Information Theory. Sanov's Theorem.

Ivan Matic

July 21, 2018

We will introduce two problems from information theory related to data compression and message transmission. The entropy of a probability distribution will naturally arise in these contexts. We will see how the entropy relates to large deviation events that can be used to design successful compression algorithms.

Data Compression

All three images from Figure 1 below are of format 1200×2320. However, when saved in Portable Network Graphics format, their files have sizes of 2.9MB, 3.5MB, and 6.7MB, in increasing order.

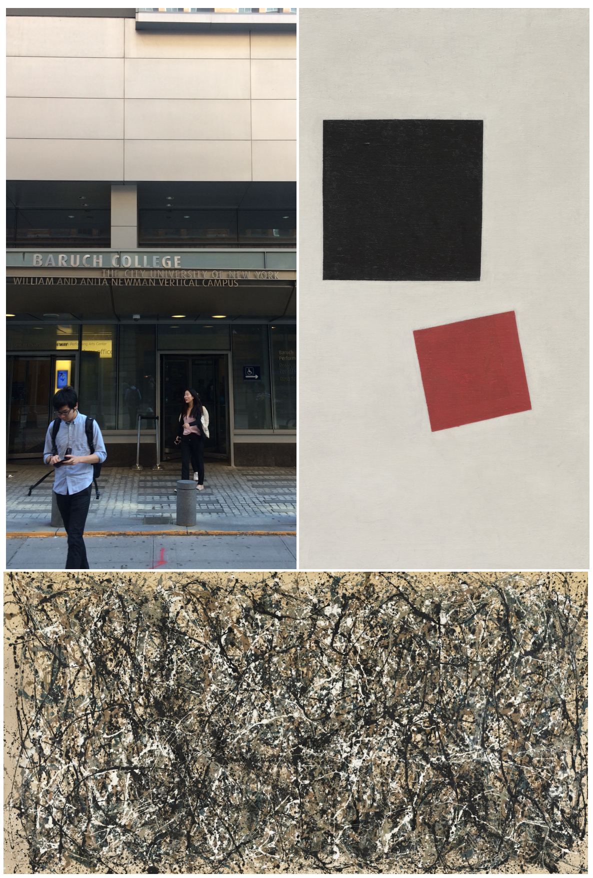

Figure 1: Three images: 1) A smartphone, *A Photograph of Baruch College*; 2) Kazimir Malevich, *Painterly Realism of a Boy with a Knapsack - Color Masses in the Fourth Dimension*; 3) Jackson Pollock, *One: Number 31, 1950*

Very few will be surprised to learn that the painting of "a Boy with a Kanpsack" is the one that takes the fewest bits in memory. And yes, Pollock is the worst.

Let us imagine that we want to store some big quantity of data, such as a picture, a text document, or a book. First, we have to start with an un-compressed data, identify its smallest building block, and called it *symbol*. The set of all symbols is *Alphabet*.

We will denote the alphabet by $\mathcal{A}$. Our raw, un-compressed data can be modeled as an element of the set $\mathcal{A}^N$ and our goal is to find a bijection between elements of sets $\mathcal{A}^N$ and binary sequences in such a way that short sequences correspond to data that is encountered more often in real life.

While it is obvious how to visualize the alphabet when we are talking about books, it is worth pointing out that we have several options of doing so.

Let us talk about pictures. One alphabet is created when each pixel is represented by a number that corresponds to its color. However, there is another option. We partition the picture in rectangles of the format 10×10. Then each small rectangle consists of 100 pixels and every possible painting of that rectangle is a letter in the alphabet. This makes the alphabet very big. However, it offers some advantages – some symbols are extremely rare, and certain symbols are not likely to be followed by other symbols.

The case when the message consists of IID letters

We will first simplify the problem a lot and our later generalizations will look more realistic.

The alphabet $\mathcal{A} = \{a_1, a_2, \dots, a_k\}$ is fixed. A message of length n is simply an element of $\mathcal{A}^n$. We will assume that we live in a simplified world in which each message must be of length n.

Definition 1. *A lossless data compression is an injective function*

$$f : \mathcal{A}^n \to \bigcup_{i=1}^{N} \{0,1\}^i$$

for some positive integer N.

