August 30 **August 2019**

Tutorial 3: The Various Facets of Shannon Entropy

Course Instructor: Himanshu Tyagi *Prepared by: Karthik*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

Contents

3.1 Preliminaries: Weak Law of Large Numbers

We begin this tutorial by recalling the statement of the weak law of large numbers. All logarithms appearing in this document are to the base 2.

Theorem 3.1.1. *(A version of the weak law of large numbers) Suppose $X_1, X_2, \ldots$ are independent and identically distributed (iid) random variables with common distribution P. Let $E[X_1] < \infty$ and Var$(X_1) < \infty$. Then, we have*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{i.p.} E[X_1] \quad \text{as } n \to \infty,$$

i.e., for any $\epsilon > 0$, we have

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - E[X_1]\right| > \epsilon\right) \longrightarrow 0 \quad \text{as } n \to \infty.$$

Remark 1.

1. Note that
$$\text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\text{Var}(X_1) \longrightarrow 0 \quad \text{as } n \to \infty.$$
Thus, the sequence of random variables given by $\left\{\frac{1}{n}\sum_{i=}^{n} X_i\right\}_{n\geq 1}$ converges in some sense to a zero variance random variable, which is a constant random variable. The above theorem captures that this convergence is in the sense of "in probability", and that the constant is the mean of the underlying distribution from which the X_i's are drawn.

2. The above theorem conveys the fact that the probability of the random variable $\frac{1}{n}\sum_{i=}^{n} X_i$ deviating from its mean (which is $E[X_1]$) becomes vanishingly small for larger and larger values of n. Interpreted in another way, suppose that $E[X_1]$ is unknown, and our interest is to estimate this mean using samples $X_1, X_2, \ldots$. Then, the above theorem claims that the probability of error in estimation becomes vanishingly small with larger and larger number of samples.

Example 3.1.2 (Cover and Thomas, Problem 3.4)**.** Suppose that $X_1, X_2, \ldots$ are drawn iid according to the distribution given by

$$P(X_1 = 1) = \frac{1}{2}, \quad P(X_1 = 2) = \frac{1}{4}, \quad P(X_1 = 3) = \frac{1}{4}.$$

Find the limiting value of $\left(\prod_{i=1}^{n} X_i\right)^{\frac{1}{n}}$.

Let $Z_n = \left(\prod_{i=1}^{n} X_i\right)^{\frac{1}{n}}$. Then,

$$\begin{aligned}\log Z_n &= \frac{1}{n}\sum_{i=1}^{n} \log X_i \\ &\xrightarrow{i.p.} E[\log X_1].\end{aligned}$$

Thus, we have $Z_n \xrightarrow{i.p.} 2^{E[\log X_1]}$. It is left as exercise to compute $E[\log X_1]$.

3.2 Entropy as Large Probability Lower and Upper Bounds on $-\log p(X)$

Recall that if X is a discrete random variable with probability mass function p taking values in $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$, its Shannon entropy (measured in bits) is defined as

$$\begin{aligned} H(X) &\coloneqq \sum_{i=1}^{n} p(x_i) \log \frac{1}{p(x_i)} \\ &= E[-\log p(X)]. \end{aligned}$$

Note that $-\log p(X)$ is a nonnegative random variable, and thus its expectation is well-defined (i.e., it is not of the form $\infty - \infty$). It is sometimes instructive to write $H(X)$ as $H(P)$ when $X \sim P$ since the Shannon entropy of a random variable does not depend on the specific values it takes, but on the probabilities with which it takes those values. Thus, we may use the phrase "entropy of distribution P" when referring to the entropy of $X \sim P$.

