A Mini-Introduction to Information Theory

Christopher Ting

Hiroshima University

✉: cting@hiroshima-u.ac.jp

☺: http://cting.x10host.com/

☎: +81 082-424-6451

A1棟 131-1

Table of Contents

What is information theory?

- Information theory is concerned with
  1. representing data in a compact fashion (a task known as **data compression** or **source coding**)
  2. transmitting and storing data in a way that is robust to errors (a task known as **error correction** or **channel coding**).
- Quantification of the **amount of information** in **events**, random variables, and distributions.
- Use of **probabilities** p_i, $i = 1, 2, \ldots, m$.

Communication Channel

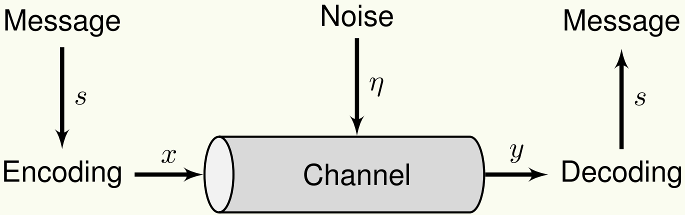


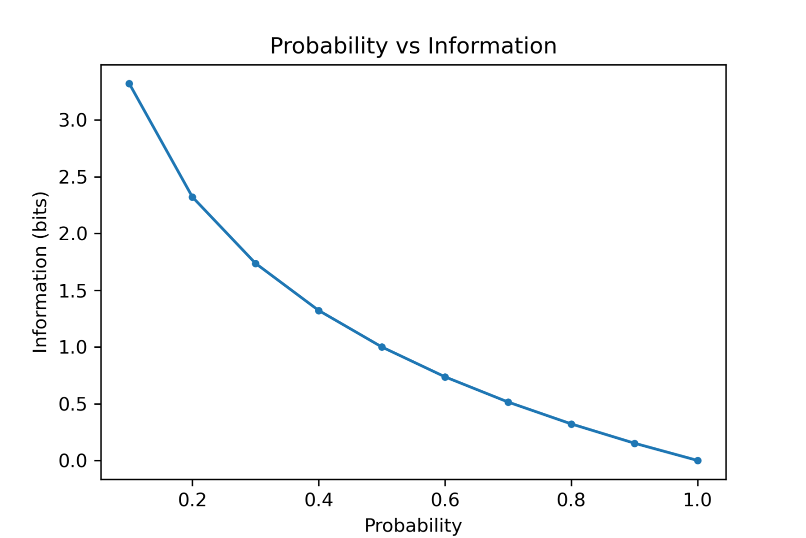
Figure: A message (data) is **encoded** before being used as input to a **communication channel**, which adds noise. The channel output is **decoded** by a receiver to recover the message.

How to measure information?

- The intuition behind quantifying information is the idea of measuring how much **surprise** there is in an event.
- Those **events** that are rare (low **probability**) are more surprising and therefore have more information than those events that are common (high probability).
- Rare events are more **uncertain** or more **surprising** and require more information to represent them than common events.
- Based on the above-mentioned intuition, if the probability of an event is p, then the information or surprise S of seeing the event is

$$\text{Surprise} \propto \frac{1}{p}.$$

Plot


Probability vs Information
Information (bits)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.2
0.4
0.6
0.8
1.0
Probability

Shannon Information

- We can calculate the **amount of information** in an event using the **probability** of the **event**.
- **Information** and can be calculated for a **discrete event** x as follows:

$$\text{surprise}(x) \equiv \text{information}(x) := \log_2\left(\frac{1}{p(x)}\right) = -\log_2 p(x),$$

  where $\log_2$ is the base-2 logarithm and $p(x)$ is the probability of the event x.
- The choice of the **base-2 logarithm** means that the unit of the information measure is in **bits**.
- In the information processing sense, Shannon information is the number of bits required to describe the event.

Asking the Right Question

- Suppose you receive a message that consists of a string of symbols a or b, say

$$aababbaaaab\cdots \tag{1}$$

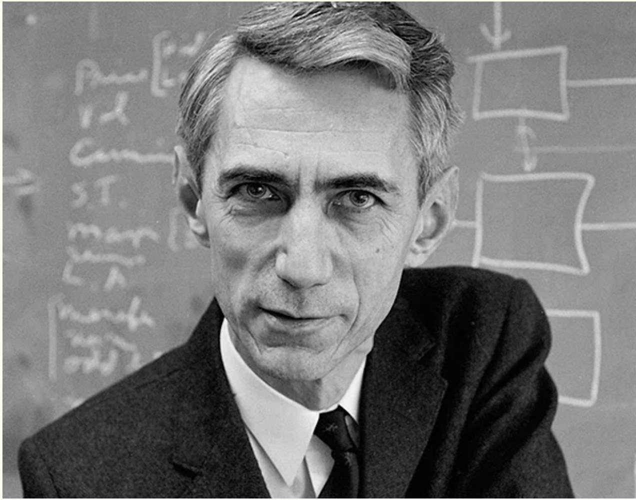
- Suppose that a occurs with probability p, and b with probability $1-p$. How many **bits of information** can you extract from a long message with N letters?
- Each message of length N in this question is the **event**.