For each message $\overrightarrow{x} \in \mathcal{A}^n$ we denote by $\ell\left(\overrightarrow{x}\right)$ the length of the message $\overrightarrow{x}$.

It is easy to create lossless data compressions in which all messages have the same length, although the use of the word *compression* is hardly justified with such behavior. The length ℓ of each message has to satisfy

$$\ell \geq n \cdot \log_2 |\mathcal{A}| .$$

Now we will assume that a message is random in the following sense

Definition 2. *A random iid message is a sequence X_1, X_2, $\dots$, X_n of independent $\mathcal{A}$-valued random variables with the same distribution.*

Let us fix a sequence $(p_1, p_2, \dots, p_k)$ of positive numbers with sum 1 and assume that each of the random variables has the same distribution $\mathbb{P}$ that satisfies

$$\mathbb{P}(X = a_i) = p_i, \quad \text{for} \quad i \in \{1, 2, \dots, k\}.$$

We will divide all messages in two classes: *Typical* and *Atypical*. To motivate the choice for definition of a typical message, let us fix a sequence $\overrightarrow{x} = (x_1, \dots, x_n)$ and calculate the probability that a random iid message $\overrightarrow{X}$ exactly matches the chosen sequence $\overrightarrow{x}$.

The independence implies that this probability is equal to

$$\begin{aligned} \mathbb{P}\left(\overrightarrow{X} = \overrightarrow{x}\right) &= \mathbb{P}\left(X_1 = x_1\right) \cdots \mathbb{P}\left(X_n = x_n\right) \\ &= 2^{\sum_{j=1}^{n} \log_2 \mathbb{P}\left(X_j = x_j\right)} \\ &= 2^{\sum_{i=1}^{k} n_i \log_2 p_i}, \end{aligned}$$

where n_i is the number of times that the symbol a_i appears in the message $\overrightarrow{X}$. A typical message $\overrightarrow{x_\star}$ is one of those in which the frequency $\frac{n_i}{n}$ of the symbol a_i is close to its probability p_i. If we are dealing with a typical message, then we have $n_i \approx p_i n$ and the probability of such a message is

$$\mathbb{P}\left(\overrightarrow{X} = \overrightarrow{x_\star}\right) \approx 2^{n \sum_{i=1}^{k} p_i \log p_i}.$$

Definition 3. *The quantity*

$$H = -\sum_{i=1}^{k} p_i \log_2 p_i. \tag{1}$$

is called the entropy *of the alphabet* $\mathcal{A}$.

A typical message appears with probability of order 2^{-nH}. The messages with this exact probability are still rare. However, it turns out that messages with probabilities in a narrow interval around 2^{-nH} are actually incredibly common. These are called ε-typical messages.

Definition 4. *For fixed* ε, *we say that* $\overrightarrow{x} = (x_1, \dots, x_n)$ *is* ε*-typical if*

$$\mathbb{P}\left(\overrightarrow{X} = \overrightarrow{x}\right) \in \left(2^{-n(H+\varepsilon)}, 2^{-n(H-\varepsilon)}\right).$$

The set of all ε*-typical sequences* $\overrightarrow{x}$ *is denoted by* $\mathcal{A}_\varepsilon$.

We will prove that typical messages occur with very high probability. We will also prove that the set of all typical messages is very small. The set of all messages has $|\mathcal{A}|^n$ elements. However, we will see that the set of typical messages has only 2^{nH} messages.

Then when we see a typical message, we preface it with the number 0 and use nH bits to compress it. If the message is atypical, we use 1 as the first bit and are free to do a rather sloppy job with the rest. Atypical messages are numerous but rare, and the average performance of the described algorithm is quite acceptable.

We will now provide the formal statement of the theorem and its proof.