Given any $\epsilon > 0$, let $L_\epsilon(P)$ denote the smallest number of bits required to represent the value of a discrete random variable $X \sim P$ correctly with probability at least $1 - \epsilon$. In other words, suppose X takes values in the discrete set $\mathcal{X}$. Then,

$$L_\epsilon(P) = \min \left\{ \left\lceil \log |A| \right\rceil : A \subseteq \mathcal{X}, \quad P(A) \geq 1 - \epsilon \right\}.$$

In the above expression for $L_\epsilon(P)$, $|A|$ denotes the number of elements in A. Suppose $\mathcal{X} = \{x_1, x_2, \ldots\}$, and the elements of $\mathcal{X}$ are ordered in decreasing order of their probabilities. Suppose further that $x_{(i)}$ denotes the element with the ith largest probability. Then, if

$$n^* \coloneqq \min \left\{ n : \sum_{i=1}^{n} p(x_{(i)}) \geq 1 - \epsilon \right\}, \quad A^* \coloneqq \{x_{(1)}, \ldots, x_{(n^*)}\},$$

it follows that $L_\epsilon(P) = \left\lceil \log |A^*| \right\rceil$. However, the above computation of $L_\epsilon(P)$ requires us to go through the probabilities of all the elements of $\mathcal{X}$ in order to sort them in decreasing order, an exercise that may be practically infeasible when the size of $\mathcal{X}$ is large. Thus, it may not be possible to compute $L_\epsilon(P)$ exactly in many practical scenarios.

This brings us to the question of whether we can find good upper and lower bounds for $L_\epsilon(P)$ that give us an idea of what its actual value is. Our goal in this section is to demonstrate that the entropy of the distribution P serves as a good upper and lower bound for $L_\epsilon(P)$ (up to constant additive factors which do not contribute significantly and may thus be neglected).

3.2.1 Entropy as Large Probability Lower and Upper Bounds for $-\log p(X)$

Let us recall the Chebyshev's inequality. Suppose X is a random variable with $E[X] < \infty$ and $\mathrm{Var}(X) < \infty$. Then, for any $t > 0$,

$$P\Big(|X - E[X]| > t \Big) \leq \frac{\mathrm{Var}(X)}{t^2}.$$

In particular, given any $\epsilon > 0$, substituting $t = \sqrt{\frac{\mathrm{Var}}{\epsilon}}$, we get

$$P\left(|X - E[X]| > \sqrt{\frac{\mathrm{Var}(X)}{\epsilon}} \right) \leq \epsilon,$$

which may also be expressed as

$$P\left(E[X] - \sqrt{\frac{\mathrm{Var}(X)}{\epsilon}} \leq X \leq E[X] + \sqrt{\frac{\mathrm{Var}(X)}{\epsilon}}\right) \geq 1 - \epsilon.$$

Thus, the set of values of X which lie in the interval $\left[E[X] - \sqrt{\frac{\mathrm{Var}}{\epsilon}}, \quad E[X] + \sqrt{\frac{\mathrm{Var}}{\epsilon}}\right]$ is a subset of $\mathcal{X}$ with probability at least $1 - \epsilon$. Thus, Chebyshev's inequality gives us a set which has large probability. An important point to note here is that even if the underlying distribution P of the random variable X is not known (in which case we cannot compute the set A^* described before), we may still assert that **there exists a set which has large probability** as a consequence of Chebyshev's inequality.

Let X be a discrete random variable with pmf p. Applying Chebyshev's inequality to the random variable $Z = -\log p(X)$ noting that $E[Z] = H(X)$, we get that for any $\epsilon > 0$,

$$P\left(H(X) - \sqrt{\frac{\mathrm{Var}(-\log p(X))}{\epsilon}} \leq -\log p(X) \leq H(X) + \sqrt{\frac{\mathrm{Var}(-\log(p(X))}{\epsilon}}\right) \geq 1 - \epsilon.$$

In particular, it follows from the above inequality that

$$P\left(-\log p(X) \geq H(X) - \sqrt{\frac{\mathrm{Var}(-\log p(X))}{\epsilon}}\right) \geq 1 - \epsilon,$$
$$P\left(-\log p(X) \leq H(X) + \sqrt{\frac{\mathrm{Var}(-\log p(X))}{\epsilon}}\right) \geq 1 - \epsilon.$$

Thus, as long as $\mathrm{Var}(-\log p(X))$ is not too large, we see that it is $H(X)$ that is the dominating term in both the high probability upper and lower bounds for the random variable $Z = -\log p(X)$.

