



Institut
Mines-Télécom

On Minimum Entropy and Gaussian Transport

Entropy 2018: From Physics to Information
Sciences and Geometry

Barcelona, Spain, May 16th 2018

Olivier Rioul

<olivier.rioul@telecom-paristech.fr>



entropy

an Open Access Journal by MDPI





Outline

Introduction

What is Entropy?

Max/Min Entropy Principles

Equivalence to the Entropy Power Inequality

A Proof that Shannon Missed

Generalization to Linear Transformations

Shannon vs. Rényi

A Proof that Shannon Missed (Revisited)

Generalization to Rényi Entropies



Outline

Introduction

What is Entropy?

Max/Min Entropy Principles

Equivalence to the Entropy Power Inequality

A Proof that Shannon Missed

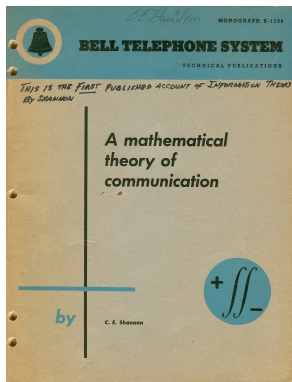
Generalization to Linear Transformations

Shannon vs. Rényi

A Proof that Shannon Missed (Revisited)

Generalization to Rényi Entropies

Shannon's 1948 seminal paper



The Bell System Technical Journal

Vol. XXVII

July, 1948

No. 3

A Mathematical Theory of Communication

By C. E. SHANNON

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist¹ and Hartley² on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.



Outline

Introduction

What is Entropy?

Max/Min Entropy Principles

Equivalence to the Entropy Power Inequality


A Proof that Shannon Missed

Generalization to Linear Transformations

Shannon vs. Rényi

A Proof that Shannon Missed (Revisited)

Generalization to Rényi Entropies



$$h(X) = \int f(x) \log \frac{1}{f(x)} dx \text{ where } X \sim f$$

“You should call it entropy [...] no one really knows what entropy really is, so in a debate you will always have the advantage.”



John von Neumann (1903–1957)




$$h(X) = \int f(x) \log \frac{1}{f(x)} dx \text{ where } X \sim f$$

"You should call it entropy [...] no one really knows what entropy really is, so in a debate you will always have the advantage."



John von Neumann (1903–1957)

"In the continuous case it is convenient to work not with the entropy H of an ensemble but with a derived quantity which we will call the entropy power." [Shannon'1948]



Entropy-Power

Definition ([Shannon48])

- *Power*: $P(X) = \mathbb{E}(X^2)$
- *Entropy-Power*: $N(X)$
same entropy

power of a *Gaussian* X^* having the

Entropy-Power

Definition ([Shannon48])

- *Power*: $P(X) = \mathbb{E}(X^2)$
- *Entropy-Power*: $N(X)$
same **entropy**

power of a *Gaussian* X^* having the

Entropy-Power

Definition ([Shannon48])

- *Power*: $P(X) = \mathbb{E}(X^2)$
 - *Entropy-Power*: $N(X) = P(X^*)$ power of a *Gaussian* X^* having the same entropy $h(X^*) = h(X)$
- Since $h(X^*) = \frac{1}{2} \log(2\pi e P(X^*))$:

$$N(X) = \frac{e^{2h(X)}}{2\pi e}$$

Entropy-Power

Definition ([Shannon48])

- *Power*: $P(X) = \mathbb{E}(X^2)$
- *Entropy-Power*: $N(X) = P(X^*)$ power of a *Gaussian* X^* having the same entropy $h(X^*) = h(X)$

- Since $h(X^*) = \frac{1}{2} \log(2\pi e P(X^*))$:

$$N(X) = \frac{e^{2h(X)}}{2\pi e}$$

“entropy power”

Entropy-Power

Definition ([Shannon48])