Theorem 1. *For every given $\omega > 0$, there exists n_0 such that for every $n \geq n_0$ there is a lossless data compression in which the average length of a message satisfies*

$$\mathbb{E}\left[\ell\left(\overrightarrow{X}\right)\right] \leq nH + n\omega. \tag{2}$$

Proof. Let us choose $\varepsilon = \frac{\omega}{\lceil \log_2 |\mathcal{A}| \rceil + 2}$ and denote by A_ε the set of all ε-typical messages. We will study this typical set and prove that it has the following two properties:

(i) *Large probability.* For fixed ε, there exists n_0 such that when $n \geq n_0$ the following holds

$$\mathbb{P}\left(\overrightarrow{X} \in \mathcal{A}_\varepsilon\right) > 1 - \varepsilon. \tag{3}$$

Exercise 1. *Prove that there exists n_0 such that for every $n \geq n_0$ the inequality (3) holds.*

(ii) *Small cardinality.* The number of elements of $\mathcal{A}_\varepsilon$ satisfies

$$|\mathcal{A}_\varepsilon| < 2^{n(H+\varepsilon)}. \tag{4}$$

Then, upon receiving the message $\overrightarrow{X}$, the strategy for compression consists of checking whether $\overrightarrow{X}$ is typical, and depending on the answer proceed as follows:

1. If it is, then the first character of compressed message is 0, followed by an element of $\{0,1\}^{n\lceil H+\varepsilon \rceil}$.

2. If it is not, then the first character of the compressed message is 1, followed by an element of $\{0,1\}^{n\lceil \log_2 |\mathcal{A}| \rceil}$.

The average length of the compressed message is

$$\begin{aligned}
\mathbb{E}\left[\ell\left(\overrightarrow{X}\right)\right] &= (1 + n(H+\varepsilon))\,\mathbb{P}\left(\overrightarrow{X} \in \mathcal{A}_\varepsilon\right) \\
&\quad + (1 + n\lceil \log_2 |\mathcal{A}| \rceil)\,\mathbb{P}\left(\overrightarrow{A} \notin \mathcal{A}_\varepsilon\right) \\
&\leq 1 + n(H+\varepsilon) + n\lceil \log_2 |\mathcal{A}| \rceil \varepsilon \\
&\leq nH + \varepsilon n\left(\lceil \log_2 |\mathcal{A}| \rceil + 2\right) = n(H+\omega).
\end{aligned}$$

It remains to prove the inequalities (3) and (4). The first of the two inequalities follows from the weak law of large numbers and is left as an exercise. The relation (4) follows from

$$1 = \sum_{\overrightarrow{x} \in \mathcal{A}^n} \mathbb{P}\left(\overrightarrow{X} = \overrightarrow{x}\right) \geq \sum_{\overrightarrow{x} \in \mathcal{A}_\varepsilon} \mathbb{P}\left(\overrightarrow{X} = \overrightarrow{x}\right) \geq |\mathcal{A}_\varepsilon| \cdot 2^{-n(H+\varepsilon)}.$$

□

Exercise 2. *Prove that*

$$|\mathcal{A}_\varepsilon| > (1-\varepsilon) \cdot 2^{n(H-\varepsilon)}.$$

The algorithm we constructed creates a lossless compression. It is very effective when the raw data is consists of components that are independent from each other. When it comes to paintings, Pollock would be most likely to produce a work that can be successfully compressed with the algorithm described above. Fortunately, the PNG format uses a better algorithm.

Method of Types

Our next task is to analyze the success of a compression algorithm in situations that a raw data does not exactly follow the distribution $\mathbb{P}$. Every sequence $\overrightarrow{x} \in \mathcal{A}^n$ that corresponds to a raw data generates a probability measure $\mathbb{P}_{\overrightarrow{x}}$ on $\mathcal{A}$. To each element $a \in \mathcal{A}$ we define $\mathbb{P}_{\overrightarrow{x}}(a)$ as the frequence in which a appears in the sequence $\overrightarrow{x}$. This probability measure is called *type* of $\overrightarrow{x}$ and its formal definition is

Definition 5. *For $\overrightarrow{x} \in \mathcal{A}^n$, the* type *of $\overrightarrow{x}$ is the probability measure $\mathbb{P}_{\overrightarrow{x}}$ defined as*

$$\mathbb{P}_{\overrightarrow{x}}(a) = \frac{1}{n} \sum_{i=1}^{k} 1_{x_i}(a). \tag{5}$$

The set of all types on $\mathcal{A}^n$ is denoted by $\mathcal{P}_n$. The sequences $\overrightarrow{x}$ and $\overrightarrow{y}$ generate the same type (i.e. $\mathbb{P}_{\overrightarrow{x}} = \mathbb{P}_{\overrightarrow{y}}$) if they can be obtained from each other by permuting their components. For a given $\mathbb{P} \in \mathcal{P}_n$ we denote by $T(\mathbb{P}) \subseteq \mathcal{A}^n$ the set of sequences $\overrightarrow{x} \in \mathcal{A}^n$ for which $\mathbb{P}_x = \mathbb{P}$.