Remark 2. It may seem puzzling to the reader that the $\frac{1}{\epsilon}$ term in the lower and upper bounds may dominate $H(X)$ when ϵ is small. We shall return to this point later and demonstrate later that the effect of ϵ will vanish when we consider n iid copies $X_1, \ldots, X_n$ of the random variable $X \sim P$, and let $n \to \infty$. This was the genius of Shannon.

3.2.2 Entropy and $L_\epsilon(P)$

Let X be a discrete random variable with pmf p. In this section, we show the following important results:

1. Any high probability upper bound for the random variable $Z = -\log p(X)$ is also an upper bound for $L_\epsilon(P)$.
2. Any high probability lower bound for the random variable $Z = -\log p(X)$ is also a lower bound for $L_\epsilon(P)$.

Together, the above results will imply that $H(X)$ forms good upper and lower bound for $L_\epsilon(P)$ when $X \sim P$, thus implying that a good estimate for $L_\epsilon(P)$ is $H(X)$.

Lemma 3.2.1. *Let $X \sim P$ be a discrete random variable with pmf p taking values in the set $\mathcal{X}$. Fix $\epsilon > 0$. Suppose that $\lambda > 0$ is a constant such that the set*

$$A_\lambda \coloneqq \{x \in \mathcal{X} : -\log p(x) \leq \lambda\}$$

satisfies $P(A_\lambda) \geq 1 - \epsilon$. Then, $L_\epsilon(P) \leq \lambda$.

Proof. Note that

$$\begin{aligned} 1 &\geq P(A_\lambda) \\ &= \sum_{x \in A_\lambda} p(x) \\ &\geq \sum_{x \in A_\lambda} 2^{-\lambda} \\ &= |A_\lambda| \cdot 2^{-\lambda}, \end{aligned}$$

from which it follows that $|A_\lambda| \leq 2^\lambda$. Thus, $\lceil \log |A_\lambda| \rceil \leq \lambda$, which implies the desired result. □

The above Lemma says that if there exists a constant $\lambda > 0$ that serves as a high probability upper bound for the random variable $Z = -\log p(X)$, then the same constant λ also serves as an upper bound for $L_\epsilon(P)$. However, we know from Chebyshev's inequality that one such λ exists, which is $\lambda = H(X) + \sqrt{\frac{\mathrm{Var}(-\log p(X))}{\epsilon}}$. Thus, it immediately follows from the above Lemma that

$$L_\epsilon(P) \leq H(X) + \sqrt{\frac{\mathrm{Var}(-\log p(X))}{\epsilon}}.$$

Lemma 3.2.2. *Let $X \sim P$ be a discrete random variable with pmf p taking values in the set $\mathcal{X}$. Fix $\epsilon > 0$. Suppose that $\lambda > 0$ is a constant such that the set*

$$B_\lambda \coloneqq \{x \in \mathcal{X} : -\log p(x) \geq \lambda\}$$

satisfies $P(B_\lambda) \geq 1 - \frac{\epsilon}{2}$. Then, $L_\epsilon(P) \geq \lambda - \log \frac{1}{1-\epsilon}$.

Proof. Note that

$$\begin{aligned} 1 - \frac{\epsilon}{2} &\leq P(B_\lambda) \\ &= \sum_{x \in B_\lambda} p(x) \\ &\leq \sum_{x \in B_\lambda} 2^{-\lambda} \\ &= |B_\lambda| \cdot 2^{-\lambda}, \end{aligned}$$

from which it follows that $|B_\lambda| \geq 2^\lambda \left(1 - \frac{\epsilon}{2}\right)$.

Let $A \subset \mathcal{X}$ be any set satisfying $P(A) \geq 1 - \frac{\epsilon}{2}$. Note that A and B_λ are both high probability sets. Therefore, their intersection must have high probability. In particular, we have

$$\begin{aligned} P(A \cap B_\lambda) &= P(A) + P(B_\lambda) - P(A \cup B_\lambda) \\ &\geq P(A) + P(B_\lambda) - 1 \\ &\geq 1 - \epsilon. \end{aligned}$$

