Introduction to Continuous Entropy

Charles Marsh
Department of Computer Science
Princeton University
crmarsh@princeton.edu

December 13, 2013

Abstract

Classically, Shannon entropy was formalized over discrete probability distributions. However, the concept of entropy can be extended to continuous distributions through a quantity known as *continuous* (or *differential*) entropy. The most common definition for continuous entropy is seemingly straightforward; however, further analysis reveals a number of shortcomings that render it far less useful than it appears. Instead, *relative entropy* (or *KL divergence*) proves to be the key to information theory in the continuous case, as the notion of comparing entropy across probability distributions retains value. Expanding off this notion, we present several results in the field of *maximum entropy* and, in particular, conclude with an information-theoretic proof of the Central Limit Theorem using continuous relative entropy.

1 Introduction

Much discussion of information theory implicitly or explicitly assumes the (exclusive) usage of discrete probability distributions. However, many of information theory's key results and principles can be extended to the continuous case–that is, to operate over continuous probability distributions. In particular, *continuous* (or *differential*) entropy is seen as the continuous-case extension of Shannon entropy. In this paper, we define and evaluate continuous entropy, relative entropy, maximum entropy, and several other topics in continuous information theory, concluding with an information-theoretic proof of the Central Limit Theorem using the techniques introduced throughout.

1.1 Goals

More specifically, our goals are as follows:

1. Introduce and evaluate a definition for continuous entropy.

2. Discuss some results of maximum entropy (i.e., for distributions with fixed mean, fixed variance, finite support, etc.).
3. Derive the Central Limit Theorem using information-theoretic principles.

2 Continuous Entropy

2.1 A Definition

Information theory truly began with Shannon entropy, i.e., entropy in the discrete case. While we will not review the concept extensively, recall the definition:

Definition (Shannon entropy)**.** *The Shannon entropy $h(X)$ of a discrete random variable X with distribution $P(x)$ is defined as:*

$$H(X) = \Sigma_i P(x_i) \log \tfrac{1}{P(x_i)}$$

The formula for continuous entropy is a (seemingly) logical extension of the discrete case. In fact, we merely replace the summation with an integral.

Definition (Continuous entropy)**.** *The continuous entropy $h(X)$ of a continuous random variable X with density $f(x)$ is defined as:*

$$h(X) = \int_S f(x) \log \tfrac{1}{f(x)} dx$$

where S is the support set of the random variable.[6] As shorthand, we can also write $H(f) = H(X)$, where random variable X has distribution $f(x)$.

To see how continuous entropy operates in the wild, consider the following example.

2.2 An Example: The Uniform Distribution

Allow f to be the uniform distribution on $[a, b]$. That is:

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{for } x \in [a, b] \\ 0, & \text{else} \end{cases}$$

Let's solve for the continuous entropy of this distribution.

$$\begin{aligned} h(f) &= \int_S f(x) \log \frac{1}{f(x)} dx \\ &= \int_a^b \frac{1}{b-a} \log (b-a) dx \\ &= \frac{1}{b-a} \log (b-a) \int_a^b dx \\ &= \log (b-a) \end{aligned}$$

Informally, the continuous entropy of the uniform distribution is equal to the log of the width of the interval.

2.3 Weaknesses

The definition of continuous entropy provided seems to follow quite naturally from Shannon entropy. But rigorously, how well does it help up? Is it a "good" extension of Shannon entropy?

As we'll show, there are a number of 'kinks' or weaknesses with our definition of continuous entropy. In the discrete case, we had a set of axioms from which we derived Shannon entropy and thus a bunch of nice properties that it exhibited. In the continuous case, however, our definition is highly problematic–to the point that, on its own, it may not be an entirely useful mathematical quantity.

2.3.1 Shannon entropy in the Limit

As mentioned earlier, Shannon entropy was *derived* from a set of axioms. But our definition for continuous entropy was provided with no such derivation. Where does the definition actually *come* from?

The natural approach to deriving continuous entropy would be to take discrete entropy in the limit of n, the number of symbols in our distribution. This is equivalent to rigorously defining integrals in calculus using a Reimannian approach: it makes sense that the continuous case would come from extending the discrete case towards infinity.