- *Power*: $P(X) = \mathbb{E}(X^2)$
- *Entropy-Power*: $N(X) = P(X^*)$ power of a *Gaussian* X^* having the same entropy $h(X^*) = h(X)$

- Since $h(X^*) = \frac{1}{2} \log(2\pi e P(X^*))$:

$$N(X) = \frac{e^{2h(X)}}{2\pi e}$$

“entropy power”

- *Scaling Property*:

$$P(aX) = a^2 P(X) \quad N(aX) = a^2 N(X)$$



Entropy-Power Inequality (EPI)

for any $X \perp\!\!\!\perp Y$:

$$P(X + Y) = P(X) + P(Y)$$

Entropy-Power Inequality (EPI)

for any $X \perp\!\!\!\perp Y$:

$$P(X + Y) = P(X) + P(Y)$$

Theorem (stated by Shannon, 1948)

$$N(X + Y) \geq N(X) + N(Y)$$

with equality iff X, Y are Gaussian.

The following result is derived in Appendix 6.

Theorem 15: Let the average power of two ensembles be N_1 and N_2 and let their entropy powers be \bar{N}_1 and \bar{N}_2 . Then the entropy power of the sum, \bar{N}_3 , is bounded by

$$\bar{N}_1 + \bar{N}_2 \leq \bar{N}_3 \leq N_1 + N_2.$$

White Gaussian noise has the peculiar property that it can

Entropy-Power Inequality (EPI)

for any $X \perp\!\!\!\perp Y$:

$$P(X + Y) = P(X) + P(Y)$$

Theorem (stated by Shannon, 1948)

$$e^{2h(X+Y)} \geq e^{2h(X)} + e^{2h(Y)}$$

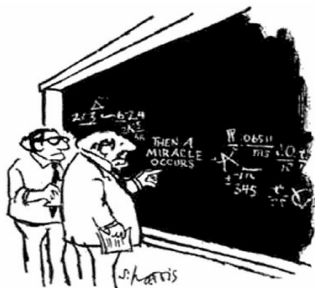
with equality iff X, Y are Gaussian.

The following result is derived in Appendix 6.

Theorem 15: Let the average power of two ensembles be N_1 and N_2 and let their entropy powers be \bar{N}_1 and \bar{N}_2 . Then the entropy power of the sum, \bar{N}_3 , is bounded by

$$\bar{N}_1 + \bar{N}_2 \leq \bar{N}_3 \leq N_1 + N_2.$$

White Gaussian noise has the peculiar property that it can



"I THINK YOU SHOULD BE MORE EXPLICIT
HERE IN STEP TWO."

© 1988 WILEY-INTERSCIENCE

Distributed by John Wiley & Sons, Inc.

Shannon's 1948 "proof" (Appendix 6).

A variational argument: $h(X + Y)$ for fixed $h(X)$ and $h(Y)$ has a stationary point when X, Y are Gaussian. □

This does not exclude local minima/maxima/saddle points.

The EPI has a Long History

- 1948 Stated and “proved” by Shannon in his seminal paper
- 1959 Stam’s proof using Fisher information
- 1965 Blachman’s exposition of Stam’s proof in IEEE Trans. IT
- 1978 Lieb’s proof using strengthened Young’s inequality
- 1991 Dembo-Cover-Thomas’ review of Stam’s & Lieb’s proofs
- 1991 Carlen-Soffer 1D variation of Stam’s proof
- 2000 Szarek-Voiculescu variant with Brunn-Minkowski inequality
- 2006 Guo-Shamai-Verdú proof based on the I-MMSE relation
- 2007 Rioul’s proof based on Mutual Information
- 2014 Wang-Madiman strengthening using Rényi entropies
- 2016 Courtade’s strengthening
- 2017 Yet another **simple** proof!