Furthermore, since $(A \cap B_\lambda) \subseteq B_\lambda$, we have that $-\log p(x) \geq \lambda$ for all $x \in A \cap B_\lambda$. Therefore, by arguments similarly as before, we get that $|A \cap B_\lambda| \geq 2^\lambda (1 - \epsilon)$. Then, we have

$$\begin{aligned} \lceil \log |A| \rceil &\geq \log |A| \\ &\geq \log |A \cap B_\lambda| \\ &\geq \lambda - \log \frac{1}{1-\epsilon}. \end{aligned}$$

Since the above is true for all sets $A \subseteq \mathcal{X}$ satisfying $P(A) \geq 1 - \epsilon$, it follows that $L_\epsilon(P) \geq \lambda - \log \frac{1}{1-\epsilon}$. □

The above Lemma says that if there exists a constant $\lambda > 0$ that serves as a high probability lower bound for the random variable $Z = -\log p(X)$, then the same constant λ also serves as a lower bound for $L_\epsilon(P)$ (up to additive factors which can be neglected as we shall soon demonstrate). However, we know from Chebyshev's inequality that one such λ exists, which is $\lambda = H(X) - \sqrt{\frac{\mathrm{Var}(-\log p(X))}{\epsilon/2}}$. Thus, it immediately follows from the above Lemma that

$$L_\epsilon(P) \geq H(X) - \sqrt{\frac{\mathrm{Var}(-\log p(X))}{\epsilon/2}} - \log \frac{1}{1-\epsilon}.$$

3.2.3 From a Single Random Variable to $n > 1$ IID Copies

Until now, we have considered upper and lower bounds for $L_\epsilon(P)$ which is the minimum number of bits required to represent the value of a random variable $X \sim P$. We shall now consider $n > 1$ iid copies of the random variable X. Let these be denoted as $X_1, \ldots, X_n$, with $X_i \sim P$ for each $i = 1, \ldots, n$. We shall denote the joint entropy of the collection $X_1, \ldots, X_n$ by $H(X_1, \ldots, X_n)$ or as $H(P^n)$, where P^n is the joint distribution of the n iid copies $X_1, \ldots, X_n$.

Exercise: Show that for $X_1, \ldots, X_n$ drawn iid from a distribution P,

$$H(X_1, \ldots, X_n) = nH(X_1)$$

(or $H(P^n) = nH(P)$).

Given any $\epsilon > 0$, we shall denote by $L_\epsilon(P^n)$ the minimum number of bits required to represent the value of $(X_1, \ldots, X_n)$. That is, $L_\epsilon(P^n)$ is the minimum number of bits required to represent n symbols drawn iid from the distribution P.

While in the previous sections, we derived upper and lower bounds for $L_\epsilon(P)$, we now derive large probability upper and lower bounds for the quantity

$$R_\epsilon(P) \coloneqq \lim_{n\to\infty} \frac{L_\epsilon(P^n)}{n}.$$

Note that the above quantity represents the average minimum number of bits required to represent **a symbol** from the distribution P.

Towards deriving upper and lower bounds for $R_\epsilon(P)$, suppose $X \sim P$ is a discrete random variable with pmf p, and $X_1, \ldots, X_n$ are iid copies of X, then

$$\mathrm{Var}(-\log p(X_1, \ldots, X_n)) = n\mathrm{Var}(-\log p(X)),$$

where $p(X_1, \ldots, X_n)$ denotes the joint pmf of $(X_1, \ldots, X_n)$. It then follows from Chebyshev's and Lemmas 3.2.2 and 3.2.1 that

$$nH(X) - \sqrt{n\frac{\mathrm{Var}(-\log p(X))}{\epsilon/2}} - \log \frac{1}{1-\epsilon} \leq L_\epsilon(P^n) \leq nH(X) + \sqrt{n\frac{\mathrm{Var}(-\log(p(X))}{\epsilon}}.$$

Dividing throughout by n and letting $n \to \infty$, we get that for every $\epsilon > 0$ (note how the effect of ϵ does not manifest in the regime of $n \to \infty$),

$$R_\epsilon(P) = \lim_{n\to\infty} \frac{L_\epsilon(P^n)}{n} = H(X).$$

Thus, $H(P)$ is the least number of bits required (on the average) to represent the value of a random variable $X \sim P$. It is a fundamental quantity in information theory that represents the extent to which the information content (uncertainty) of a distribution may be represented in the most compact manner. Also, entropy may be interpreted as "compressibility" of a probability distribution. Higher the entropy, less compressible the distribution is.