Definition of Shannon Entropy

- For large N, the message will consist very nearly of pN occurrences of a and $(1-p)N$ occurrences of b.
- The number of such messages is

$$\frac{N!}{(pN)!((1-p)N)!} = 2^{NS} \qquad (2)$$

where S is the **Shannon entropy** per letter [Shannon (1948)].



Heroes of Tech: Claude Shannon

Multinomial Distribution

- In the 2-letter case, we can also write pN as n_a and $(1-p)N$ as n_b, leading to

$$S = \frac{1}{N}\log_2\left(\frac{N!}{n_a!n_b!}\right).$$

- In general, when we have m letters, the number of possible ways W to write a message of length N becomes

$$W = \frac{N!}{n_1!\,n_2!\,\ldots\,n_m!}.$$

- We want to obtain the following formula

$$S = -\sum_{i=1}^{m} p_i \log_2 p_i, \tag{3}$$

where $p_i = \dfrac{n_i}{N}$ for $i = 1, \ldots, m$.

Gamma Function

- Stirling's approximation is
$$\ln n! \cong n \ln n - n. \tag{4}$$
- To prove (4), we first consider the **Gamma function**
$$\Gamma(x+1) = \int_0^\infty t^x e^{-t} dt. \tag{5}$$
- Note that
$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt,$$
and by the **method of integration by parts**, we obtain the **recursion formula** for the Gamma function:
$$\Gamma(x+1) = x\Gamma(x).$$

Proof of the Recursion Formula

- Let $u = t^x$ for (5). So $du = xt^{x-1}dt$.
- And $dv = e^{-t}dt$, which integrates to $v = -e^{-t}$.
- The integration by parts formula is $\int u dv = uv - \int v du$. Hence

$$
\begin{aligned}
\Gamma(x+1) &= -t^x e^{-t}\Big|_0^\infty + \int_0^\infty e^{-t} x t^{x-1} dt \\
&= 0 + x \int_0^\infty t^{x-1} e^{-t} dt \\
&= x\Gamma(x).
\end{aligned}
$$

□

Factorial

- If x is a positive integer n, the recursive formula allows us to obtain

$$\Gamma(n+1) = n!.$$

- The proof is simple.

$$\begin{aligned}\Gamma(n+1) &= n\Gamma(n) = n(n-1)\Gamma(n-1) = n(n-1)(n-2)\Gamma(n-2) = \cdots \\ &= n(n-1)(n-2)\cdots 1\Gamma(1) \\ &= n!\Gamma(1).\end{aligned}$$

- From the definition, $\Gamma(1) = \displaystyle\int_0^\infty e^{-t}dt = -e^{-t}\Big|_0^\infty = 1.$ $\square$

Critical Point

- The **integrand** of the **Gamma function** is a function of t

$$f(t) = t^x e^{-t}. \tag{6}$$

- Differentiation yields

$$f'(t) = xt^{x-1}e^{-t} - t^x e^{-t}. \tag{7}$$

- To find the **critical point**, we let $f'(t) = 0$. Thus, we solve for t that satisfies the **first-order condition**:

$$xt^{x-1}e^{-t} - t^x e^{-t} = 0.$$

- The solution is $x = t$. Hence $f(x) = \left(\dfrac{x}{e}\right)^x$.

Maximum

- To examine whether the **critical point** $t = x$ is the maximum or minimum point, we differential $f'(t)$ to obtain

$$f''(t) = x(x-1)t^{x-2}e^{-t} - xt^{x-1}e^{-t} - \left(xt^{x-1}e^{-t} - t^x e^{-t}\right).$$

- At $t = x$,

$$\begin{aligned} f''(x) &= x(x-1)x^{x-2}e^{-x} - xx^{x-1}e^{-x} - \left(xx^{x-1}e^{-x} - x^x e^{-x}\right) \\ &= (x-1)x^{x-1}e^{-x} - x^x e^{-x} - \left(x^x e^{-x} - x^x e^{-x}\right) \\ &= e^{-x}\left((x-1)x^{x-1} - x^x\right) \\ &= -e^{-x}x^{x-1} \\ &< 0. \end{aligned}$$

- Therefore at $t = x$, the integrand $f(t)$ is at its maximum.