Applications of the EPI:

- nonGaussian Capacity [Shannon'48]
Gaussian means worst noise / Gaussian means best signal



Applications of the EPI:

- nonGaussian Capacity [Shannon'48]
Gaussian means worst noise / Gaussian means best signal
- multi-user capacity region outer bounds



Applications of the EPI:

- nonGaussian Capacity [Shannon'48]
Gaussian means worst noise / Gaussian means best signal
- multi-user capacity region outer bounds
- strengthening the Central Limit Theorem [Barron86]



Applications of the EPI:

- nonGaussian Capacity [Shannon'48]
Gaussian means worst noise / Gaussian means best signal
- multi-user capacity region outer bounds
- strengthening the Central Limit Theorem [Barron86]
- blind deconvolution / source separation [Donoho81]



Outline

Introduction

What is Entropy?

Max/Min Entropy Principles

Equivalence to the Entropy Power Inequality

A Proof that Shannon Missed

Generalization to Linear Transformations

Shannon vs. Rényi

A Proof that Shannon Missed (Revisited)

Generalization to Rényi Entropies





Ingredients

- random $X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \in \mathbb{R}^n$ with independent components X_i

Ingredients

- random $X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \in \mathbb{R}^n$ with independent components X_i
- a linear transformation: $X \mapsto \mathbf{A}X$

Ingredients

- random $X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \in \mathbb{R}^n$ with independent components X_i
- a linear transformation: $X \mapsto \mathbf{A}X$
- consider $h(\mathbf{A}X)$:

Ingredients

- random $X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \in \mathbb{R}^n$ with independent components X_i
- a linear transformation: $X \mapsto \mathbf{A}X$
- consider $h(\mathbf{A}X)$:
 - assume it is nondegenerate: $h(\mathbf{A}X) > -\infty$

Ingredients

- random $X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \in \mathbb{R}^n$ with independent components X_i
- a linear transformation: $X \mapsto \mathbf{A}X$
- consider $h(\mathbf{A}X)$:
 - assume it is nondegenerate: $h(\mathbf{A}X) > -\infty$
 - $\implies \mathbf{A}$ has full row rank

Ingredients

- random $X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \in \mathbb{R}^n$ with independent components X_i
- a linear transformation: $X \mapsto \mathbf{A}X$
- consider $h(\mathbf{A}X)$:
 - assume it is nondegenerate: $h(\mathbf{A}X) > -\infty$
 - $\implies \mathbf{A}$ has full row rank
 - \mathbf{A} is an $m \times n$ matrix with $m \leq n$

Ingredients

- random $X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \in \mathbb{R}^n$ with independent components X_i
- a linear transformation: $X \mapsto \mathbf{A}X$
- consider $h(\mathbf{A}X)$:
 - assume it is nondegenerate: $h(\mathbf{A}X) > -\infty$
 - $\implies \mathbf{A}$ has full row rank
 - \mathbf{A} is an $m \times n$ matrix with $m \leq n$
- $\max h(\mathbf{A}X)$ or $\min h(\mathbf{A}X)$?

Max/Min Entropy Principle

Let X^* be **Gaussian** with independent components X_i^* of same variances: $\text{Var}(X_i^*) = \text{Var}(X_i)$.

Theorem (Maximum Entropy Principle)

$$h(\mathbf{A}X) \leq h(\mathbf{A}X^*) \quad \text{with equality iff } X \text{ is Gaussian}$$

Max/Min Entropy Principle

Let X^* be **Gaussian** with independent components X_i^* of same variances: $\text{Var}(X_i^*) = \text{Var}(X_i)$.

Theorem (Maximum Entropy Principle)

$$h(\mathbf{AX}) \leq h(\mathbf{AX}^*) \quad \text{with equality iff } X \text{ is Gaussian}$$

■ Proof: $h(\mathbf{AX}^*) - h(\mathbf{AX}) = D_{\text{KL}}(\mathbf{AX} \parallel \mathbf{AX}^*) \geq 0 \quad \square$

Max/Min Entropy Principle

Let X^* be **Gaussian** with independent components X_i^* of same variances: $\text{Var}(X_i^*) = \text{Var}(X_i)$.

