Maximum Entropy Principle, Normal Distribution, and Central Limit Theorem

Xu Chenrui

20/12/2023

1 Introduction

The Maximum Entropy Principle states that, given some constraints, the probability distribution that maximizes entropy is the one that best represents the current system. Intuitively, as entropy is a measure of uncertainty or randomness in a probability distribution, the principle asserts that the most unbiased distribution—that is, the one with the highest entropy—should always be chosen. The principle has found applications in a wide range of fields, including natural language processing, image recognition, speech recognition, and financial analysis..

During the lecture, we've already proved that some classical probability distribution maximize the entropy of the system under certain conditions. For example, the normal distribution maximize the entropy of the system given variance constraints. But can we go further to derive the original PDF of normal distribution by the maximum entropy principle? Such kind of question is philosophically more important: it require us to use the intuitive beliefs to produce feasible tools. It will also provide new perspectives towards the basic formulas, and the relationships between information theories and probability theories.

In this report, we will investigate the mathematical foundations of Maximum Entropy. We will also discuss some applications of Maximum Entropy in basic probability theories. By better understanding the Maximum Entropy Principle and its applications, we can find more essential relationships between probability theories and information theories. .

Questions

The project is guided by following questions:

- How to derive the general formula of the best (entropy maximizing) distribution of collections of r.v. given some constraints?
- Can we further derive the PDF of Normal distribution (as well as other classical distributions) by Maximum Entropy Principle (MEP)?
  - Why is this distribution so "common"?
  - Why there is a "π" in the PDF of normal distribution?

(Remark: we can abstractly "define" the normal distribution as somehow the "best" distribution among all distributions given certain variance constraints.)

- The Central Limit Theorem (CLT) implies that the sum of a large number of iid r.v. will tend to be normally distributed, so are there any relationships between the MEP and the CLT? Can we use the MEP to prove the CLT?

Notations

- **MEP**: Maximum Entropy Principle
- **CLT**: Central Limit Theorem
- $\mathbb{X}_{\mathbb{R}}(C_1, C_2, ...C_n)$: the collection of continuous random variables defined on $\mathbb{R}$ under constraints $(C_1, C_2, ...C_n)$, where C_i is the i^{th} moment constraint. In this paper, we only need to consider the case $n = 2$.
- **EMD** (Entropy Maximizing Distribution): we call the distribution with density f as **EMD** if f maximize the differential entropy among all $X \in \mathbb{X}_{\mathbb{R}}(C_1, C_2, ...C_n)$.

2 Maximum Entropy Distributions

In this section, we derive the general formula of the the Entropy Maximizing Distribution of collections of r.v. given some constraints.

2.1 Problem Formulation

We formulate the problem as an optimization problem. For the simplicity of further explanation, we first consider the **Discrete Case**:

$$\begin{aligned}
max. \quad & H(p) = -\sum_{i=1}^{n} p_i \log p_i; \\
s.t. \quad & p_i \geq 0; \\
& \sum_{i=1}^{n} p_i = 1; \\
& \sum_{i=1}^{n} p_i r_{ij} = \alpha_i;
\end{aligned}$$

where the last line illustrate the constraints.
We solve the optimal solution by Lagrangian multiplier method:

$$L(p) = -\sum_{i=1}^{n} p_i \log p_i + \lambda_0 \left(\sum_{i=1}^{n} p_i - 1 \right) + \sum_{j=1}^{m} \lambda_j \left(\sum_{i=1}^{n} p_i r_{ij} - \alpha_i \right).$$

We take derivative with respect to pi and let it to be 0:

$$L'(p) = -1 - \log p_i + \lambda_0 + \sum_{j=1}^{m} \lambda_j r_{ij} = 0.$$

Then we get the solution:

$$\tilde{p_i} = \frac{e^{\sum_{j=1}^{m} \lambda_j r_{ij}}}{e^{1-\lambda_0}} = e^{\lambda_0 - 1 + \sum_{j=1}^{m} \lambda_j r_{ij}}.$$

Now we extend our method to the **Continuous Case**:

$$\begin{aligned} max. \quad & h(p) = -\int_s f \log f \\ s.t. \quad & f(x) \geq 0 \\ & \int_s f(x)dx = 1 \\ & \int_s f(x) r_i(x) dx = \alpha_i \end{aligned}$$