To begin, we discretize our continuous distribution f into bins of size Δ. By the Mean Value Theorem, we get that there exists an x_i such that $f(x_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} f(x)dx$. This implies that we can approximate f by a Reimann sum:

$$\int_{-\infty}^{\infty} f(x)dx = \lim_{\Delta \to 0} \Sigma_{i=-\infty}^{\infty} f(x_i)\Delta$$

Claim. *Continuous entropy differs from Shannon entropy in the limit by a potentially infinite offset.*

Proof.

$$\begin{aligned}
H^{\Delta} &= -\lim_{\Delta \to 0} \Sigma_{i=-\infty}^{\infty} f(x_i)\Delta \log\left(f(x_i)\Delta\right) \\
&= -\lim_{\Delta \to 0} \Sigma_{i=-\infty}^{\infty} f(x_i)\Delta \log\left(f(x_i)\right) - \lim_{\Delta \to 0} \Sigma_{i=-\infty}^{\infty} f(x_i)\Delta \log \Delta \\
\lim_{\Delta \to 0} \Sigma_{i=-\infty}^{\infty} f(x_i)\Delta &= \int_{-\infty}^{\infty} f(x)dx = 1 \\
\lim_{\Delta \to 0} \Sigma_{i=-\infty}^{\infty} f(x_i)\Delta \log\left(f(x_i)\right) &= \int_{-\infty}^{\infty} f(x) \log f(x) dx \\
H^{\Delta} &= -\int_{-\infty}^{\infty} f(x) \log f(x) dx - \lim_{\Delta \to 0} \Sigma_{i=-\infty}^{\infty} f(x_i)\Delta
\end{aligned}$$

Ideally, we'd have that H^{Δ} were equal to our definition for continuos entropy, as it represents Shannon entropy in the limit. But note that $\log(\Delta) \to -\infty$ as

$\Delta \to 0$. As a result, the right term will explode. So instead, we need a special definition for continuous entropy:

$$h(f) = \lim_{\Delta \to 0} (H^{\Delta} + \log \Delta) = -\int_{-\infty}^{\infty} f(x) \log f(x) dx$$

In this sense, continuous entropy differs in the limit by an infinite offset! □

This demonstrates that the formula for continuous entropy **is not a derivation of anything**, unlike Shannon entropy–it's merely the result of replacing the summation with an integration.[8] This result may not be a problem in and of itself, but it helps to explain some of the proceeding difficulties with the definition.

2.3.2 Variable Under Change of Coordinates

$h(X)$ is *variant* under change of variables. Depending on your coordinate system, a distribution might have a different continuous entropy. This shouldn't be the case–but it is. Informally, this means that the same underlying distribution, represented with different variables, might not have the same continuous entropy.

To understand why, note that the probability contained in a differential area should *not* alter under change of variables. That is, for x, y:

$$|f_Y(y)dy| = |f_X(x)dx|$$

Further, define $g(x)$ to be the mapping from x to y, and $g^{-1}(y)$, its inverse. Then, we get:

Lemma 2.1. $f_Y(y) = \frac{d}{dy}(g^{-1}(y)) f_X(g^{-1}(y))$

Proof.

$$\begin{aligned} f_Y(y) &= \frac{dx}{dy} f_X(x) \\ &= \frac{d}{dy}(x) f_X(x) \\ &= \frac{d}{dy}(g^{-1}(y)) f_X(g^{-1}(y)) \end{aligned}$$ □

We'll use this fact in the following example[2]: Say, abstractly, that you have an infinite distribution of circles. Let $p(x)$ be the distribution of their radii and $q(w)$, the distribution of their areas. Further, $x(w) = \sqrt{\frac{w}{\pi}}$ and $w(x) = \pi x^2$. You'd expect this distribution to have the same continuous entropy regardless of its representation, In fact, we'll show that $H(p) \neq H(q)$.

Claim. $H(p) \neq H(q)$

Proof.

$$\begin{aligned}
p(x) &= |\frac{d}{dx}(g^{-1}(x))|q(w) \\
&= w'(x)q(w) \\
&= 2x\pi q(w) \\
\text{Thus: } q(w) &= \frac{p(x)}{2x\pi} \\
\text{Therefore: } H(w) &= \int q(w) \log \frac{1}{q(w)} dw \\
&= \int \frac{p(x)}{2x\pi} \log \frac{2x\pi}{p(x)} (2x\pi dx) \\
&= \int p(x)(\log (2x\pi) - \log (p(x)))dx \\
&= \int p(x) \log (2x\pi)dx + \int p(x) \log \frac{1}{p(x)} dx \\
&= H(x) + \int p(x) \log (2x\pi)dx \\
&\neq H(x)
\end{aligned}$$

□

To quote Shannon: "The scale of measurements sets an arbitrary zero corresponding to a uniform distribution over a unit volume".[8] The implication here is that all continuous entropy quantities are somehow relative to the coordinate system in-use. Further, one could extend this argument to say that *continuous entropy is useless* when viewed on its own. In particular, relative entropy between distributions could be the valuable quantity (which we'll see later on).