Theorem (Maximum Entropy Principle)

$$h(\mathbf{AX}) \leq h(\mathbf{AX}^*) \quad \text{with equality iff } X \text{ is Gaussian}$$

- Proof: $h(\mathbf{AX}^*) - h(\mathbf{AX}) = D_{\text{KL}}(\mathbf{AX} \parallel \mathbf{AX}^*) \geq 0 \quad \square$
- known in the 19th century (Gibbs' inequality)

Max/Min Entropy Principle

Let X^* be **Gaussian** with independent components X_i^* of same variances: $\text{Var}(X_i^*) = \text{Var}(X_i)$.

Theorem (Maximum Entropy Principle)

$$h(\mathbf{A}X) \leq h(\mathbf{A}X^*) \quad \text{with equality iff } X \text{ is Gaussian}$$

- Proof: $h(\mathbf{A}X^*) - h(\mathbf{A}X) = D_{\text{KL}}(\mathbf{A}X \parallel \mathbf{A}X^*) \geq 0 \quad \square$
- known in the 19th century (Gibbs' inequality)
- components need not be independent

Max/Min Entropy Principle

Let X^* be **Gaussian** with independent components X_i^* of same variances: $\text{Var}(X_i^*) = \text{Var}(X_i)$.

Theorem (Maximum Entropy Principle)

$$h(\mathbf{AX}) \leq h(\mathbf{AX}^*) \quad \text{with equality iff } X \text{ is Gaussian}$$

- Proof: $h(\mathbf{AX}^*) - h(\mathbf{AX}) = D_{\text{KL}}(\mathbf{AX} \parallel \mathbf{AX}^*) \geq 0 \quad \square$
- known in the 19th century (Gibbs' inequality)
- components need not be independent
- E. T. Jaynes, "Information theory and statistical mechanics," Physical Review, vol. 106, no. 4, pp. 620–630, 1957.



Max/Min Entropy Principle

Let X^* be **Gaussian** with independent components X_i^* of same variances: $\text{Var}(X_i^*) = \text{Var}(X_i)$.

Theorem (Maximum Entropy Principle)

$$h(\mathbf{AX}) \leq h(\mathbf{AX}^*) \quad \text{with equality iff } X \text{ is Gaussian}$$

- Proof: $h(\mathbf{AX}^*) - h(\mathbf{AX}) = D_{\text{KL}}(\mathbf{AX} \parallel \mathbf{AX}^*) \geq 0 \quad \square$
- known in the 19th century (Gibbs' inequality)
- components need not be independent
- E. T. Jaynes, "Information theory and statistical mechanics," Physical Review, vol. 106, no. 4, pp. 620–630, 1957.
- J. P. Burg, "Maximum entropy spectral analysis," Ph.D., Stanford, Dept. of Geophysics, Stanford, CA, USA, 1975.



Max/Min Entropy Principle

Let X^* be **Gaussian** with independent components X_i^* of same variances: $\text{Var}(X_i^*) = \text{Var}(X_i)$.

Theorem (Maximum Entropy Principle)

$$h(\mathbf{AX}) \leq h(\mathbf{AX}^*) \quad \text{with equality iff } X \text{ is Gaussian}$$

Max/Min Entropy Principle

Let X^* be **Gaussian** with independent components X_i^* of same variances: $\text{Var}(X_i^*) = \text{Var}(X_i)$.

Theorem (Maximum Entropy Principle)

$$h(\mathbf{AX}) \leq h(\mathbf{AX}^*) \quad \text{with equality iff } X \text{ is Gaussian}$$

Let X^* be **Gaussian** with independent components X_i^* of same entropies: $h(X_i^*) = h(X_i)$.

Theorem (Minimum Entropy Principle)

$$h(\mathbf{AX}) \geq h(\mathbf{AX}^*) \quad \text{with equality iff } X \text{ is Gaussian. . .}$$

Max/Min Entropy Principle

Let X^* be **Gaussian** with independent components X_i^* of same entropies: $h(X_i^*) = h(X_i)$.