Similarly we use Lagrangian multiplier method (but here, take the derivative of $L(f)$ w.r.t f) to solve that:

$$\tilde{f}(x) = e^{\lambda_0 - 1 + \sum_{j=1}^{m} \lambda_j r_{ij}(x)}. \tag{1}$$

3 Maximum Entropy and Normal Distribution

In this section, we derive the formula of the density of the "Normal Distribution" by **MEP**. Here we use the following definition:

Definition: Normal Distribution The normal distribution is defined to be the **EMD** (Entropy Maximizing Distribution) of the collection of r.v. $\mathbb{X}_{\mathbb{R}}(\mathbb{E}X = \mu, \mathbb{E}X^2 = \sigma^2 + \mu^2)$, where μ and σ are given constant (please refer to the introduction section for the notation).

Theorem 1 The probability density function of the "Normal Distribution" is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

(Remark: this consist with what we have learned about $X \sim \mathcal{N}(\mu, \sigma^2)$.)

3.1 Analysis on the Constraints

By referring to the general formula that we derived in section 2, we can assume that $f(x) = e^{\lambda_0 - 1 + \lambda_1 x + \lambda_2 x^2}$. Now we only need to determined the $\lambda_i' s$ in the method of Lagrangian multiplier.

3.1.1 Analysis on the Probability (0th moment) Constraint

First, we analyze the Probability Constraint. The idea to eliminate the variable x in the formula is to make the integrand easily calculated. After observation we find the intergrand is very similar to the Gaussian integral, so we try to rephrase the e- exponential quadratic formula from standard form to normal form, and put the constant out of the integral. Then we will have:

$$
\begin{aligned}
\int_{-\infty}^{\infty} f(x)dx &= \int_{-\infty}^{\infty} e^{\lambda_0 - 1 + \lambda_1 x + \lambda_2 x^2} dx \\
&= \int_{-\infty}^{\infty} \exp\left(\lambda_2 \left[\left(x + \frac{\lambda_1}{2\lambda_2}\right)^2 - \frac{\lambda_1^2 - 4\lambda_0\lambda_2 + 4\lambda_2}{4\lambda_2^2}\right]\right) dx \\
&= \int_{-\infty}^{\infty} \exp\left(\lambda_2 \left(x + \frac{\lambda_1}{2\lambda_2}\right)^2 - \frac{\lambda_1^2 - 4\lambda_0\lambda_2 + 4\lambda_2}{4\lambda_2}\right) dx \\
&= \exp\left(-\frac{\lambda_1^2 - 4\lambda_0\lambda_2 + 4\lambda_2}{4\lambda_2}\right) \int_{-\infty}^{\infty} \exp\left(\lambda_2 \left(x + \frac{\lambda_1}{2\lambda_2}\right)^2\right) dx \\
&= \exp\left(-\frac{\lambda_1^2 - 4\lambda_0\lambda_2 + 4\lambda_2}{4\lambda_2}\right) \int_{-\infty}^{\infty} \exp\left(-(-\lambda_2)\left(x + \frac{\lambda_1}{2\lambda_2}\right)^2\right) dx \\
&= \exp\left(-\frac{\lambda_1^2 - 4\lambda_0\lambda_2 + 4\lambda_2}{4\lambda_2}\right) \int_{-\infty}^{\infty} \exp\left(-\left(\sqrt{-\lambda_2}\left(x + \frac{\lambda_1}{2\lambda_2}\right)\right)^2\right) dx \\
&= \frac{1}{\sqrt{-\lambda_2}} \exp\left(-\frac{\lambda_1^2 - 4\lambda_0\lambda_2 + 4\lambda_2}{4\lambda_2}\right) \cdot \\
&\quad \int_{-\infty}^{\infty} \exp\left(-\left(\sqrt{-\lambda_2}\left(x + \frac{\lambda_1}{2\lambda_2}\right)\right)^2\right) d\left(\sqrt{-\lambda_2}\left(x + \frac{\lambda_1}{2\lambda_2}\right)\right)
\end{aligned}
$$