Scaling the Gamma Function

- Starting from (5), let us divide both sides of the equation by the maximum value of $(x/e)^x$, so that the new **integrand** is a function that has a maximum value of 1 where $t = x$.

$$\left(\frac{e}{x}\right)^x \Gamma(x+1) = \int_0^\infty \left(\frac{t}{x}\right)^x e^{-(t-x)} dt.$$

- Now, make a small **change of variable**. Let $s = t - x$, so that

$$\left(\frac{e}{x}\right)^x \Gamma(x+1) = \int_{-x}^\infty \left(1 + \frac{s}{x}\right)^x e^{-s} ds = \int_{-x}^\infty g(s) ds.$$

- Since we want to obtain an approximation for large values of x, let us try to obtain an expansion of $g(s)$ as a series in s/x.

Gaussian Integral

- A convenient way of obtaining the expansion is to take the logarithm of $g(x)$:

$$\ln g(s) = x \ln \left(1 + \frac{s}{x}\right) - s.$$

- If $|s| < x$, the **Maclaurin expansion** of the natural logarithm is

$$\ln g(s) = x \left(\frac{s}{x} - \frac{1}{2} \left(\frac{s}{x}\right)^2 + \cdots \right) - s. \tag{8}$$

- If x is sufficiently large, this becomes $\ln g(s) \approx -\dfrac{s^2}{2x}$, and we obtain

$$\left(\frac{e}{x}\right)^x \Gamma(x+1) \approx \int_{-\infty}^{\infty} \exp\left(-\frac{s^2}{2x}\right) ds.$$

- The integral on the right-hand side is the well known **Gaussian integral** and the result is $\sqrt{2\pi x}$.

Stirling Approximation

- Thus for large x,

$$\Gamma(x+1) \approx \left(\frac{x}{e}\right)^x \sqrt{2\pi x}.$$

- If x is an integer,

$$n! \approx \left(\frac{n}{e}\right)^n \sqrt{2\pi n}$$

- Taking logarithms on both sides, we obtain

$$\ln n! \approx \left(n + \frac{1}{2}\right) \ln n - n + \ln \sqrt{2\pi},$$

- Since n is large, we obtain the **Stirling approximation**

$$\ln n! \approx n \ln n - n.$$

Remarks

- For very large N, we can make the further approximation

$$\ln N! \approx N \ln N \tag{9}$$

- For smaller numbers, the following approximation is remarkably good:

$$\ln n! \approx \left(n + \frac{1}{2}\right) \ln n - n + \frac{1}{12n} + \ln \sqrt{2\pi}.$$

- This approximation can be obtained when (8) is expanded as

$$\ln g(s) \approx x \left(\frac{s}{x} - \frac{1}{2}\left(\frac{s}{x}\right)^2 + \frac{1}{3}\left(\frac{s}{x}\right)^3 + \cdots \right) - s.$$

Shannon Information for m Letters

- For very large N, the **Shannon information** is

$$S = \frac{1}{N}\log_2 W_2 = \frac{1}{N}\log_2 \frac{N!}{n_1!\,n_2!\,\cdots\,n_m!} = \frac{1}{N}\log_2 \frac{N!}{(Np_1)!\,(Np_2)!\,\ldots\,(Np_m)!}$$

$$= \frac{1}{N}\left(\log_2 N! - \sum_{i=1}^{m} \log_2((Np_i)!)\right).$$

- Now, the logarithmic base transformation formula is

$$x = 2^{\log_2 x} \quad \Longrightarrow \quad \ln x = \ln\left(2^{\log_2 x}\right) = \log_2 x \cdot \ln 2$$

- For very large N, we use (9) to obtain

$$\log_2 N! = \frac{1}{\ln 2}\ln N! \approx \frac{1}{\ln 2} N \ln N.$$

Calculation with the Properties of Logarithm

- Given that probabilities sum to 1, we obtain

$$
\begin{aligned}
S = \frac{1}{N}\log_2 W &= \frac{1}{N}\left(N\log_2 N - \sum_{i=1}^{m} Np_i \log_2(Np_i)\right) \\
&= \log_2 N - \sum_{i=1}^{m} p_i \log_2(Np_i) = \log_2 N - \log_2 N \sum_{i=1}^{m} p_i - \sum_{i=1}^{m} p_i \log_2 p_i \\
&= \left(1 - \sum_{i=1}^{m} p_i\right)\log_2 N - \sum_{i=1}^{m} p_i \log_2 p_i \\
&= -\sum_{i=1}^{m} p_i \log_2 p_i
\end{aligned}
$$