Exercise: Show that if X is a discrete random variable taking values in a finite set $\mathcal{X}$, then

$$H(X) \leq \log |\mathcal{X}|,$$

with equality above attained if and only if $X \sim \text{unif}(\mathcal{X})$.

3.3 The Asymptotic Equipartition Property (AEP)

As the previous exercise demonstrates, a uniform distribution on a finite set $\mathcal{X}$ is the least compressible distribution since its entropy is the largest among all probability distributions on $\mathcal{X}$. In other words, the uniform distribution cannot be compressed any further. In this section, we demonstrate a very important property that if a distribution P on a finite set is not the uniform distribution, then there is scope for further compression in the sense that we may extract a high probability set under the distribution P whose elements are nearly uniformly distributed. For all practical purposes, then, we may focus our attention on this high probability set when doing computations under the distribution P.

The above property is known as the asymptotic equipartition property, also known as the AEP, which is stated below formally.

Theorem 3.3.1. *(Asymptotic equipartition property) Let $X \sim P$ be a discrete random variable with pmf p. Further, let $E[X] < \infty$ and $Var(X) < \infty$. Suppose $X_1, X_2, \ldots$ represent iid copies of X. Then*

$$-\frac{1}{n} \log p(X_1, \ldots, X_n) \xrightarrow{i.p.} H(X) \quad \text{as } n \to \infty.$$

In other words, for every $\epsilon > 0$,

$$P\left(\left| -\frac{1}{n} \log p(X_1, \ldots, X_n) - H(X) \right| > \epsilon \right) \longrightarrow 0 \quad \text{as } n \to \infty.$$

Proof. The proof follows from the weak law of large numbers applied to the sequence

$$Z_n := -\frac{1}{n} \log p(X_1, \ldots, X_n) = -\frac{1}{n} \sum_{i=1}^{n} \log p(X_i), \quad n \geq 1,$$

and noting that $E[-\log p(X)] = H(X)$. □

The above theorem is perhaps the most important theorem in all of information theory. We shall now interpret the statement of the theorem in slightly greater detail. Recall that the theorem says that for any choice of $\epsilon > 0$, we have

$$P\left(\left| -\frac{1}{n} \log p(X_1, \ldots, X_n) - H(X) \right| > \epsilon \right) \longrightarrow 0 \quad \text{as } n \to \infty.$$

This means that given every choice of $\epsilon > 0$, there exists an integer N large enough (this integer N may vary with the choice of ϵ) such that

$$P\left(\left| -\frac{1}{n} \log p(X_1, \ldots, X_n) - H(X) \right| > \epsilon \right) \leq \epsilon \quad \text{for all } n \geq N.$$

In other words, for every choice of $\epsilon > 0$, there exists N sufficiently large such that

$$P\left(H(X) - \epsilon \leq -\frac{1}{n}\log p(X_1, \ldots, X_n) \leq H(X) + \epsilon\right) \geq 1 - \epsilon \quad \text{for all } n \geq N,$$

or equivalently,

$$P\left(2^{-n(H(X)+\epsilon)} \leq p(X_1, \ldots, X_n) \leq 2^{n(H(X)-\epsilon)}\right) \geq 1 - \epsilon \quad \text{for all } n \geq N.$$

That is, for all sufficiently large values of n, the value of the joint pmf of $(X_1, \ldots, X_n)$ is roughly of the order of $2^{-nH(X)}$ with high probability. Let us denote the set of all values of $(X_1, \ldots, X_n)$ which have this value of pmf by $A_\epsilon^{(n)}$. That is,

$$A_\epsilon^{(n)} \coloneqq \{(x_1, \ldots, x_n) \in \mathcal{X}^n : 2^{-n(H(X)+\epsilon)} \leq p(x_1, \ldots, x_n) \leq 2^{n(H(X)-\epsilon)}\}.$$

The AEP says that $P(A_\epsilon^{(n)}) \geq 1 - \epsilon$ for all n sufficiently large. The set $A_\epsilon^{(n)}$ is known as the **ϵ-typical set**.