2.3.3 Scale Variant

Generalizing this result, we can also get that continuous entropy is not scale invariant.

Theorem 2.2. *If $Y = \alpha X$, then $h(Y) = h(X) + \log |\alpha|$.[14]*

Proof.

$$\begin{aligned}
h(Y) &= h(X) - \mathbb{E}[\log |\frac{dx}{dy}|] \\
&= h(X) - \mathbb{E}[\log \frac{1}{\alpha}] \\
&= h(X) + \log |\alpha|
\end{aligned}$$

□

2.3.4 Negativity & Information Content

With Shannon entropy, we had this wonderful intuition in which it represented the 'information content' of a discrete distribution. That is, Shannon entropy

could also be defined as the "expected value of the information of the distribution" or the number of bits you'd need to reliably encode n symbols. In the continuous case, this intuition deteriorates as $h(X)$ does not give you the amount of information in X.

To see why, note that $h(X)$ can be **negative**! For example: if X is uniformly distributed in $[0, \frac{1}{2}]$, then $h(X) = \log(\frac{1}{2} - 0) = \log \frac{1}{2} = -1$. If entropy can be negative, how can this quantity have any relationship to the information content of X?

2.4 An Alternative Definition

E.T. Jaynes[8] argued that we should define an invariant factor $m(X)$ that defines the density (*note: not probability density*) of a discrete distribution in the limit.

Definition. *Suppose we have a discrete set $\{x_i\}$ of an increasingly dense distribution. The invariant factor $m(X)$ is defined as:*

$$\lim_{n\to\infty} \frac{1}{n}(\textit{number of points in } a < x < b) = \int_a^b m(x)dx$$

This would give us an alternative definition of continuous entropy that is invariant under change of variables.

Definition. *Let X be a random variable with probability distribution $p(X)$. An alternative definition of the entropy $H(X)$ follows:*

$$H(X) = -\int_S p(x) \log \frac{p(x)}{m(x)} dx$$
$$\textit{where } S \textit{ is the support set of } X.$$

We provide this definition solely for educational purposes. The rest of the paper will assume that $H(X) = \int_S p(x) \log \frac{1}{p(x)} dx$.

2.5 Continuous Relative Entropy (KL divergence)

Despite the aforementioned flaws, there's hope yet for information theory in the continuous case. A key result is that definitions for relative entropy and mutual information follow naturally from the discrete case and retain their usefulness.

Let's go ahead and define relative entropy in the continuous case, using the definition in [6].

Definition. *The relative entropy $D(f||g)$ of two PDFs f and g is defined as:*

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

where S is the support set of f. Note that $D(f||g) = 0$ if $supp(g) \notin supp(f)$.

2.5.1 Non-Negativity of Relative Entropy

Importantly, relative entropy remains non-negative in the continuous case. We prove this using Jensen's Inequality[4].

Theorem 2.3. *For any two distributions f and g:*

$$D(f||g) \geq 0$$

Proof.

$$\begin{aligned}
D(p||q) &= \int p(x) \log \frac{p(x)}{q(x)} dx \\
&= \mathop{\mathbb{E}}_{p}[\log \frac{p(X)}{q(X)}] \\
&= \mathop{\mathbb{E}}_{p}[-\log \frac{q(X)}{p(X)}] \\
&\geq -\log \mathop{\mathbb{E}}_{p}[\frac{q(X)}{p(X)}] \text{ by Jensen's Inequality} \\
&= -\log \int p(x) \frac{q(x)}{p(x)} dx \\
&= -\log \int q(x) dx \\
&= -\log 1 \\
&= 0
\end{aligned}$$

□

2.5.2 Using Relative Entropy to Prove Upper Bounds

Before we advance, it's worth formalizing a key lemma that follows from the non-negativity of relative entropy.