Summary: Shannon Information

- Therefore, we have derived that the **Shannon information** S in **bits** is,

$$S = -\sum_{i=1}^{m} p_i \log_2 p_i. \tag{10}$$

- Note that we can also write S as

$$S = \sum_{i=1}^{m} p_i \log_2 \left(\frac{1}{p_i}\right).$$

- We may interpret S as the average **surprise**, which is commonly known as **information entropy**.

Special Case: Binary

- When $m = 2$, **information entropy** is

$$S = -p\log_2 p - (1-p)\log_2(1-p) = -p\log_2 p - \log_2(1-p) + p\log_2(1-p). \qquad (11)$$

- Thus, S is a function of p. Let's find the **critical point** of $S(p)$.

$$\begin{aligned}\frac{dS}{dp} &= -p\frac{1}{p\ln 2} - \log_2 p + \frac{1}{(1-p)\ln 2} + \log_2(1-p) - \frac{p}{(1-p)\ln 2}\\ &= -\frac{1}{\ln 2} - \log_2 p + \log_2(1-p) + \frac{1}{(1-p)\ln 2} - \frac{p}{(1-p)\ln 2}\\ &= -\log_2 p + \log_2(1-p)\end{aligned}$$

- The **first-order condition** $\dfrac{dS}{dp} = 0$ gives rise to

$$\log_2\left(\frac{1-p}{p}\right) = 0.$$

Maximum Information Entropy

- Since $\log_2 1 = 0$ regardless of base, it must be that

$$\frac{1-p}{p} = 1.$$

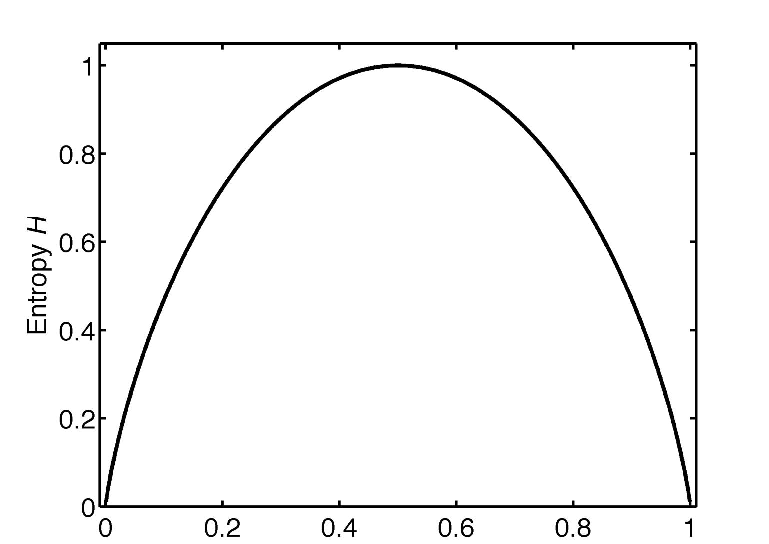
- It turns out that the **critical point** $p = \dfrac{1}{2}$.

- To examine whether it is a maximum or minimum,

$$\frac{d^2 S}{dp^2} = -\frac{1}{p \ln 2} - \frac{1}{(1-p)\ln 2} < 0 \qquad \text{for all } 0 < p < 1.$$

- Therefore, we have an information maximum of $S(1/2) = \dfrac{1}{2} + \dfrac{1}{2} = 1$ bit.

Graph of Binary Information Entropy



General Case: Uniform Probability 1/3

- The probabilities are $p_1, p_2, \ldots, p_m$.
- Without loss of generality, suppose $p_1 < p_2$. Let $\epsilon > 0$ be a tiny positive number such that
$$p_1 + \varepsilon < p_2 - \varepsilon.$$
- That is, these two probabilities are getting closer to each other, and their difference is getting smaller.
- The entropy of $\{p_1 + \varepsilon, p_2 - \varepsilon, p_3, \ldots, p_m\}$ minus the entropy of $\{p_1, p_2, p_3, \ldots, p_m\}$ is

$$\begin{aligned}
&-(p_1+\varepsilon)\log_2(p_1+\varepsilon) - (p_2-\varepsilon)\log_2(p_2-\varepsilon) - [-p_1\log_2 p_1 - p_2\log_2 p_2] \\
&= -p_1\log_2\left(\frac{p_1+\varepsilon}{p_1}\right) - \varepsilon\log_2(p_1+\varepsilon) - p_2\log_2\left(\frac{p_2-\varepsilon}{p_2}\right) + \varepsilon\log_2(p_2-\varepsilon)
\end{aligned}$$