Theorem (Minimum Entropy Principle)

$$h(\mathbf{AX}) \geq h(\mathbf{AX}^*) \quad \text{with equality iff } X \text{ is Gaussian.} \dots$$

Max/Min Entropy Principle

Let X^* be **Gaussian** with independent components X_i^* of same entropies: $h(X_i^*) = h(X_i)$.

Theorem (Minimum Entropy Principle)

$$h(\mathbf{A}X) \geq h(\mathbf{A}X^*) \quad \text{with equality iff } X \text{ is Gaussian. . .}$$

... or \mathbf{A} is *trivial*

- Closeness to normality by linear filtering

Max/Min Entropy Principle

Let X^* be **Gaussian** with independent components X_i^* of same entropies: $h(X_i^*) = h(X_i)$.

Theorem (Minimum Entropy Principle)

$$h(\mathbf{AX}) \geq h(\mathbf{AX}^*) \quad \text{with equality iff } X \text{ is Gaussian. . .}$$

... or \mathbf{A} is *trivial*

- Closeness to normality by linear filtering
- D. Donoho, "On minimum entropy deconvolution," in Applied Time Series Analysis II, Acad. Press, 1981, pp. 565–608.



Max/Min Entropy Principle

Let X^* be **Gaussian** with independent components X_i^* of same entropies: $h(X_i^*) = h(X_i)$.

Theorem (Minimum Entropy Principle)

$$h(\mathbf{AX}) \geq h(\mathbf{AX}^*) \quad \text{with equality iff } X \text{ is Gaussian. . .}$$

... or \mathbf{A} is *trivial*

- Closeness to normality by linear filtering
- D. Donoho, "On minimum entropy deconvolution," in Applied Time Series Analysis II, Acad. Press, 1981, pp. 565–608.
- R. Zamir & M. Feder, "A generalization of the entropy power inequality," IEEE Trans. IT, 39(5):1723–1728, Sep. 1993.



Max/Min Entropy Principle

Let X^* be **Gaussian** with independent components X_i^* of same entropies: $h(X_i^*) = h(X_i)$.

Theorem (Minimum Entropy Principle)

$$h(\mathbf{AX}) \geq h(\mathbf{AX}^*) \quad \text{with equality iff } X \text{ is Gaussian. . .}$$

... or \mathbf{A} is *trivial*

- Closeness to normality by linear filtering
- D. Donoho, "On minimum entropy deconvolution," in Applied Time Series Analysis II, Acad. Press, 1981, pp. 565–608.
- R. Zamir & M. Feder, "A generalization of the entropy power inequality," IEEE Trans. IT, 39(5):1723–1728, Sep. 1993.
- Application to deconvolution / blind separation



Max/Min Entropy Principle

Let X^* be **Gaussian** with independent components X_i^* of same entropies: $h(X_i^*) = h(X_i)$.

Theorem (Minimum Entropy Principle)

$$h(\mathbf{AX}) \geq h(\mathbf{AX}^*) \quad \text{with equality iff } X \text{ is Gaussian. . .}$$

... or \mathbf{A} is *trivial*

- Closeness to normality by linear filtering
- D. Donoho, "On minimum entropy deconvolution," in Applied Time Series Analysis II, Acad. Press, 1981, pp. 565–608.
- R. Zamir & M. Feder, "A generalization of the entropy power inequality," IEEE Trans. IT, 39(5):1723–1728, Sep. 1993.
- Application to deconvolution / blind separation
- Proof: involved!





Outline

Introduction

What is Entropy?

Max/Min Entropy Principles

Equivalence to the Entropy Power Inequality

A Proof that Shannon Missed

Generalization to Linear Transformations

Shannon vs. Rényi

A Proof that Shannon Missed (Revisited)

Generalization to Rényi Entropies

Simplest nontrivial case: $(m, n) = (1, 2)$

Take $\mathbf{A} = \begin{pmatrix} a & b \end{pmatrix}$ with nonzero a, b (nontrivial mixture).