We can easily show that the following are true.

Proposition 3.3.2. *For all sufficiently large values of n, we have:*

1. $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$.
2. $|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}$.

Proof. Exercise. □

Thus, for all sufficiently large values of n, the ϵ-typical set $A_\epsilon^{(n)}$ has roughly $2^{nH(X)}$ elements, with each element of $A_\epsilon^{(n)}$ having probability approximately equal to $2^{-nH(X)}$. Thus,

The ϵ-typical set is a high probability set whose elements are nearly uniformly distributed.

3.3.1 How Large is the ϵ-typical set? Does it Contain the Most Likely Sequence?

One of the most often encountered questions about the ϵ-typical set $A_\epsilon^{(n)}$ are about its size relative to the whole set $\mathcal{X}^n$ and whether or not it contains the most likely sequence. These questions mostly stem from the fact that $A_\epsilon^{(n)}$ has high probability for all sufficiently large n, and our notion of large probability sets is (perhaps) to visualise such sets as containing a large number of elements and/or to contain the most likely sequence. We shall demonstrate shortly that these notions are not true.

We first answer the question about the size of the ϵ-typical set relative to the whole set $\mathcal{X}^n$. Note that from Proposition 3.3.2, for every $\epsilon > 0$, for all sufficiently large values of n, we have

$$\begin{aligned}\frac{|A_\epsilon^{(n)}|}{|\mathcal{X}^n|} &\leq \frac{2^{n(H(X)+\epsilon)}}{2^{n\log|\mathcal{X}|}} \\ &= 2^{n(H(X)+\epsilon-\log|\mathcal{X}|)}.\end{aligned}$$

If $X \sim P$, where P is not equal to the uniform distribution on $\mathcal{X}$, then it follows that $H(X) < \log|\mathcal{X}|$, and therefore

$$\frac{|A_\epsilon^{(n)}|}{|\mathcal{X}^n|} \longrightarrow 0 \quad \text{as } n \to \infty,$$

provided $H(X)+\epsilon < \log|\mathcal{X}|$. This is certainly true for any choice of ϵ sufficiently small[1], say $\epsilon = \frac{\log|\mathcal{X}|-H(X)}{10^6}$. Thus, the important point to note here is the following:

If $X \sim P$ where P is not the uniform distribution, then the size of the ϵ-typical set relative to the whole set $\mathcal{X}^n$ is vanishingly small for all n sufficiently large.

We now turn our attention to answer the second question about whether the most likely sequence is present in the ϵ-typical set. Towards this, fix $\epsilon = 0.01$, and let $X_1, X_2, \ldots, X_n$ be iid Bernoulli distributed random variables distributed with mean 0.3. Then, the most likely sequence is the all-zeros with probability $p(0,\ldots,0) = (0.7)^n$. It can be shown that the entropy of Ber(0.3) distribution is roughly about 0.881 bits/symbol, and thus the 0.01-typical set is given by

$$A_{0.01}^{(n)} = \{(x_1,\ldots,x_n) \in \{0,1\}^n : (0.53)^n \leq p(x_1,\ldots,x_n) \leq (0.5466)^n\},$$

and it is clear that $(0,\ldots,0) \notin A_{0.01}^{(n)}$ for any $n \geq 1$.

3.4 Entropy and Hashing

We saw in the previous sections that the Shannon entropy represents the compressibility limit of any distribution. Higher the Shannon entropy, lesser our ability to compress. Also, we saw that any distribution P which is not the uniform distribution may be compressed further in the sense that a subset (known as the typical set) of size roughly $2^{H(P)}$ (or $2^{nH(P)}$ for n iid copies case) may be extracted and shown to have the properties that (a) each of its elements is nearly uniformly distributed, and (b) it has high probability. The typical set is then the most compressed version of the original set, and for all practical purposes, can not be compressed any further. It is therefore enough to work with the typical set.

We now provide an alternative viewpoint to Shannon entropy. This viewpoint arises as the answer to the following question on guessing.