General Case: Uniform Probability 2/3

- Note that
$$p_1 + \varepsilon = p_1\left(1 + \frac{\varepsilon}{p_1}\right) \qquad \text{and} \qquad p_2 - \varepsilon = p_2\left(1 - \frac{\varepsilon}{p_2}\right).$$

- So, we can rewrite the difference as
$$-p_1 \log_2\left(1 + \frac{\varepsilon}{p_1}\right) - \varepsilon\left(\log_2 p_1 + \log_2\left(1 + \frac{\varepsilon}{p_1}\right)\right)$$
$$-p_2 \log_2\left(1 - \frac{\varepsilon}{p_2}\right) + \varepsilon\left(\log_2 p_2 + \log_2\left(1 - \frac{\varepsilon}{p_2}\right)\right).$$

- Recalling that $\log_2(1 + x) = \dfrac{1}{\ln 2}x + O(x^2)$ for small x, the difference simplifies to
$$-\frac{1}{\ln 2}\varepsilon - \varepsilon \log_2 p_1 + \frac{1}{\ln 2}\varepsilon + \varepsilon \log_2 p_2 + O(\varepsilon^2) = \varepsilon \log_2(p_2/p_1) + O(\varepsilon^2).$$

General Case: Uniform Probability 3/3

- Since $\varepsilon > 0$ and $p_1 < p_2$, the entropy difference

$$\varepsilon \log_2(p_2/p_1) + O(\varepsilon^2) > 0.$$

- Therefore we have shown that by using ε to make the first two probabilities get closer to each other, the entropy becomes larger.
- The same argument and calculation apply to every pair from $\{p_1, p_2, p_3, \ldots, p_m\}$.
- The implication is that **maximum entropy** is obtained when

$$p_1 = p_2 = \cdots = p_m = \frac{1}{m}.$$

Communication over a Noisy Channel

- Alice is trying to communicate with Bob, and she sends a message that consists of many letters, each being an instance of a random variable X whose possible values are $x_1, \ldots, x_m$.
- She sends the message over a noisy telephone connection, and what Bob receives is many copies of a random variable Y, drawn from an alphabet with letters $y_1, \cdots, y_r$.
- How many **bits** of **information** does Bob get after Alice has transmitted a message with N letters?

Sending x and Hearing y

- Suppose that $P_{X,Y}(x_i, y_j)$ is the probability that Alice sends $X = x_i$ whereas Bob hears $Y = y_j$.
- The probability that Bob hears $Y = y_j$ is

$$P_Y(y_j) = \sum_i P_{X,Y}(x_i, y_j). \tag{12}$$

- It is a sum over all possible letters of what Alice has sent.

Conditional Probability

- If Bob does hear $Y = y_j$, his estimate of the probability that Alice sent x_i is the **conditional probability**

$$P_{X|Y}(x_i|y_j) = \frac{P_{X,Y}(x_i, y_j)}{P_Y(y_j)}. \tag{13}$$

- From Bob's point of view, once he has heard $Y = y_j$, his estimate of the remaining entropy in Alice's signal is the **Shannon entropy** of the conditional probability distribution.

$$S_{X|Y=y_j} = -\sum_i P_{X|Y}(x_i|y_j) \log_2(P_{X|Y}(x_i|y_j)). \tag{14}$$

Average Remaining Entropy

- Averaging over all possible values of Y, the **average remaining entropy**, once Bob has heard Y, is

$$\begin{aligned}\sum_j P_Y(y_j) S_{X|Y=y_j} &= -\sum_j P_Y(y_j) \sum_i \frac{P_{X,Y}(x_i, y_j)}{P_Y(y_j)} \log_2\left(\frac{P_{X,Y}(x_i, y_j)}{P_Y(y_j)}\right) \\ &= -\sum_{i,j} P_{X,Y}(x_i, y_j) \log_2 P_{X,Y}(x_i, y_j) + \sum_{i,j} P_{X,Y}(x_i, y_j) \log_2 P_Y(y_j) \\ &= S_{XY} - S_Y . \end{aligned} \tag{15}$$

- Here S_{XY} is the entropy of the **joint distribution** $P_{X,Y}(x_i, y_j)$ for the pair X, Y .