Theorem (MinEnt for $(m, n) = (1, 2)$)

For any two independent X, Y , letting X^*, Y^* independent Gaussian s.t. $h(X^*) = h(X)$, $h(Y) = h(Y^*)$,

$h(aX + bY) \geq h(aX^* + bY^*)$ with equality iff X, Y are Gaussian.

Simplest nontrivial case: $(m, n) = (1, 2)$

Take $\mathbf{A} = (a \ b)$ with nonzero a, b (nontrivial mixture).

Theorem (MinEnt for $(m, n) = (1, 2)$)

For any two independent X, Y , letting X^*, Y^* independent Gaussian s.t. $h(X^*) = h(X), h(Y) = h(Y^*)$,

$$h(aX + bY) \geq h(aX^* + bY^*) \text{ with equality iff } X, Y \text{ are Gaussian.}$$

Definition (Entropy Power [Shannon'48])

Entropy Power = Power of a Gaussian noise with the same entropy:

$$N(X) = \text{Var}(X^*) \quad \text{where} \quad h(X^*) = h(X)$$

i.e., since $h(X^*) = \frac{1}{2} \log(2\pi e \text{Var}(X^*))$,

$$N(X) = \exp(2h(X))/2\pi e$$

Simplest nontrivial case: $(m, n) = (1, 2)$

Take $\mathbf{A} = (a \ b)$ with nonzero a, b (nontrivial mixture).

Theorem (MinEnt for $(m, n) = (1, 2)$)

For any two independent X, Y , letting X^*, Y^* independent Gaussian s.t. $h(X^*) = h(X)$, $h(Y) = h(Y^*)$,

$N(aX + bY) \geq N(aX^* + bY^*)$ with equality iff X, Y are Gaussian.

Definition (Entropy Power [Shannon'48])

Entropy Power = Power of a Gaussian noise with the same entropy:

$$N(X) = \text{Var}(X^*) \quad \text{where} \quad h(X^*) = h(X)$$

i.e., since $h(X^*) = \frac{1}{2} \log(2\pi e \text{Var}(X^*))$,

$$N(X) = \exp(2h(X)) / 2\pi e$$

Simplest nontrivial case: $(m, n) = (1, 2)$

Take $\mathbf{A} = (a \ b)$ with nonzero a, b (nontrivial mixture).

Theorem (MinEnt for $(m, n) = (1, 2)$)

For any two independent X, Y , letting X^*, Y^* independent Gaussian s.t. $h(X^*) = h(X)$, $h(Y) = h(Y^*)$,

$N(aX + bY) \geq N(aX^* + bY^*)$ with equality iff X, Y are Gaussian.

Definition (Entropy Power [Shannon'48])

Entropy Power = Power of a Gaussian noise with the same entropy:

$$N(X) = \text{Var}(X^*) \quad \text{where} \quad h(X^*) = h(X)$$

i.e., since $h(X^*) = \frac{1}{2} \log(2\pi e \text{Var}(X^*))$,

$$N(X) = \exp(2h(X))/2\pi e$$

$$N(X^*) = \text{Var}(X^*)$$

Simplest nontrivial case: $(m, n) = (1, 2)$

Take $\mathbf{A} = (a \ b)$ with nonzero a, b (nontrivial mixture).

Theorem (MinEnt for $(m, n) = (1, 2)$)

For any two independent X, Y , letting X^*, Y^* independent Gaussian s.t. $h(X^*) = h(X)$, $h(Y) = h(Y^*)$,

$$N(aX + bY) \geq N(aX^*) + N(bY^*) \text{ with equality iff } X, Y \text{ are Gaussian.}$$

Definition (Entropy Power [Shannon'48])

Entropy Power = Power of a Gaussian noise with the same entropy:

$$N(X) = \text{Var}(X^*) \quad \text{where} \quad h(X^*) = h(X)$$

i.e., since $h(X^*) = \frac{1}{2} \log(2\pi e \text{Var}(X^*))$,

$$N(X) = \exp(2h(X)) / 2\pi e$$

$$N(X^*) = \text{Var}(X^*)$$

Simplest nontrivial case: $(m, n) = (1, 2)$

Take $\mathbf{A} = (a \ b)$ with nonzero a, b (nontrivial mixture).