Theorem 2. *The number $|\mathcal{P}_n|$ satisfies*

$$|\mathcal{P}_n| = \binom{n+|\mathcal{A}|-1}{n} \quad \text{and} \tag{6}$$
$$|\mathcal{P}_n| < (n+1)^{|\mathcal{A}|}. \tag{7}$$

Exercise 3. *Prove the equality* (6) *and the inequality* (7).

Assume that $\overrightarrow{X} = (X_1, \dots, X_n)$ is an iid random message where each component is drawn according to the probability $\mathbb{Q}$. For a fixed sequence $\overrightarrow{x} \in \mathcal{A}^n$ we have

$$\begin{aligned}
\mathbb{Q}\left(\overrightarrow{X} = \overrightarrow{x}\right) &= \prod_{i=1}^{n} \mathbb{Q}\left(X = x_i\right) = 2^{\sum_{i=1}^{n} \log_2 \mathbb{Q}(X=x_i)} \\
&= 2^{\sum_{i=1}^{n} \sum_{a\in\mathcal{A}} 1_{x_i}(a) \log_2 \mathbb{Q}(X=a)} \\
&= 2^{\sum_{a\in\mathcal{A}} \sum_{i=1}^{n} 1_{x_i}(a) \log_2 \mathbb{Q}(X=a)} \\
&= 2^{n \sum_{a\in\mathcal{A}} \log_2 \mathbb{Q}(X=a) \frac{1}{n} \sum_{i=1}^{n} 1_{x_i}(a)} \\
&= 2^{n \sum_{a\in\mathcal{A}} \mathbb{P}_{\overrightarrow{x}}(a) \log_2 \mathbb{Q}(X=a)}.
\end{aligned}$$

The last quantity can be written in terms of the entropy of $\mathbb{P}_{\overrightarrow{x}}$ in the following way:

$$\begin{aligned}
\sum_{a\in\mathcal{A}} \mathbb{P}_{\overrightarrow{x}}(a) \log_2 \mathbb{Q}\left(X=a\right) &= \sum_{a\in\mathcal{A}} \mathbb{P}_{\overrightarrow{x}}(a) \log_2 \mathbb{P}_{\overrightarrow{x}}\left(X=a\right) \\
&\quad - \sum_{a\in\mathcal{A}} \mathbb{P}_{\overrightarrow{x}}(a) \log_2 \frac{\mathbb{P}_{\overrightarrow{x}}\left(X=a\right)}{\mathbb{Q}\left(X=a\right)} \\
&= -H\left(\mathbb{P}_{\overrightarrow{x}}\right) - H\left(\mathbb{P}_{\overrightarrow{x}} | \mathbb{Q}\right),
\end{aligned}$$

where $H(\mu)$ is an entropy of the measure μ and $H(\mu|\nu)$ is a relative entropy of the measure μ with respect to ν. These are defined as

$$H(\mu) = -\mathbb{E}_\mu \left[\log_2 \mu\right] \tag{8}$$
$$H(\mu|\nu) = \mathbb{E}_\mu \left[\log_2 \frac{\mu}{\nu}\right]. \tag{9}$$

Exercise 4. *The random variable X has values in the set $\{-2, 0, 2\}$ and the expectation 1. What is the smallest and the largest entropy that the distribution of X can have?*

The previous calculation can be summarized as

Theorem 3. *Fix $\overrightarrow{x} \in \mathcal{A}^n$. If $\overrightarrow{X}$ is a random vector with iid components of distribution $\mathbb{Q}$, the following holds:*

$$\mathbb{Q}\left(\overrightarrow{X} = \overrightarrow{x}\right) = 2^{-n\left(H\left(\mathbb{P}_{\overrightarrow{x}}\right) + H\left(\mathbb{P}_{\overrightarrow{x}} | \mathbb{Q}\right)\right)}. \tag{10}$$

Exercise 5. *Prove that $H(\mu|\nu) \geq 0$. Prove that $H(\mu|\nu) = 0$ if and only if $\mu = \nu$.*