- S_Y is the entropy of the probability distribution $P_Y(y_j) = \sum_i P_{X,Y}(x_i, y_j)$ for Y only.

Mutual Information

- $S_{XY} - S_Y$, is called the **conditional entropy** $S_{X|Y}$.
- It is the entropy that remains in the probability distribution X once Y is known.
- From the left-hand side of (15), which is a sum of ordinary entropies $S_{X|Y=y_j}$ with positive coefficients ($P_Y(y_j)$), it must be that the conditional entropy satisfies

$$S_{XY} - S_Y \geq 0. \tag{16}$$

- Since S_X is the total information content in Alice's message, and $S_{XY} - S_Y$ is the information content that Bob still does not have after observing Y, it follows that the information about X that Bob *does* gain when he receives Y is the difference or

$$I(X;Y) = S_X - S_{XY} + S_Y. \tag{17}$$

- Here $I(X;Y)$ is called the **mutual information** between X and Y. It measures how much we learn about X by observing Y, and vice versa.

Motivation for Relative Information

- Suppose that we are observing a random variable X, for example the final state in the decays of a radioactive nucleus.
- We have a theory that predicts a **probability distribution** Q_X. for the final state.
- The **probability** to observe state $X = x_i$, where i runs over a set of possible outcomes $\{1, 2, \ldots, m\}$, is $q_i = Q_X(x_i)$.
- But maybe our theory is wrong and the decay is actually described by some different probability distribution P_X, such that the probability of $X = x_i$ is $p_i = P_X(x_i)$.

How Correct is Q_X?

- If the correct **probability distribution** is P_X, then after observing N decays, we will see outcome x_i approximately $p_i N$ times.
- Believing Q_X to be the correct distribution, we will judge the probability of what we have seen to be

$$\mathcal{P} = \prod_{i=1}^{m} q_i^{p_i N} \frac{N!}{\prod_{j=1}^{m}(p_j N)!}. \tag{18}$$

- Here $\frac{N!}{\prod_{j=1}^{m}(p_j N)!}$ is the number of sequences in which outcome x_i occurs $p_i N$ times
- Assuming that the initial hypothesis Q_X is correct, $\prod_{i=1}^{m} q_i^{p_i N}$ is the probability of any specific such sequence.

Kullback-Liebler Divergence

- We have already calculated that for large N, $\dfrac{N!}{\prod_{j=1}^{s}(p_j N)!} \sim 2^{-N\sum_i p_i \log_2 p_i}$, and $\displaystyle\prod_{i=1}^{m} q_i^{p_i N} = 2^{N\sum_i p_i \log_2 q_i}$, so

$$\mathcal{P} \sim 2^{-N\sum_i p_i(\log_2 p_i - \log_2 q_i)}. \tag{19}$$

- Let (19) be $2^{-D(P_X||Q_X)}$ where the **Kullback-Liebler divergence** is given by

$$D(P_X||Q_X) = \sum_i p_i(\log_2 p_i - \log_2 q_i). \tag{20}$$

- From the derivation, $D(P_X||Q_X)$ is clearly nonnegative, and zero only if $P_X = Q_X$, i.e., if the initial hypothesis is correct.

Remarks

- If the initial hypothesis is wrong, we will be sure of this once

$$N D(P_X || Q_X) \gg 1. \tag{21}$$

- The **Kullback-Liebler divergence** $D(P_X || Q_X)$ is an important measure of the difference between two probability distributions P_X and Q_X of the random variable X.
- Note that it is asymmetric, i.e., $D(P_X || Q_X) \neq D(Q_X || P_X)$.
- This asymmetry is a result of our assumption that Q_X is our initial hypothesis whereas P_X is the correct answer.

Example of Kullback-Liebler Divergence

x	0	1	2
Distribution $P(x)$	9/25	12/25	4/25
Distribution $Q(x)$	1/3	1/3	1/3

$$
\begin{aligned}
D(P \parallel Q) &= \sum_{x \in \mathcal{X}} P(x) \log_2 \left(\frac{P(x)}{Q(x)} \right) \\
&= \frac{9}{25} \log_2 \left(\frac{9/25}{1/3} \right) + \frac{12}{25} \log_2 \left(\frac{12/25}{1/3} \right) + \frac{4}{25} \log_2 \left(\frac{4/25}{1/3} \right) \approx 0.123,
\end{aligned}
$$

$$
\begin{aligned}
D(Q \parallel P) &= \sum_{x \in \mathcal{X}} Q(x) \log_2 \left(\frac{Q(x)}{P(x)} \right) \\
&= \frac{1}{3} \log_2 \left(\frac{9/25}{1/3} \right) + \frac{1}{3} \log_2 \left(\frac{12/25}{1/3} \right) + \frac{1}{3} \log_2 \left(\frac{4/25}{1/3} \right) \approx 0.141,
\end{aligned}
$$