Theorem (MinEnt for $(m, n) = (1, 2)$)

For any two independent X, Y , letting X^*, Y^* independent Gaussian s.t. $h(X^*) = h(X)$, $h(Y) = h(Y^*)$,

$$N(aX + bY) \geq N(aX^*) + N(bY^*) \quad \text{with equality iff } X, Y \text{ are Gaussian.}$$

Definition (Entropy Power [Shannon'48])

Entropy Power = Power of a Gaussian noise with the same entropy:

$$N(X) = \text{Var}(X^*) \quad \text{where} \quad N(X^*) = N(X)$$

i.e., since $h(X^*) = \frac{1}{2} \log(2\pi e \text{Var}(X^*))$,

$$N(X) = \exp(2h(X))/2\pi e$$

$$N(X^*) = \text{Var}(X^*)$$

Simplest nontrivial case: $(m, n) = (1, 2)$

Take $\mathbf{A} = (a \ b)$ with nonzero a, b (nontrivial mixture).

Theorem (MinEnt for $(m, n) = (1, 2)$)

For any two independent X, Y , letting X^*, Y^* independent Gaussian s.t. $h(X^*) = h(X)$, $h(Y) = h(Y^*)$,

$N(aX + bY) \geq N(aX) + N(bY)$ with equality iff X, Y are Gaussian.

Definition (Entropy Power [Shannon'48])

Entropy Power = Power of a Gaussian noise with the same entropy:

$$N(X) = \text{Var}(X^*) \quad \text{where} \quad N(X^*) = N(X)$$

i.e., since $h(X^*) = \frac{1}{2} \log(2\pi e \text{Var}(X^*))$,

$$N(X) = \exp(2h(X))/2\pi e$$

$$N(X^*) = \text{Var}(X^*)$$

Simplest nontrivial case: $(m, n) = (1, 2)$

Take $\mathbf{A} = (a \ b)$ with nonzero a, b (nontrivial mixture).

Theorem (MinEnt for $(m, n) = (1, 2)$)

For any two independent X, Y ,

$$N(X + Y) \geq N(X) + N(Y) \text{ with equality iff } X, Y \text{ are Gaussian.}$$

Definition (Entropy Power [Shannon'48])

Entropy Power = Power of a Gaussian noise with the same entropy:

$$N(X) = \text{Var}(X^*) \quad \text{where} \quad N(X^*) = N(X)$$

i.e., since $h(X^*) = \frac{1}{2} \log(2\pi e \text{Var}(X^*))$,

$$N(X) = \exp(2h(X)) / 2\pi e$$

$$N(X^*) = \text{Var}(X^*)$$

Simplest nontrivial case: $(m, n) = (1, 2)$

Take $\mathbf{A} = \begin{pmatrix} a & b \end{pmatrix}$ with nonzero a, b (nontrivial mixture).

Theorem (**Entropy-Power Inequality** [Shannon'48])

For any two independent X, Y ,

$$N(X + Y) \geq N(X) + N(Y) \text{ with equality iff } X, Y \text{ are Gaussian.}$$

Definition (Entropy Power [Shannon'48])

Entropy Power = Power of a Gaussian noise with the same entropy:

$$N(X) = \text{Var}(X^*) \quad \text{where} \quad N(X^*) = N(X)$$

i.e., since $h(X^*) = \frac{1}{2} \log(2\pi e \text{Var}(X^*))$,

$$N(X) = \exp(2h(X))/2\pi e \quad N(X^*) = \text{Var}(X^*)$$



Outline

Introduction

What is Entropy?