For simplicity, we can change the variable $y = \sqrt{-\lambda_2}\left(x + \frac{\lambda_1}{2\lambda_2}\right)$. Notice that that probability constraint states that $\int_{-\infty}^{\infty} f(x)dx = 1$. Therefore, we have

$$
\begin{aligned}
\int_{-\infty}^{\infty} f(x)dx &= \frac{1}{\sqrt{-\lambda_2}} \exp\left(-\frac{\lambda_1^2 - 4\lambda_0\lambda_2 + 4\lambda_2}{4\lambda_2}\right) \int_{-\infty}^{\infty} e^{-y^2} dy \\
&= \frac{1}{\sqrt{-\lambda_2}} \exp\left(-\frac{\lambda_1^2 - 4\lambda_0\lambda_2 + 4\lambda_2}{4\lambda_2}\right) \sqrt{\pi} = 1
\end{aligned}
$$

Note that $\int_{-\infty}^{\infty} e^{-y^2} dy$ is the Gaussian Integral, we can easily calculate it by double integration on the polar coordinates:

$$
\begin{aligned}
\int_{-\infty}^{\infty} e^{-y^2} dx &= \sqrt{\int_{-\infty}^{\infty} e^{-y^2} dy \cdot \int_{-\infty}^{\infty} e^{-x^2} dx} = \sqrt{\int_0^{2\pi} \int_0^{\infty} e^{-r^2} r dr d\theta} \\
&= \sqrt{2\pi \left[-\frac{1}{2} e^{-r^2}\right]_0^{\infty}} = \sqrt{\pi}.
\end{aligned}
$$

In conclusion, we have

$$\frac{1}{\sqrt{-\lambda_2}} \exp\left(-\frac{\lambda_1^2 - 4\lambda_0\lambda_2 + 4\lambda_2}{4\lambda_2}\right)\sqrt{\pi} = 1 \tag{2}$$

3.1.2 Analysis on the Mean (1st moment) Constraint

By the similar process as the previous analysis, we have:

$$\begin{aligned}
\int_{-\infty}^{\infty} xf(x)dx &= \int_{-\infty}^{\infty} xe^{\lambda_0 - 1 + \lambda_1 x + \lambda_2 x^2} dx \\
&= \frac{1}{\sqrt{-\lambda_2}} \exp\left(-\frac{\lambda_1^2 - 4\lambda_0\lambda_2 + 4\lambda_2}{4\lambda_2}\right) \int_{-\infty}^{\infty} x \exp\left(-\left(\sqrt{-\lambda_2}(x + \frac{\lambda_1}{2\lambda_2})\right)^2\right) \\
&\quad d\left(\sqrt{-\lambda_2}(x + \frac{\lambda_1}{2\lambda_2})\right) \\
&= \frac{1}{\sqrt{-\pi\lambda_2}} \int_{-\infty}^{\infty} \left(\sqrt{-\lambda_2}(x + \frac{\lambda_1}{2\lambda_2}) - \frac{\lambda_1\sqrt{-\lambda_2}}{2\lambda_2}\right) \\
&\qquad \exp\left(-\left(\sqrt{-\lambda_2}(x + \frac{\lambda_1}{2\lambda_2})\right)^2\right) d\left(\sqrt{-\lambda_2}(x + \frac{\lambda_1}{2\lambda_2})\right)
\end{aligned}$$

Similarly, we set $y = \sqrt{-\lambda_2}(x + \frac{\lambda_1}{2\lambda_2})$, then:

$$\begin{aligned}
\int_{-\infty}^{\infty} xf(x)dx &= \frac{1}{\sqrt{-\pi\lambda_2}} \int_{-\infty}^{\infty} \left(y - \frac{\lambda_1\sqrt{-\lambda_2}}{2\lambda_2}\right) e^{-y^2} dy \\
&= \frac{1}{\sqrt{-\pi\lambda_2}} \left(\left(\int_{-\infty}^{\infty} ye^{-y^2} dy\right) - \left(\frac{\lambda_1\sqrt{-\lambda_2}}{2\lambda_2} \int_{-\infty}^{\infty} e^{-y^2} dy\right)\right) \\
&= \frac{1}{\sqrt{-\pi\lambda_2}} \left(0 - \frac{\lambda_1\sqrt{-\lambda_2}}{2\lambda_2}\sqrt{\pi}\right) = -\frac{\lambda_1}{2\lambda_2} = \mu
\end{aligned}$$