Joint Probability

- The **joint probability** distribution is denoted by $P_{X,Y}(x_i, y_j)$, for two possibly correlated random variables X and Y.
- The separate probability distributions for X and for Y are obtained by "**integrating out**" or summing over the other variable:

$$P_X(x_i) = \sum_j P_{X,Y}(x_i, y_j), \qquad P_Y(y_j) = \sum_i P_{X,Y}(x_i, y_j). \tag{22}$$

- We define a second probability distribution for X, Y by ignoring the **correlations** between them:

$$Q_{X,Y}(x_i, y_j) = P_X(x_i) P_Y(y_j). \tag{23}$$

Sub-additivity of Entropy

- Now we define and calculate the **entropy** between these two distributions:

$$\begin{aligned} S(P_{X,Y}|Q_{X,Y}) :=& \sum_{i,j} P_{X,Y}(x_i, y_j)(\log_2 P_{X,Y}(x_i, y_j) - \log_2(P_X(x_i)P_Y(y_j))) \\ =& \sum_{i,j} P_{X,Y}(x_i, y_j)(\log_2 P_{X,Y}(x_i, y_j) - \log_2 P_X(x_i) - \log_2 P_Y(y_j)) \\ =& S_X + S_Y - S_{XY} = I(X;Y). \end{aligned} \quad (24)$$

- Thus **mutual information** $I(X;Y) \geq 0$, with equality only if the two distributions are the same, meaning that X and Y were uncorrelated to begin with.

- The property

$$S_X + S_Y - S_{XY} \geq 0 \quad (25)$$

is called **sub-additivity** of entropy.

Main Idea

- Financial market is an **information channel**.
- Prices traded over a time period form a message.
- A sequence of prices is a message generated by traders, which reflect their individual information.
- What is the amount of information in such message of m unique prices?

Tick-by-Tick Prices

+ Over a time period, every trade is recorded.

$$\pi_1, \pi_2, \ldots, \pi_N.$$

Time	Price	Vol
75952	38290	1
75952	38295	1
75952	38295	1
75952	38295	2
75952	38290	1
75952	38290	1
75956	38285	1
75957	38280	1
75957	38280	1
75957	38280	1

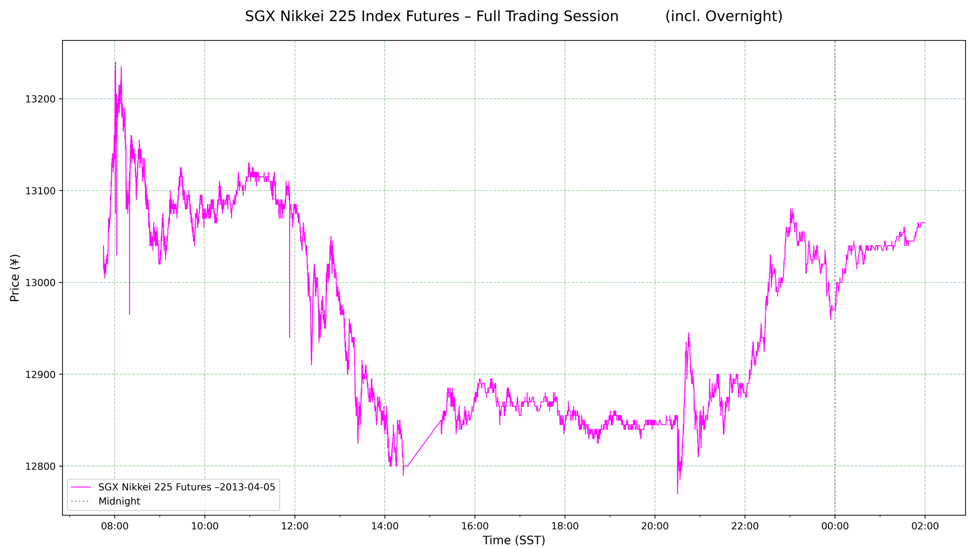
Time	Price	Vol
75958	38275	1
75958	38270	1
75958	38275	1
75958	38275	3
75958	38275	1
75958	38275	1
75959	38270	1
75959	38270	1
80000	38265	1
80000	38260	1