Lemma 2.4. *If f and g are continuous probability distributions, then:*

$$h(f) \leq -\int f(x) \log g(x) dx$$

Proof. Using relative entropy.

$$\begin{aligned}
D(f||g) &= \int f(x) \log \frac{f(x)}{g(x)} dx \\
&= \int f(x)(\log(f(x)) - \log(g(x)))dx \\
&= \int f(x) \log(f(x))dx - \int f(x) \log(g(x))dx \\
&= -\int f(x) \log \frac{1}{(f(x))} dx - \int f(x) \log(g(x))dx \\
&= -h(x) - \int f(x) \log(g(x))dx \\
&= -h(x) - \int f(x) \log(g(x))dx \\
&\geq 0 \\
\text{Therefore: } h(x) &\leq -\int f(x) \log(g(x))dx
\end{aligned}$$

□

We can use this lemma to prove upper bounds on the entropy of probability distributions given certain constraints. Examples will follow in the proceeding sections.

2.6 Continuous Mutual Information

We can use our definition of relative entropy to define mutual information for continuous distributions as well. Recall that in the discrete case, we had:

$$I(X;Y) = D(p(x,y)||p(x)p(y))$$

We'll use this statement to define mutual information for continuous distributions[6].

Definition. *The mutual information $I(X;Y)$ of two random variables X and Y drawn from continuous probability distributions is defined as:*

$$\begin{aligned}
I(X;Y) &= D(p(x,y)||p(x)p(y)) \\
&= \int p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dx
\end{aligned}$$

3 Maximum Entropy

Now that we've defined and analyzed continuous entropy, we can now focus on some interesting results that follow from our formulation. Recall that the entropy of a continuous distribution is a highly problematic quantity as it is variant under change of coordinates, potentially non-negative, etc. The true quantity of interest, then, is the *relative* entropy between (sets of) distributions. This leads us to examine the problem of *maximum entropy*, defined in [5] as follows:

Definition. *The maximum entropy problem is to find a probability distribution that maximizes entropy satisfying some set of constraints.*

Intuitively, the maximum entropy problem focuses on finding the "most random" distribution under some conditions. For example, finding the maximum entropy among all distributions with mean λ, or all distributions with variance σ^2. (Incidentally, both of these constraints yield interesting solutions.)

We further motivate maximum entropy by noting the following from [16]:

1. Maximizing entropy will minimize the amount of "prior information" built into the probability distribution.
2. Physical systems tend to move towards maximum entropy as time progresses.

3.1 Maximum Entropy on an Interval

The first constraint we will examine is that of finite support. That is, lets find the distribution of maximum entropy for all distributions with support limited to the interval $[a, b]$.

Recall that in the discrete case, entropy is maximized when a set of events are equally likely, i.e., uniformly distributed. Intuitively, as the events are equiprobable, we can't make any educated guesses about which event might occur; thus, we learn a lot when we're told which event occurred.

In the continuous case, the result is much the same.

Claim. *The uniform distribution is the maximum entropy distribution on any interval* $[a, b]$.

Proof. From [14]: Suppose $f(x)$ is a distribution for $x \in (a, b)$ and $u(x)$ is the uniform distribution on that interval. Then:

$$\begin{aligned} D(f||u) &= \int f(x) \log \frac{f(x)}{u(x)} dx \\ &= \int f(x)(\log(f(x)) - \log(u(x)))dx \\ &= -h(x) - \int f(x) \log(u(x))dx \\ &= -h(x) + \log(b-a) \geq 0 \text{ by Theorem 2.3} \end{aligned}$$

Therefore, $\log(b-a) \geq h(x)$. That is, no distribution with finite support can have greater entropy than the uniform on the same interval. □

3.2 Maximum Entropy for Fixed Variance

Maximizing entropy over all distributions with fixed variance σ^2 is particularly interesting. Variance seems like the most natural quantity to vary when discussing entropy. Intuitively, if entropy is interpreted as a barometer for the

'randomness' of a distribution, then it would hopefully have some significant relationship to variance.

Recall (or see [13]) that the normal distribution with mean μ and variance σ^2 is defined as:

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

We will prove (from [5]) that the normal maximizes entropy.

Theorem 3.1. *The normal distribution maximizes entropy for all distributions with fixed variance σ^2 and mean μ.*

Proof. Again, we use relative entropy. Consider some distribution f and the normal distribution ϕ.

It is easily verified that the normal distribution ϕ with mean μ and variance σ^2 has entropy equal to $h(\phi) = \frac{1}{2}\log(2\pi e\sigma^2)$.

Combining this result with Lemma 2.4, we get:

$$\begin{aligned}
h(f) &\leq -\int f(x)\log(\phi(x))dx \\
&\leq -\int f(x)\log(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{(x-\mu)^2}{2\sigma^2}\})dx \\
&\leq -\int f(x)(\log(\exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}) + \log(\frac{1}{\sqrt{2\pi\sigma^2}}))dx \\
&\leq -\int f(x)(-\frac{(x-\mu)^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2))dx \\
&\leq \int f(x)\frac{(x-\mu)^2}{2\sigma^2}dx + \frac{1}{2}\log(2\pi\sigma^2)\int f(x)dx \\
&\leq \frac{1}{2\sigma^2}\int f(x)(x-\mu)^2dx + \frac{1}{2}\log(2\pi\sigma^2) \\
\text{As } &\int f(x)(x-\mu)^2dx \text{ is the variance of } f\text{:} \\
&\leq \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2} \\
&= \frac{1}{2}\log(2\pi e\sigma^2) \\
&= h(\phi)
\end{aligned}$$

Therefore, the entropy of f must be less than or equal to the entropy of the normal distribution with identical mean and variance. □

3.3 Maximum Entropy for Fixed Mean

As another example, consider the following problem in which we put a constraint on the mean of the distribution: *Find the continuous probability density function p of maximum entropy on $(0, \infty)$ with mean $\frac{1}{\lambda}$.*

Claim. *The exponential distribution with parameter λ maximizes entropy on $(0, \infty)$ for distributions with mean $\frac{1}{\lambda}$*

Proof. Consider the exponential distribution q with parameter λ (and, consequently, mean $\frac{1}{\lambda}$). It is easily verified that $h(q) = \log \frac{1}{\lambda} + 1$.

Let p be some other distribution on $(0, \infty)$ with mean $\frac{1}{\lambda}$. Then, by Lemma 2.4:

$$\begin{aligned}
h(p) &\leq -\int p(x) \log (q(x)) dx \\
&\leq -\int p(x) \log (\lambda e^{-\lambda x}) dx \\
&\leq -\int p(x) (\log \lambda + \log e^{-\lambda x}) dx \\
&\leq \int p(x) (\log \frac{1}{\lambda} - \log e^{-\lambda x}) dx \\
&\leq \log \frac{1}{\lambda} + \int p(x) \lambda x dx \\
&\leq \log \frac{1}{\lambda} + \lambda \int p(x) x dx \\
&\leq \log \frac{1}{\lambda} + \lambda \, \mathbb{E}[X] \\
&\leq \log \frac{1}{\lambda} + 1 \\
&= h(q)
\end{aligned}$$

□

4 The Central Limit Theorem

We start with an informal definition of the Central Limit Theorem, motivated by [7].

Definition. *The Central Limit Theorem (CLT) states that the distribution of the mean of a sample of i.i.d. random variables will approach normal in the limit. Specifically, if our variables X_i have mean μ and variance σ^2, the arithmetic mean will approach normal with parameters $(\mu, \sigma^2/n)$.*

The CLT has massive implications within statistics. Intuitively, it says that the distribution of the standardized sum of a bunch of X_is will be normal *regardless* of the shape of the X_is themselves. This allows us to make normality assumptions fairly often when handling real-world data.

In this paper, we prove the version of the CLT defined in [2].

Claim. *Let $X_1, X_2, ...$ be i.i.d. random variables with mean μ and variance σ^2. Further, let $S_n = \sqrt{n}[(\Sigma_{i=1}^{n} X_i)/n - \mu]$ be the standardized sum (that is, the convolution of the X_is divided by $\sqrt{n}$). We claim that the underlying distribution of S_n approaches normal with mean 0 and variance σ^2 as $n \to \infty$.*

For the rest of the proof, we assume that $\mu = 0$, and thus $S_n = \Sigma_{i=1}^{n} X_i/\sqrt{n}$, as μ is simply a shifting factor. If S_n is normal for $\mu = 0$, then it will be normal for any μ, as this factor just modifies the center of the distribution.

4.1 Overview

Typically, proofs of the CLT use inverse Fourier transforms or moment generating functions, as in [11]. In this paper, we'll use information-theoretic principles.

The broad outline of the proof will be to show that the relative entropy of S_n with respect to a normal distribution ϕ goes to zero.

To see that this is sufficient to prove the CLT, we use Pinkser's Inequality (from [10]).

Theorem 4.1 (Pinsker's Inequality)**.** *The variational distance between two probability mass functions P and Q, defined as:*

$$d(P, Q) = \Sigma_{x \in X} |P(x) - Q(x)|$$

is bounded above the relative entropy between the two distributions in the sense that

$$D(P||Q) \geq \tfrac{1}{2} d^2(P, Q)$$

Thus, if $\lim_{n \to \infty} D(S_n||\phi) = 0$, then the distance $d(P, Q)$ between the two distributions goes to 0. In other words, S_n approaches the normal.

(Note: from here onwards, we'll define $D(X) = D(f||\phi)$, where X has distribution f.)

To begin the proof, we provide a number of definitions and useful lemmas.

4.2 Fisher Information and its Connection to Entropy

Fisher Information is a useful quantity in the proof of the Central Limit Theorem. Intuitively, Fisher Information is the minimum error involved in estimating a parameter of a distribution. Alternatively, it can be seen as a measurement of how much information a random variable X carries about a parameter θ upon which it depends.

We provide the following definitions. While they will be necessary in our proofs, it is not imperative that you understand their significance.

Definition. *The standardized Fisher information of a random variable Y with density $g(y)$ and variance σ^2 is defined as*

$$J(Y) = \sigma^2 \, \mathbb{E}[\rho(Y) - \rho_\sigma(Y)]^2$$

where $\phi = g'/g$ is the score function for Y and $\rho_\sigma = \phi'/\phi$ is the score function for the normal with the same mean and variance as Y.[2]

Definition. *The Fisher information is defined in [2] as*

$$I(Y) = \mathbb{E}[\rho^2(Y)]$$

Alternatively, from [12]:

$$I(Y) = \int_{-\infty}^{\infty} (\frac{f'(y)}{f(y)})^2 f(y) dy$$

where the two quantities are related by $I = (J+1)/\sigma^2$.

4.3 Relationship Between Relative Entropy and Fisher Information

From [1], we can relate relative entropy to Fisher Information through the following lemma.

Lemma 4.2. *Let* X *be a random variable with finite variance. Then:*

$$\begin{aligned} D(X) &= \int_0^1 J(\sqrt{t}X + \sqrt{1-t}Z)\frac{dt}{2t},\ t \in (0,1) \\ &= \int_0^\infty J(X + \sqrt{\tau}Z)\frac{d\tau}{1+\tau},\ \tau = \frac{t}{1-t},\ \tau \in (0,1) \end{aligned}$$

This connection will be key in proving the Central Limit Theorem.

4.4 Convolution Inequalities

Again from [1] (drawing on [3] and [15]), we have the following result:

Lemma 4.3. *If* Y_1 *and* Y_2 *are random variables and* $\alpha_i \geq 0$, $\alpha_1 + \alpha_2 = 1$, *then* $I(\sqrt{\alpha_1}Y_1 + \sqrt{\alpha_2}Y_2) \leq \alpha_1 I(Y_1) + \alpha_2 I(Y_2)$.

Using this result, we can prove something stronger.

Lemma 4.4. *If* Y_1 *and* Y_2 *have the same variance, then*

$$\begin{gathered} J(\sqrt{\alpha_1}Y_1 + \sqrt{\alpha_2}Y_2) \leq \alpha_1 J(Y_1) + \alpha_2 J(Y_2) \\ \textit{and} \\ J(\Sigma_i \sqrt{\alpha_i} Y_i) \leq \Sigma_i \alpha_i J(Y_i) \end{gathered}$$

Proof. Recall that $I = (J+1)/\sigma^2$. Then:

$$\begin{aligned} I(\sqrt{\alpha_1}Y_1 + \sqrt{\alpha_2}Y_2) &\leq \alpha_1 I(Y_1) + \alpha_2 I(Y_2) \\ (J(\sqrt{\alpha_1}Y_1 + \sqrt{\alpha_2}Y_2) + 1)/\sigma^2 &\leq \alpha_1 (J(Y_1)+1)/\sigma^2 + \alpha_2 (J(Y_2)+1)/\sigma^2 \\ J(\sqrt{\alpha_1}Y_1 + \sqrt{\alpha_2}Y_2) + 1 &\leq \alpha_1 (J(Y_1)+1) + \alpha_2 (J(Y_2)+1) \\ J(\sqrt{\alpha_1}Y_1 + \sqrt{\alpha_2}Y_2) + 1 &\leq \alpha_1 J(Y_1) + \alpha_2 J(Y_2) + (\alpha_1 + \alpha_2) \\ J(\sqrt{\alpha_1}Y_1 + \sqrt{\alpha_2}Y_2) + 1 &\leq \alpha_1 J(Y_1) + \alpha_2 J(Y_2) + 1 \\ J(\sqrt{\alpha_1}Y_1 + \sqrt{\alpha_2}Y_2) &\leq \alpha_1 J(Y_1) + \alpha_2 J(Y_2) \end{aligned}$$

□

This argument can be extended to yield the stronger statement.

Next, we apply Lemma 4.4 to prove a number of helpful convolution inequalities.

Lemma 4.5. $D(\Sigma_i \sqrt{\alpha_i} X_i) \leq \Sigma_i \alpha_i D(X_i)$

Lemma 4.6. $H(\Sigma_i \sqrt{\alpha_i} X_i) \geq \Sigma_i \alpha_i H(X_i)$

Proof. From [1]. Let $Y_i = X_i + \sqrt{\tau} Z_i$, where Z_i is the normal with the same variance as X_i. By combining Lemma 4.5 with the equation:

$$D(X) = \int_0^\infty J(X + \sqrt{\tau} Z) \frac{d\tau}{1+\tau}$$

We get Lemma 4.5: $D(\Sigma_i \sqrt{\alpha_i} X_i) \leq \Sigma_i \alpha_i D(X_i)$. Noting that $H(X) = \frac{1}{2} \log(2\pi e \sigma^2) - D(X)$ gives us Lemma 4.6. □

We'll need a few more results before we can complete the proof of the CLT.

Lemma 4.7. $H(\frac{X_1 + \ldots + X_m}{\sqrt{m}}) \geq H(X_1)$ *if the* X_i *are i.i.d.*

Proof. Apply Lemma 4.6 with $\alpha_i = \frac{1}{m}$. □

Lemma 4.8. *For any integers* $n = mp$, $H(S_{mp}) \geq H(S_p)$.

Proof. Returning to the standardized sum, we note that $S_{mp} = \Sigma_{i=0}^m S_p / \sqrt{m}$. If we apply Lemma 4.6 with $X_i = S_p$ and $\alpha_i = 1/m$, we get:

$$H(S_{mp}) \geq H(S_p)$$ □

4.5 Subadditivity of Relative Entropy for Standardized Sum

The main theorem follows.

Theorem 4.9. *Let* S_n *be the standardized sum. Then* $nD(S_n)$ *is a subadditive sequence, and* $D(S_{2n}) \leq D(S_n)$. *As a result, we get convergence of the relative entropy:*

$$\lim_{n \to \infty} D(S_n) = 0$$

Proof. We divide our proof into several stages.

Subadditivity. Recall that $H(S_{mp}) \geq H(S_p)$. Setting $m = 2$ and $p = n$, we get $H(S_{2n}) \geq H(S_n)$, which implies that $D(S_{2n}) \leq D(S_n)$.

Limit is infimum. Next, we prove that the limit exists and equals the infimum. Let p be such that $H(S_p) \geq \sup_n(H(S_n)) - \epsilon$. Let $n = mp + r$ where $r < p$.

$$\begin{aligned}
H(S_{mp}) &= H(\Sigma_{i=1}^{m} S_p/\sqrt{m}) \\
H(S_n) &= H(S_{mp+r}) \\
&= H(\frac{\sqrt{mp}}{\sqrt{n}} S_{mp} + \frac{\sqrt{r}}{\sqrt{n}} S_r) \\
&\geq H(\frac{\sqrt{mp}}{\sqrt{n}} S_{mp}) \text{ as samples i.i.d., entropy increases on convolution} \\
&= H(S_{mp}) + \frac{1}{2}\log(mp/n) \\
&= H(S_{mp}) + \frac{1}{2}\log(mp/(mp+r)) \\
&= H(S_{mp}) + \frac{1}{2}\log(1-(r/n)) \\
&\geq H(S_p) + \frac{1}{2}\log(1-(r/n)) \text{ by Lemma 4.8}
\end{aligned}$$

This quantity converges to $H(S_p)$ as $n \to \infty$. As a result, we get that:

$$\begin{aligned}
\lim_{n\to\infty} H(S_n) &\geq H(S_p) + 0 \\
&\geq \sup_n(H(S_n)) - \epsilon
\end{aligned}$$

If we let $\epsilon \to 0$, we get that $\lim_{n\to\infty} H(S_n) = \sup_n(H(S_n))$. From the definition of relative entropy, we have $H(S_n) = \frac{1}{2}\log(2\pi e\sigma^2) - D(S_n)$. Thus, the previous statement is equivalent to $\lim_{n\to\infty} D(S_n) = \inf_n(D(S_n))$.

Infimum is 0. The skeleton of the proof in [2] is to show that the infimum is 0 for a subsequence of the n_k's. As the limit exists, all subsequences must converge to the limit of the sequence, and thus we can infer the limit of the entire sequence given a limit of one of the subsequences.

In particular, the subsequence is $n_k = 2^k n_0$, implying that the goal is to prove $\lim_{k\to\infty} D(S_{2^k n_0}) = 0$. This is done by showing that $\lim_{k\to\infty} J(S_{2^k n_0} + \sqrt{\tau}Z) = 0$, i.e., that J goes to zero for a subsequence of the n_k's (proven by going back to the definition of Fisher Information). Using the relationship between D and J demonstrated in Lemma 4.2, we get that $\lim_{k\to\infty} D(S_{2^k n_0}) = 0$.

As the limit exists, all subsequences must converge to the limit of the sequence, and thus the limit of the entire sequence is 0.

With that established, we've proven that $\lim_{n\to\infty} D(S_n) = 0$. □

The significance of Theorem 4.9 is that the distribution of the standardized sum deviates by less and less from the normal as n increases and, in the limit, does not deviate at all. Therefore, as we sample (i.e., as we increase n), the distribution of the standardized sum approaches the normal, proving the Central Limit Theorem.

5 Conclusion

Beginning with a definition for continuous entropy, we've shown that the quantity on its own holds little value due to its many shortcomings. While the definition was–on the surface–a seemingly minor notational deviation from the discrete case, continuous entropy lacks invariance under change of coordinates, non-negativity, and other desirable quantities that helped motivate the original definition for Shannon entropy.

But while continuous entropy on its own proved problematic, comparing entropy across continuous distributions (with relative entropy) yielded fascinating results, both through maximum entropy problems and, interestingly enough, the information-theoretic proof of the Central Limit Theorem, where the relative entropy of the standardized sum and the normal distribution was shown to drop to 0 as the sample size grew to infinity.

The applications of continuous information-theoretic techniques are varied; but, perhaps best of all, they allow us a means of justifying and proving results with the same familiar, intuitive feel granted us in the discrete realm. An information-theoretic proof of the Central Limit Theorem *makes sense* when we see that the relative entropy of the standardized sum and the normal decreases over time; similarly, the normal as the maximum entropy distribution for fixed mean and variance feels intuitive as well. Calling on information theory to prove and explain these results in the continuous case results in both rigorous justifications and intuitive explanations.

Appendix

Convolution Increases Entropy. From [9]: Recall that conditioning decreases entropy. Then, for independent X and Y, we have:

$$\begin{aligned} h(X+Y|X) = h(Y|X) = h(Y) \text{ by independence} \\ h(Y) = h(X+Y|X) \leq h(X+Y) \end{aligned}$$ □

References

[1] Andrew R. Barron. Monotonic Central Limit Theorem for Densities. Technical report, Stanford University, 1984.

[2] Andrew R. Barron. Entropy and the Central Limit Theorem. *The Annals of Probability*, 14:336–342, 1986.

[3] Nelson M. Blachman. The Convolution Inequality for Entropy Powes. *IEEE Interactions on Information Theory*, pages 267–271, April 1965.

[4] R.M. Castro. Maximum likelihood estimation and complexity regularization (lecture notes). May 2011.

[5] Keith Conrad. Probability Distributions and Maximum Entropy.

[6] Natasha Devroye. University of Illinois at Chicago ECE 534 notes on Differential Entropy. 2009.

[7] Justin Domke and Linwei Wang. The Central Limit Theorem (RIT lecture notes). 2012.

[8] E.T. Jaynes. Information theory and statistical mechanics. *Brandeis University Summer Institute Lectures in Theoretical Physics*, pages 182–218, 1963.

[9] O. Johnson and Y. Suhov. Cambridge University Information and Cdoing notes. 2006.

[10] Sanjeev Khudanpur. Johns Hopkins University ECE 520.674 notes. 1999.

[11] Hank Krieger. Proof of the Central Limit Theorem (Harvey Mudd College lecture notes). 2005.

[12] Miller Smith Puckette. *Shannon Entropy and the Central Limit Theorem.* PhD thesis, Harvard University, 1986.

[13] Raul Rojas. Why the Normal Distribution? (Freis Universitat Berlin lecture notes). 2010.

[14] Besma Smida. Harvard University ES250 notes on Differential Entropy. 2009.

[15] A.J. Stam. Some Inequalities Satisfied by the Quantities of Information of Fisher and Shannon. *Information and Control*, 2:101–112, 1959.

[16] Yao Xie. Duke university ECE587 notes on Differential Entropy.