Note: for the second last line of the above steps, the first integral is easily calculated, and the second integral is the Gaussian integral (refer the previous sections). So we have:

$$-\frac{\lambda_1}{2\lambda_2} = \mu \tag{3}$$

3.1.3 Analysis on the Variance (2nd moment) Constraint

Notice that $var[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2 = \sigma^2$.
By the similar process as the previous analysis, we have:

$$
\begin{aligned}
&\int_{-\infty}^{\infty} x^2 f(x)dx \\
&= \int_{-\infty}^{\infty} x^2 e^{\lambda_0 - 1 + \lambda_1 x + \lambda_2 x^2} dx \\
&= \frac{1}{\sqrt{-\lambda_2}} \exp\Big(- \frac{\lambda_1^2 - 4\lambda_0\lambda_2 + 4\lambda_2}{4\lambda_2} \Big) \int_{-\infty}^{\infty} x^2 \exp\Big(- \Big(\sqrt{-\lambda_2}(x + \frac{\lambda_1}{2\lambda_2}) \Big)^2 \Big) \\
&\quad d\Big(\sqrt{-\lambda_2}(x + \frac{\lambda_1}{2\lambda_2}) \Big) \\
&= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} \Big(\frac{1}{-\lambda_2} \Big(\sqrt{-\lambda_2}(x + \frac{\lambda_1}{2\lambda_2}) \Big)^2 - \frac{\lambda_1}{\lambda_2} x - \frac{\lambda_1^2}{4\lambda_2^2} \Big) \\
&\quad \exp\Big(- \Big(\sqrt{-\lambda_2}(x + \frac{\lambda_1}{2\lambda_2}) \Big)^2 \Big) d\Big(\sqrt{-\lambda_2}(x + \frac{\lambda_1}{2\lambda_2}) \Big) \\
&= \frac{1}{\sqrt{\pi}} \Bigg(\frac{1}{-\lambda_2} \int_{-\infty}^{\infty} \Big(\sqrt{-\lambda_2}(x + \frac{\lambda_1}{2\lambda_2}) \Big)^2 \exp\Big(- \Big(\sqrt{-\lambda_2}(x + \frac{\lambda_1}{2\lambda_2}) \Big)^2 \Big) d\Big(\sqrt{-\lambda_2}(x + \frac{\lambda_1}{2\lambda_2}) \Big) \\
&\quad - \frac{\lambda_1}{\lambda_2} \int_{-\infty}^{\infty} x \exp\Big(- \Big(\sqrt{-\lambda_2}(x + \frac{\lambda_1}{2\lambda_2}) \Big)^2 \Big) d\Big(\sqrt{-\lambda_2}(x + \frac{\lambda_1}{2\lambda_2}) \Big) \\
&\quad - \frac{\lambda_1^2}{4\lambda_2^2} \int_{-\infty}^{\infty} \exp\Big(- \Big(\sqrt{-\lambda_2}(x + \frac{\lambda_1}{2\lambda_2}) \Big)^2 \Big) d\Big(\sqrt{-\lambda_2}(x + \frac{\lambda_1}{2\lambda_2}) \Big) \Bigg)
\end{aligned}
$$

Similarly, we set $y = \sqrt{-\lambda_2}(x + \frac{\lambda_1}{2\lambda_2})$, then:

$$
\begin{aligned}
\int_{-\infty}^{\infty} x^2 f(x)dx &= \frac{1}{\sqrt{\pi}} \Bigg(\frac{1}{-\lambda_2} \int_{-\infty}^{\infty} y^2 e^{-y^2} dy - \frac{\lambda_1}{\lambda_2} \int_{-\infty}^{\infty} y e^{-y^2} dy - \frac{\lambda_1^2}{4\lambda_2^2} \int_{-\infty}^{\infty} e^{-y^2} dy \Bigg) \\
&= \frac{1}{\sqrt{\pi}} \Bigg(\frac{1}{-\lambda_2} \frac{\sqrt{\pi}}{2} - \frac{\lambda_1}{\lambda_2} \Big(- \frac{\lambda_1}{2\lambda_2} \sqrt{\pi} \Big) - \frac{\lambda_1^2}{4\lambda_2^2} \sqrt{\pi} \Bigg) \\
&= -\frac{1}{2\lambda_2^2} + \frac{\lambda_1^2}{4\lambda_2^2} = \mu^2 + \sigma^2
\end{aligned}
$$