Time	Price	Vol
80000	38260	1
80000	38260	1
80000	38255	1
80000	38255	1
80000	38260	1
80000	38260	1
80000	38265	1
80000	38265	1
80000	38265	1
80000	38265	1

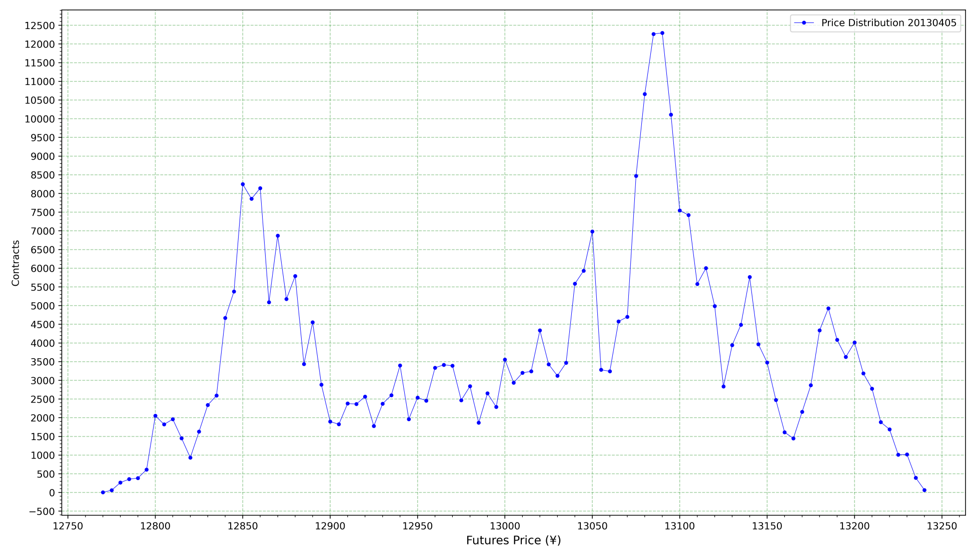
Offshore Nikkei Futures

- Launched in 1986
- First equity index futures contract in Asia
- First futures contract based on the Japanese stock market.
- In 2014, extended trading from 7:30 AM to next day's 2 AM Singapore time.
- Contract size multiplier: ¥500
- Minimum price fluctuation is 5 index points

NK Futures Time Series of Traded Prices



Contracts Traded at a Given Price



5 Questions

1. What is the information entropy of the NK futures' tick-by-tick price data assigned to you? The random variable X here is the number of contracts traded at a given price.

2. What is the Kullback-Liebler divergence if the hypothesized probability Q_X is uniform?

3. Consider the price change Y from one tick to the next. Let the possible outcomes be UP, DOWN, and NO CHANGE. What is the information entropy of price movements for your data? (The first trade is assumed to be NO CHANGE.)

4. What is the corresponding Kullback-Liebler divergence for Question 3 if the correct probability distribution of price movement is equally likely?

5. What is the value of the conditional entropy $S(X|Y)$?

Requirements

- Write a report to answer all the questions.
- You may use Japanese to answer the questions.
- You must use the provided `demo3B.py` and `demo3C.py` to plot the time series of futures prices and the volume distribution for each unique price traded.
- These two figures must be included in the report.
- You must use the provided latex template file to write your report.
- You must submit all the python codes you have written.
- Finally, submit the tex source and the pdf file generated by it.

Quantities of Information 1/2

- Surprise: $-\log_2 p$ bits
- Stirling's aproximation for large N: $\ln N! \approx N \ln N$.
- Shannon entropy = Shannon information: $S = -\sum_{i=1}^{m} p_i \log_2 p_i$
- Maximum entropy: Equal or uniform probability
- Shannon entropy of the conditional probability distribution

$$S_{X|Y=y_j} = -\sum_i P_{X|Y}(x_i|y_j)\log_2(P_{X|Y}(x_i|y_j)).$$

Quantities of Information 2/2

- Entropy of the joint probability distribution

$$S_{XY} = -\sum_{i,j} P_{X,Y}(x_i, y_j) \log_2 P_{X,Y}(x_i, y_j)$$

- Conditional entropy: $S(X|Y) = S_{XY} - S_Y$
- Mutual information: $I(X, Y) = S(P_X|Q_Y) = S_X - S_{XY} + S_Y$
- Kullback-Liebler divergence: $D(P_X||Q_X) = \sum_i p_i(\log_2 p_i - \log_2 q_i)$

Keywords

References

Shannon, C.E., 1948. A mathematical theory of communication. The Bell System Technical Journal 27, 379–423, 623–656. URL: https://people.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf.