2^{n\mathbb{H}(p+\varepsilon)}}{n+1},$$

by Corollary 40.1.5 (iii). We have

$$\begin{aligned}\mathbb{E}[B] &\geq \sum_{k=\lfloor n(p-\varepsilon)\rfloor}^{\lceil n(p-\varepsilon)\rceil} \mathbb{P}[Z=k]\, \mathbb{E}\Big[B \,\Big|\, Z=k\Big] \geq \sum_{k=\lfloor n(p-\varepsilon)\rfloor}^{\lceil n(p-\varepsilon)\rceil} \mathbb{P}[Z=k]\left(\left\lfloor \lg \binom{n}{k} \right\rfloor - 1\right) \\ &\geq \sum_{k=\lfloor n(p-\varepsilon)\rfloor}^{\lceil n(p-\varepsilon)\rceil} \mathbb{P}[Z=k]\left(\lg \frac{2^{n\mathbb{H}(p+\varepsilon)}}{n+1} - 2\right) \\ &= (n\mathbb{H}(p+\varepsilon) - \lg(n+1))\, \mathbb{P}[|Z - np| \leq \varepsilon n] \\ &\geq (n\mathbb{H}(p+\varepsilon) - \lg(n+1))\left(1 - 2\exp\left(-\frac{n\varepsilon^2}{4p}\right)\right),\end{aligned}$$

since $\mu = \mathbb{E}[Z] = np$ and $\mathbb{P}\Big[|Z - np| \geq \frac{\varepsilon}{p}pn\Big] \leq 2\exp\left(-\frac{np}{4}\left(\frac{\varepsilon}{p}\right)^2\right) = 2\exp\left(-\frac{n\varepsilon^2}{4p}\right)$, by the Chernoff inequality. In particular, fix $\varepsilon > 0$, such that $\mathbb{H}(p+\varepsilon) > (1-\delta/4)\mathbb{H}(p)$, and since p is fixed $n\mathbb{H}(p) = \Omega(n)$, in particular, for

n sufficiently large, we have $-\lg(n+1) \geq -\frac{\delta}{10} n\mathbb{H}(p)$. Also, for n sufficiently large, we have $2\exp\left(-\frac{n\varepsilon^2}{4p}\right) \leq \frac{\delta}{10}$. Putting it together, we have that for n large enough, we have

$$\mathbb{E}[B] \geq \left(1 - \frac{\delta}{4} - \frac{\delta}{10}\right) n\mathbb{H}(p)\left(1 - \frac{\delta}{10}\right) \geq (1-\delta)n\mathbb{H}(p),$$

as claimed.

As for the upper bound, observe that if an input sequence x has probability q, then the output sequence $y = \mathtt{Ext}(x)$ has probability to be generated which is at least q. Now, all sequences of length $|y|$ have equal probability to be generated. Thus, we have the following (trivial) inequality $2^{|\mathtt{Ext}(x)|} q \leq 2^{|\mathtt{Ext}(x)|} \mathbb{P}[y = \mathtt{Ext}(X)] \leq 1$, implying that $|\mathtt{Ext}(x)| \leq \lg(1/q)$. Thus,

$$\mathbb{E}[B] = \sum_x \mathbb{P}[X = x] \, |\mathtt{Ext}(x)| \leq \sum_x \mathbb{P}[X = x] \lg \frac{1}{\mathbb{P}[X = x]} = \mathbb{H}(X). \qquad \blacksquare$$

40.3. Bibliographical Notes

The presentation here follows [MU05, Sec. 9.1-Sec 9.3].

References

[MU05] M. Mitzenmacher and U. Upfal. *Probability and Computing – randomized algorithms and probabilistic analysis*. Cambridge, 2005.