So we have:

$$
-\frac{1}{2\lambda_2^2} + \frac{\lambda_1^2}{4\lambda_2^2} = \mu^2 + \sigma^2 \tag{4}
$$

3.2 A Summary of the Previous Results and Deriving the PDF of Normal Distribution

We gather together the results of the three previous analysis (2),(3) and (4):

$$\begin{cases} \dfrac{1}{\sqrt{-\lambda_2}} \exp\left(-\dfrac{\lambda_1^2 - 4\lambda_0\lambda_2 + 4\lambda_2}{4\lambda_2}\right)\sqrt{\pi} = 1 \\ -\dfrac{\lambda_1}{2\lambda_2} = \mu \\ -\dfrac{1}{2\lambda_2^2} + \dfrac{\lambda_1^2}{4\lambda_2^2} = \mu^2 + \sigma^2 \end{cases}$$

Then we can solve the following results:

$$\begin{cases} \lambda_0 = \log\left(\dfrac{1}{\sqrt{2\pi}\sigma}\right) - \dfrac{\mu^2}{2\sigma^2} + 1 \\ \lambda_1 = \dfrac{\mu}{\sigma^2} \\ \lambda_2 = -\dfrac{1}{2\sigma^2} \end{cases}$$

By plugging λ_0, λ_1 and λ_2 into the formula (1) of $f(x)$ that we obtained in section 2, we have:

$$f(x) = e^{\lambda_0 - 1 + \lambda_1 x + \lambda_2 x^2} = \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

That finished the proof of theorem 1. □

4 Using Maximum Entropy to Prove CLT

The Central Limit Theorem (CLT) is one of the most important theorems in elementary probability theories. It implies that the sum of a large number of iid r.v. will tend to be normally distributed. In this section, we discuss the relationship between MEP and CLT, specifically, we try to prove the CLT by MEP. First, we state the version of CLT that we want to discuss here:

Theorem 2 (Central Limit Theorem) If $\bar{X}$ is the mean of a sequence of iid random sample $X_1, X_2, \ldots, X_n$ of size n from a distribution with a finite mean μ and a finite positive variance σ^2, then the distribution of

$$S_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\sum_{i=1}^{n}(X_i - \mu)}{\sqrt{n}\sigma}$$

converges to $\mathcal{N}(0, 1)$ as $n \to \infty$.

4.1 Intuition

Notice that, as $n \to \infty$

$$\mathbb{E}(W) = \mathbb{E}\left(\frac{\sum_{i=1}^{n}(X_i - \mu)}{\sqrt{n}\sigma}\right) = 0$$

$$\mathbb{E}(W^2) = \mathbb{E}^2(W) + var(W) = var\left(\frac{\sum_{i=1}^{n}(X_i - \mu)}{\sqrt{n}\sigma}\right) = 1$$

And further there are no more other constraints assumed in the CLT, therefore, by maximum entropy principle, W should follow the distribution $\mathcal{N}(0,1)$, as we have derived above.

4.2 Notations, Assumptions and Preliminaries for the Proof

Recall the relative entropy $D(f\|g)$ between two densities f and g, which is defined by

$$D(f\|g) = \int_S f \log \frac{f}{g},$$

by convention, $0 \log \frac{0}{0}$ is set to be 0.
Denote ϕ as the PDF of the standard normal distribution. Let $S_n = \frac{\sum_{i=1}^{n}(X_i-\mu)}{\sqrt{n}\sigma}$, and suppose f_n, defined on $\mathbb{R}$, is the PDF of each S_n. Then we have $\int_{\mathbb{R}} f_n(x)dx = 1$ and $\int_{\mathbb{R}} f_n(x)x^2dx = \mathbb{E}[S_n^2] = 1$.
We also claim the following lemma without proof:

Lemma Let X and Y be two r.v. in $\mathbb{X}_{\mathbb{R}}(C_1, C_2, ...C_n)$. If the distributions of X and Y both maximizing the entropy among $\mathbb{X}_{\mathbb{R}}(C_1, C_2, ...C_n)$, then $X \triangleq Y$ in distribution with probability 1.