Imagine that you are creating an account on Google for the first time. You are asked to choose a username and enter a password. Of course, you would not want Google to store your password for future logins. So, suppose that the Google server employs an algorithm to convert your password into an l-bit string of zeros and ones. In one of the subsequent login attempts, suppose that you forget your password and query the Google server to return you the password. Assuming that the Google server has a list of passwords used by users with high probability available at its disposal, the server then applies the same algorithm as before to create l-bit strings of zeros and ones for each one of these commonly used passwords, and if it finds a match between the l-bit strings of any one of the commonly used passwords with the l-bit string corresponding to your password, it returns that password to you (if there are multiple matchings, it sends one of them randomly). Else, it declares error. Given $\epsilon > 0$, what is the minimum number of bits l that the algorithm needs to use so that you can recover your password with probability at least $1-\epsilon$?

We will soon see that the answer turns out to be Shannon entropy. Before we get to the answer, we shall introduce a few notations. Let $\epsilon > 0$ be given.

- Let $X \sim P$ be a discrete random variable representing the password you enter. It takes values in the set of all possible passwords, which we denote by $\mathcal{X}$. Let p denote the pmf of X.

[1] You should always think of ϵ as being very small and should not be bothered about its presence when interpreting important results.

- The algorithm that the Google server uses to convert the password entered into an l-bit string of zeros and ones may be represented by a function

$$F : \mathcal{X} \to \{0,1\}^l.$$

chosen uniformly at random from the set of all such possible functions. This function is known as a hash. In other words, each time a user enters a password, the Google server samples an l-bit string uniformly at random from the set of all 2^l possible l-bit strings, and assigns the sampled string as hash to the entered password.

- Let A denote the set of commonly used passwords that the Google server has at its disposal. Let us assume that Google has pre-computed the probability of this set, and it satisfies $P(A) \geq 1 - \frac{\epsilon}{2}$. As a user, you do not know what this set A is. Neither don you know what the distribution P is. These are known only at the server.

- If X denotes the password you entered at the time of creating the account, and $\hat{X} \in A$ is such that $f(\hat{X}) = f(X)$, then the Google server returns $\hat{X}$ as your password. If there are multiple such $\hat{X}$, it returns one of them at random. Else, it gives up.

We would like to determine the smallest value of l such that $P(\hat{X} \neq X) \leq \epsilon$. Clearly, we have

$$\begin{aligned}
P(\hat{X} \neq X) &= \underbrace{P(\hat{X} \neq X,\ X \in A)}_{\text{your password is a commonly used password}} + \underbrace{P(\hat{X} \neq X,\ X \notin A)}_{\text{your password is not a commonly used password}} \\
&\leq P(\hat{X} \neq X,\ X \in A) + P(X \notin A) \\
&\leq P(\hat{X} \neq X,\ X \in A) + \frac{\epsilon}{2} \\
&= \sum_{x \in A} \frac{p(x)}{P(A)} \sum_{x' \in A : x' \neq x} \underbrace{P(F(x') = F(x))}_{= \frac{1}{2^l}} + \frac{\epsilon}{2} \\
&\leq \frac{|A|}{2^l} + \frac{\epsilon}{2}.
\end{aligned}$$

Suppose that $|A| \leq 2^\lambda$ for some $\lambda > 0$. Then, from the above set of inequalities, it follows that $P(\hat{X} \neq X) \leq \epsilon$ if

$$l \geq \lambda - \log \frac{\epsilon}{2}.$$

From Lemma 3.2.1, we know that the conditions $|A| \leq 2^\lambda$ and $P(A) \geq 1 - \frac{\epsilon}{2}$ may be achieved by choosing A as the set

$$A = \left\{ x \in \mathcal{X} : -\log p(X) \leq H(X) + \sqrt{\frac{\mathrm{Var}(-\log p(X))}{\epsilon/2}} \right\}$$

and

$$\lambda = H(X) + \sqrt{\frac{\mathrm{Var}(-\log p(X))}{\epsilon/2}}.$$

In summary, we have the following:

Entropy represents the minimum number of bits needed to represent the value of a random variable $X \sim P$ in order to recover it with high probability even when the distribution P is not known.

Example 3.4.1. (Hashing and the birthday paradox) We now show through an example that:

1. Given a set $\mathcal{H}$ of all possible hashes, and a hash $h^* \in \mathcal{H}$, if $H_1, \ldots, H_n$ are drawn iid uniformly at random from $\mathcal{H}$, the chance that at least of the H_i's is same as h^* is very small.
2. Given a set $\mathcal{H}$ of all possible hashes, if $H_1, \ldots, H_n$ are drawn iid uniformly at random from $\mathcal{H}$, the chance that at least two of the H_i's are identical is close to 50% even for moderate sizes of $\mathcal{H}$ (this has connections to the problem of birthday paradox).

1. Consider the set $\mathcal{H}$ of all l-bit hashes (l-bit strings of zeros and ones). Let h^* be the all zero string. If $H_1, \ldots, H_n \overset{iid}{\sim} \text{unif}(\mathcal{H})$, then

$$\begin{aligned} P(\exists\ 1 \leq i \leq n \text{ such that } H_i = h^*) &= 1 - P(H_i \neq h^* \text{ for all } 1 \leq i \leq n) \\ &\overset{(a)}{=} 1 - (P(H_1 \neq h^*))^n \\ &= 1 - \left(\frac{2^l - 1}{2^l}\right)^n, \end{aligned}$$

   where (a) above follows because H_i's are iid. Note that for $l = 20$ (i.e., 2^{20} possible hashes, which is very common to have in practice) and $n = 2^{10}$, the above probability is almost 0.

2. Consider the set $\mathcal{H}$ of all l-bit hashes (l-bit strings of zeros and ones). Let h^* be the all zero string. Let $H_1, \ldots, H_n \overset{iid}{\sim} \text{unif}(\mathcal{H})$. Denote by N the number of (i, j) pairs, $i \neq j$, such that $H_i = H_j$. That is,

$$N = \sum_{(i,j):i \neq j} 1_{\{H_i = H_j\}},$$

   where 1_A is a random variable that is one if the event A occurs, and zero otherwise. Clearly, we can see that

$$N = \left(\sum_{(i,j):i<j} 1_{\{H_i = H_j\}} \right) \times 2.$$

   We are interested in computing the probability $P(N \geq 1)$. In order to do so, we shall compute $P(N = 0)$, and then obtain the required probability as $P(N \geq 1) = 1 - P(N = 0)$.

   Now, we note that $N = 0$ if and only if for whatever value H_1 takes, $H_2 \neq H_1$ and $H_3 \neq H_1$ or H_2 and $H_4 \neq H_1$ or H_2 or H_3 and so on.

$$\begin{aligned} P(N = 0) &= P(H_2 \neq H_1, H_3 \neq H_2 \text{ or } H_1, \ldots) \\ &= \left(\frac{2^l - 1}{2^l}\right) \cdot \left(\frac{2^l - 2}{2^l}\right) \cdots \left(\frac{2^l - n + 1}{2^l}\right) \\ &= \frac{(2^l)!}{2^{ln} \cdot (2^l - n)!}. \end{aligned}$$

   It follows from the above expression that the expression for $P(N = 0)$ is not easy to evaluate in closed form even for moderate values of l and n.

   However, there are several approximations available for calculating the required probability $P(N \geq 1)$ directly. Denoting this probability as a function of n by $p(n)$, one of the very crude approximations[2] is $p(n) \approx \frac{n^2}{2 \cdot 2^l}$. Then, we note that for $l = 20$ and $n = 2^{10}$, we have $P(N \geq 1) \approx 0.5$, which is about 50% chance that there exist at least two hashes which are identical.

[2] See `https://en.wikipedia.org/wiki/Birthday_problem` for a list of commonly used approximations.

Remark 3. Aside from the existing approximations provided in the link in the footnote below, note that N is a special type of random variable. It is a sum of Bernoulli random variables. It is left as an exercise to deduce the distribution of N. However, it is important to note here that inequalities like the Chebyshev's inequality come very handy in obtaining approximations for probabilities such as $P(N \geq 1)$. See homework 1 for one such exercise.