If $\mathbb{Q}$ is a type of $\overrightarrow{x}$, i.e. $\mathbb{Q} = \mathbb{P}_{\overrightarrow{x}}$, then (10) and $H(\mu|\mu) = 0$ imply

$$\mathbb{Q}\left(\overrightarrow{X} = \overrightarrow{x}\right) = 2^{-nH(\mathbb{Q})}. \tag{11}$$

Observe that $H(\mathbb{P}_{\overrightarrow{x}})$ is constant for all sequences of the same type. Let us fix a measure $\mathbb{P} \in \mathcal{A}^n$. Then for $\overrightarrow{x} \in T(\mathbb{P})$, all values $H(\mathbb{P}_{\overrightarrow{x}})$ are equal to $H(\mathbb{P})$. Therefore we have

$$\mathbb{P}\left(\overrightarrow{X} \in T(\mathbb{P})\right) = \sum_{\overrightarrow{x} \in T(\mathbb{P})} \mathbb{P}\left(\overrightarrow{X} = \overrightarrow{x}\right) = |T(\mathbb{P})| \cdot 2^{-nH(\mathbb{P})}. \quad (12)$$

Using that the probability on the left can be at most 1 we obtain the following upper bound on the cardinality of the type

$$|T(\mathbb{P})| \leq 2^{nH(\mathbb{P})}. \quad (13)$$

The following result that is left as an exercise confirms that under the probability measure $\mathbb{P}_{\overrightarrow{x}}$, it is most likely to draw a sequence whose type class is the same as the type class of $\overrightarrow{x}$.

Exercise 6. *Prove that if* $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_n$ *are two types then*

$$\mathbb{P}\left(\overrightarrow{X} \in T(\mathbb{P})\right) \geq \mathbb{P}\left(\overrightarrow{X} \in T(\mathbb{Q})\right). \quad (14)$$

Using the inequality (14) and the equality (12) we can establish a lower bound for the probability of a type-class.

$$\begin{aligned} 1 &= \sum_{\mathbb{Q} \in \mathcal{P}_n} \mathbb{P}(T(\mathbb{Q})) \\ &\leq \mathbb{P}(T(\mathbb{P})) \cdot |\mathcal{P}_n| \\ &= |T(\mathbb{P})| \cdot 2^{-nH(\mathbb{P})} \cdot |\mathcal{P}_n|. \end{aligned}$$

Re-arranging the terms gives us

$$|T(\mathbb{P})| \geq \frac{2^{-nH(\mathbb{P})}}{|\mathcal{P}_n|},$$

which together with (7) implies

$$|T(\mathbb{P})| \geq \frac{2^{nH(\mathbb{P})}}{(n+1)^{\mathcal{A}}}. \quad (15)$$

Theorem 4 (Sanov)**.** *Assume that* $E \subseteq \mathcal{P}_n$*. Then the following holds*

$$\limsup_{n\to\infty} \frac{1}{n} \log_2 \mathbb{P}\left(\mathbb{P}_{\overrightarrow{X}} \in E\right) \leq -\inf_{\mathbb{Q} \in E} H(\mathbb{Q} | \mathbb{P}), \quad (16)$$

$$\liminf_{n\to\infty} \frac{1}{n} \log_2 \mathbb{P}\left(\mathbb{P}_{\overrightarrow{X}} \in E\right) \geq -\inf_{\mathbb{Q} \in E^\circ} H(\mathbb{Q} | \mathbb{P}), \quad (17)$$

where E° *is the interior of the set* E*.*

Proof. The upper bound follows from

$$\begin{aligned}
\mathbb{P}\left(\mathbb{P}_{\overrightarrow{X}} \in E\right) &= \sum_{\mathbb{Q}\in E} \mathbb{P}\left(\mathbb{P}_{\overrightarrow{X}} = \mathbb{Q}\right) \\
&= \sum_{\mathbb{Q}\in E} \mathbb{P}\left(\overrightarrow{X} \in T(\mathbb{Q})\right) \\
&\leq |E| \sup_{\mathbb{Q}\in E} \mathbb{P}\left(\overrightarrow{X} \in T(\mathbb{Q})\right) \\
&\leq (n+1)^{|\mathcal{A}|} \cdot \sup_{\mathbb{Q}\in E} \mathbb{P}\left(\overrightarrow{X} \in T(\mathbb{Q})\right). \qquad (18)
\end{aligned}$$

Using the identity (10) we conclude

$$\begin{aligned}
\mathbb{P}\left(\overrightarrow{X} \in T(\mathbb{Q})\right) &= \sum_{\overrightarrow{y}\in T(\mathbb{Q})} \mathbb{P}\left(\overrightarrow{X} = \overrightarrow{y}\right) \\
&= |T(\mathbb{Q})| \cdot \mathbb{P}\left(\overrightarrow{X} = \overrightarrow{y}\right) \\
&= |T(\mathbb{Q})| \cdot 2^{-n\left(H\left(\mathbb{P}_{\overrightarrow{y}}\right)+H\left(\mathbb{P}_{\overrightarrow{y}}\middle|\mathbb{P}\right)\right)} \\
&= |T(\mathbb{Q})| \cdot 2^{-n(H(Q)+H(\mathbb{Q}|\mathbb{P}))} \\
&\leq 2^{-nH(\mathbb{Q}|\mathbb{P})}. \qquad (19)
\end{aligned}$$

From (18) and (19) we obtain (16). Now we concentrate on proving (17). Let $\varepsilon > 0$ and $\mathbb{P}_\star$ an element of E° such that

$$H(\mathbb{P}_\star|\mathbb{P}) \leq \varepsilon + \inf_{\mathbb{Q}\in E^\circ} H(\mathbb{Q}|\mathbb{P}). \qquad (20)$$

Observe that $\mathbb{P}_\star$ is an element of $\mathcal{P}_n$, hence there exists $\overrightarrow{x_\star}$ such that $\mathbb{P}_\star = \mathbb{P}_{\overrightarrow{x_\star}}$. We will obtain a lower bound for the probability of the event $\left\{\mathbb{P}_{\overrightarrow{X}} \in E\right\}$ in the following way

$$\begin{aligned}
\mathbb{P}\left(\mathbb{P}_{\overrightarrow{X}} \in E\right) &\geq \mathbb{P}\left(\mathbb{P}_{\overrightarrow{X}} = \mathbb{P}_\star\right) \\
&= \mathbb{P}\left(\overrightarrow{X} \in T(\mathbb{P}_\star)\right) \\
&= |T(\mathbb{P}_\star)| \cdot \mathbb{P}\left(\overrightarrow{X} = \overrightarrow{x_\star}\right) \\
&\geq \frac{2^{nH(\mathbb{P}_\star)}}{(n+1)^{|\mathcal{A}|}} \cdot \mathbb{P}\left(\overrightarrow{X} = \overrightarrow{x_\star}\right) \\
&= \frac{2^{nH(\mathbb{P}_\star)}}{(n+1)^{|\mathcal{A}|}} \cdot 2^{-n(H(\mathbb{P}_\star)+H(\mathbb{P}_\star|\mathbb{P}))} \\
&= \frac{2^{-nH(\mathbb{P}_\star|\mathbb{P})}}{(n+1)^{|\mathcal{A}|}} \\
&\geq \frac{2^{-n\left(\varepsilon+\inf_{\mathbb{Q}\in E^\circ} H(\mathbb{Q}|\mathbb{P})\right)}}{(n+1)^{|\mathcal{A}|}}.
\end{aligned}$$

The last inequality implies

$$\liminf_{n\to\infty} \frac{1}{n} \log_2 \mathbb{P}\left(\mathbb{P}_{\overrightarrow{X}} \in E\right) \geq -\varepsilon - \inf_{Q\in E^\circ} H(\mathbb{Q}|\mathbb{P}).$$

Since the last inequality holds for every $\varepsilon > 0$ we conclude that (17) must hold as well. $\square$

Exercise 7. *Let us define* $E = \left\{ \mathbb{Q} \in \mathcal{P}_n : \int_{\mathcal{A}} a \, d\mathbb{Q}(a) \geq \alpha \right\}$.

(a) *Assume that* $X_1, \dots, X_n$ *are iid with distribution* $\mathbb{P}$ *on* $\mathcal{A}$. *Prove that*

$$\mathbb{P}\left(\frac{X_1 + \cdots + X_n}{n} \geq \alpha \right) = \mathbb{P}\left(\mathbb{P}_{\overrightarrow{X}} \in E \right).$$

(b) *Use the method of Lagrange multipliers to prove that*

$$\inf_{\mathbb{Q} \in E} H\left(\mathbb{Q} | \, \mathbb{P} \right) = \alpha\theta - \log_2 \left(\sum_{i=1}^{k} p_i 2^{\theta a_i} \right),$$

where θ *is the unique solution of* $\dfrac{\sum_{i=1}^k a_i p_i 2^{\theta a_i}}{\sum_{i=1}^k p_i 2^{\theta a_i}} = \alpha$.

(c) *Prove that Sanov's theorem implies Cramér's theorem in the case that the probability space is discrete.*

Binary Investigation

A random element is chosen from the set $S = \{s_1, \dots, s_n\}$ and painted in green. The probability that the element s_i is chosen is p_i, where $p_1, \dots, p_n$ are fixed positive real numbers that add up to 1.

In each step we are allowed to choose a subset T of S and ask the question:

"Does the green element belong to the set T?"

We receive a "Yes" or "No" for an answer. Our goal is to find out the green number.

We are in business of playing this game many times. In each game, a number from S becomes green, and we start asking questions until we figure out which one. Then we play the game again. However, the answers to our questions are pronounced in an annoying voice that we would like to hear a little bit less of. Obviously, there is still much fun left to the game that we can't resist playing it repeatedly. What is the way to play?

In our first step we must choose a set T. There are only finitely many choices for the first question (how many?). Consequently, there are only finitely many strategies in the game and we want to identify the best.

We can build a tree for each strategy. One example is shown in Figure 2.

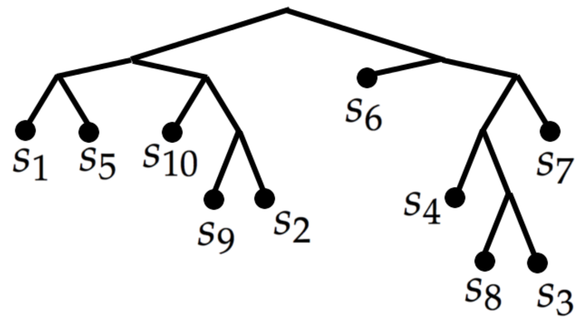

Figure 2: A tree corresponding to the following strategy:
Question 1: "Does the green number belong to the set $\{s_1, s_2, s_5, s_9, s_{10}\}$?"
Question 2 if the Answer 1 is "Yes:" "Does the green number belong to the set $\{s_1, s_5\}$?"
Question 2 if the Answer 1 is "No:" "Does the green number belong to the set $\{s_6\}$?"

Let us label by m_i the number of questions in the case that the green number is s_i. In Example from Figure 2, we have that $m_1 = m_5 = m_{10} = m_7 = 3$, $m_2 = m_4 = m_9 = 4$, $m_6 = 2$, $m_3 = m_8 = 5$.

Exercise 8. *Prove that the following equality always holds:*

$$\sum_{i=1}^{n} \frac{1}{2^{m_i}} = 1. \tag{21}$$

Once the strategy is fixed, we can denote by $N(S)$ the total number of questions asked. Then $N(S)$ is a random variable and its expectation satisfies

$$\mathbb{E}\left[N(S)\right] = \sum_{i=1}^{n} m_i p_i.$$

Conversely, the following holds:

Exercise 9. *For every sequence of n positive integers $m_1, m_2, \dots, m_n$ that satisfy (21), there exists a tree with n nodes such that the height of the i-th node is exactly m_i.*

Hint: Prove that there are two of the numbers m_i and m_j that are equal. Then consider the problem in which these two numbers are replaced with a single number equal to $m_i - 1$.

Theorem 5. *Let us denote*

$$H = -\sum_{s \in S} \mathbb{P}(s) \log_2 \mathbb{P}(s) = \sum_{i=1}^{n} -p_i \log_2 p_i.$$

The following inequalities hold

$$H \leq \mathbb{E}\left[N(S)\right] < H + 1. \tag{22}$$

Proof. We will first prove the inequality $H \leq \mathbb{E}\left[N(S)\right]$. We will prove the inequality for each strategy using the induction on the number of elements n in the set S. Let us denote by $T \subseteq S$ the first subset that we use in asking the first questions. Then T also corresponds to the event that the answer to the first question is "Yes," in which case we go to the left sub-tree. The set T^C corresponds to the event that the answer to the first question is "No". Then we have

$$\mathbb{E}\left[N(S)\right] = \mathbb{E}\left[N(S)|\, T\right] \cdot \mathbb{P}(T) + \mathbb{E}\left[N(S)|\, T^C\right] \cdot \mathbb{P}\left(T^C\right). \tag{23}$$

Assume that an element of T is chosen. Then $N(S) = 1 + N(T)$, where $N(T)$ is the number of questions necessary to find out which number from T is green. We are assuming that that the left sub-tree is used. Since T has fewer elements than S, we will use induction hypothesis for the probability defined on T with $\mathbb{P}_T(x) = \mathbb{P}(x)/\mathbb{P}(T)$. Therefore

$$\begin{aligned} \mathbb{E}\left[N(S)|\, T\right] &= 1 + \mathbb{E}\left[N(T)|\, T\right] \\ &\geq 1 - \sum_{x \in T} \frac{\mathbb{P}(x)}{\mathbb{P}(T)} \log_2 \frac{\mathbb{P}(x)}{\mathbb{P}(T)}. \end{aligned}$$

Using an equivalent inequality for T^C, the relation (23) implies

$$\begin{aligned} \mathbb{E}\left[N(S)\right] &\geq \mathbb{P}(T) + \mathbb{P}\left(T^C\right) - \sum_{x \in S} \mathbb{P}(x) \log_2 \mathbb{P}(x) \\ &\quad + \mathbb{P}(T) \log_2 \mathbb{P}(T) + \mathbb{P}\left(T^C\right) \log_2 \mathbb{P}\left(T^C\right) \\ &= 1 + H + \mathbb{P}(T) \log_2 \mathbb{P}(T) + \mathbb{P}\left(T^C\right) \log_2 \mathbb{P}\left(T^C\right) \\ &\geq H. \end{aligned}$$

Exercise 10. *For $x \in (0,1)$ prove that*

$$1 + x \log_2 x + (1-x) \log_2 (1-x) \geq 0.$$

The last inequality follows from Exercise 10.

Now we will prove that there exists a strategy for which $\mathbb{E}\left[N(S)\right] < H + 1$. Let us define

$$\hat{m}_i = \lceil -\log_2 \mathbb{P}\left(s_i\right) \rceil$$

for $i \in \{1, 2, \dots, n\}$.

Observe that

$$1 = \sum_{i=1}^{n} \frac{1}{2^{-\log_2 \mathbb{P}(s_i)}} \geq \sum_{i=1}^{n} \frac{1}{2^{\lceil -\log_2 \mathbb{P}(s_i) \rceil}} = \sum_{i=1}^{n} \frac{1}{2^{\hat{m}_i}}.$$

Exercise 11. *If $\hat{m}_1, \dots, \hat{m}_n$ are positive integers such that $\sum_{i=1}^{n} \frac{1}{2^{\hat{m}_i}} \leq 1$ there are positive integers $m_1, \dots, m_n$ such that $m_i \leq \hat{m}_i$ for all $i \in \{1, \dots, n\}$ and*

$$\sum_{i=1}^{n} \frac{1}{2^{m_i}} = 1.$$

According to result from Exercise 11 there are positive integers m_1, ..., m_n smaller than $\hat{m}_1, \dots, \hat{m}_n$, respectively, that satisfy (21). The result of Exercise 9 means that there is a tree with n nodes whose heights are $m_1, \ldots, m_n$. The Expected value of the number of questions in such strategy is

$$\begin{aligned}
\mathbb{E}\left[N(S)\right] &= \sum_{i=1}^{n} m_i p_i \leq \sum_{i=1}^{n} \hat{m}_i p_i < \sum_{i=1}^{n} \left(-\log_2 p_i + 1\right) p_i \\
&= \sum_{i=1}^{n} p_i - \sum_{i=1}^{n} p_i \log_2 p_i \\
&= 1 + H.
\end{aligned}$$

This completes the proof of the Theorem 5. $\square$