4.3 Proof of the CLT by MEP

Consider the relative entropy of f_n and ϕ, we have:

$$\begin{aligned}
D(f_n\|\phi) &= \int_{\mathbb{R}} f_n \log(f_n/\phi) \\
&= -h(S_n) - \int_{\mathbb{R}} f_n \log(\phi) \\
&= -h(S_n) - \int_{\mathbb{R}} f_n(x) \log\left(\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\right) dx \\
&= -h(S_n) + \int_{\mathbb{R}} f_n(x)\left[\log(\sqrt{2\pi}) + \frac{x^2}{2}\right] dx \\
&= -h(S_n) + \log(\sqrt{2\pi}) \int_{\mathbb{R}} f_n(x)dx + \frac{1}{2}\int_{\mathbb{R}} f_n(x)x^2dx \\
&= -h(S_n) + \frac{1}{2}\log(2\pi) + \frac{1}{2} = -h(S_n) + \frac{1}{2}\log 2\pi e
\end{aligned}$$

Notice that $\log 2\pi e$ is the differential entropy of the standard normal random variable. By **MEP** we need to maximize $h(S_n)$. In summary we have:

$$h(S_n) = h(\phi) - D(f_n \| \phi) \tag{5}$$

Therefore, to maximize the differential entropy $h(S_n)$ is equivalent to minimize $D(f_n\|\phi)$ for $n \to \infty$. Notice that $D(f_n\|\phi)$ is always larger than or equal to 0, so we might guess that the $D(f_n\|\phi)$ should converge to 0 as $n \to \infty$. In fact, there is a theorem that states this important results, which is provided by A.R.Barron [2]:

Theorem 3 (Barron) (Under the previous assumptions and notations) $nD(f_n\|\phi)$ is a subadditive sequence. In particular $D(f_{2n}\|\phi) \leq D(f_n\|\phi)$. Furthermore, the relative entropy converges to zero

$$\lim_{n\to\infty} D(f_n\|\phi) = 0$$

if and only if $D(f_n\|\phi)$ is finite for some n.

The detailed proof of the theorem can be found in Barron's original paper.
Theorem 3 implies that as $n \to \infty$, S_n will achieve the maximum differential entropy, which is exactly the differential entropy of the standard normal distribution. By the previous lemma, we claim that S_n converges to ϕ in probability as $n \to \infty$. This finish the proof of the Central Limit Theorem. □

4.4 Further Properties of the Convergence of $\{S_n\}$

In fact, the sequence constructed in the CLT has much better properties. There are several interesting theorems representing those conclusions:

Theorem 4 (Shannon-Stam) If X and Y are independent and identically distributed, then the normalized sum

$$\frac{X+Y}{\sqrt{2}}$$

has entropy at least that of X and Y.

There is a more general conclusion for integer n:

Theorem 5 (Artstein, Ball, Barthe, Naor) [1] If X_i are iid random variables with finite variance, then the normalized sums

$$W_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$$

have increasing entropy.

The above two theorems implies that, during the process of $n \to 0$ in CLT, the normalized sum S_n somehow **monotonously** converges to normal random variable. In deed, it's quite hard to identify this kind of **monotonousness** just by density function. However, the differential entropy provide us an excellent approach to illustrate the **monotonousness**. We might treat (differential) entropy as a measure of the ”goodness” or ”simplicity” of a collection of probability distribution.

5 Conclusion

In this report, we discussed the formula of the maximum entropy distribution and explored the application of the maximum entropy principle to derive the density function of the normal distribution. We also use the maximum entropy principle to prove the famous Central Limit Theorem. These results highlight the versatility of the maximum entropy principle in probability theory. The principle also provide us very intuitive and thought-provoking perspective toward the probability theories.

References

[1] Artstein, S., Ball, K. M., Barthe, F., and Naor, A. (2004). Solution of Shannon’s Problem on the Monotonicity of Entropy. Journal of the American Mathematical Society, 17(4), 975–982. https://doi.org/10.1090/S0894-0347-04-00459-X

[2] Barron, A. R. (1986). Entropy and the Central Limit Theorem. The Annals of Probability, 14(1), 336–342.https://doi.org/10.1214/aop/1176992632

[3] Cover, T. M., and Thomas, J. A. (2006). Elements of information theory (2nd ed.). Wiley-Interscience.