Max/Min Entropy Principles

Equivalence to the Entropy Power Inequality

A Proof that Shannon Missed

Generalization to Linear Transformations

Shannon vs. Rényi

A Proof that Shannon Missed (Revisited)

Generalization to Rényi Entropies



Ingredients: “Optimal Transport”

Lemma (inverse function sampling method)

If U is uniform in $[0, 1]$ and X has c.d.f. $F(x) = \mathbb{P}(X \leq x)$, then $F^{-1}(U)$ has the same distribution as X .

Proof.

$$\mathbb{P}(F^{-1}(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x). \quad \square$$

Ingredients: “Optimal Transport”

Lemma (inverse function sampling method)

If U is uniform in $[0, 1]$ and X has c.d.f. $F(x) = \mathbb{P}(X \leq x)$, then $F^{-1}(U)$ has the same distribution as X .

Corollary (monotonic increasing transport $T = F^{-1} \circ G$)

Let F, G be two c.d.f.'s. Then $X^* \sim G \implies X = T(X^*) \sim F$.

Proof.

$U = G(X^*) \sim \text{uniform}; \quad T(X^*) = F^{-1}(G(X^*)) = F^{-1}(U) \sim F.$ □

Ingredients: “Optimal Transport”

Lemma (inverse function sampling method)

If U is uniform in $[0, 1]$ and X has c.d.f. $F(x) = \mathbb{P}(X \leq x)$, then $F^{-1}(U)$ has the same distribution as X .

Corollary (monotonic increasing transport $T = F^{-1} \circ G$)

Let F, G be two c.d.f.'s. Then $X^* \sim G \implies X = T(X^*) \sim F$.

Proof.

$U = G(X^*) \sim \text{uniform}$; $T(X^*) = F^{-1}(G(X^*)) = F^{-1}(U) \sim F$. □

- nD generalization: Knöthe map, Brenier map...

Ingredients: “Optimal Transport”

Lemma (inverse function sampling method)

If U is uniform in $[0, 1]$ and X has c.d.f. $F(x) = \mathbb{P}(X \leq x)$, then $F^{-1}(U)$ has the same distribution as X .

Corollary (monotonic increasing transport $T = F^{-1} \circ G$)

Let F, G be two c.d.f.'s. Then $X^* \sim G \implies X = T(X^*) \sim F$.

Proof.

$U = G(X^*) \sim \text{uniform}$; $T(X^*) = F^{-1}(G(X^*)) = F^{-1}(U) \sim F$. □

- n D generalization: Knöthe map, Brenier map...
- Used in **optimal transport** theory

Ingredients: “Optimal Transport”

Lemma (Change of variable [Shannon'48])

For any continuous X, X^* , monotonic increasing transport $T(X^*) \sim X$,

$$h(X) = \boxed{h(T(X^*)) = h(X^*) + \mathbb{E} \log T'(X^*)}$$

Ingredients: “Optimal Transport”

Lemma (Change of variable [Shannon'48])

For any continuous X, X^* , monotonic increasing transport $T(X^*) \sim X$,

$$h(X) = \boxed{h(T(X^*)) = h(X^*) + \mathbb{E} \log T'(X^*)}$$

Proof.

Proof: make the change of variable $x = T(x^*)$ in

$$h(X) = \int f_X(x) \log \frac{1}{f_X(x)} dx = \int \underbrace{f_X(T(x^*)) T'(x^*)}_{f_{X^*}(x^*)} \log \frac{1}{f_X(T(x^*))} dx^*$$



Ingredients: “Optimal Transport”

Lemma (Change of variable [Shannon'48])

For any continuous X, X^* , monotonic increasing transport $T(X^*) \sim X$,

$$h(X) = \boxed{h(T(X^*)) = h(X^*) + \mathbb{E} \log T'(X^*)}$$

Proof.

Proof: make the change of variable $x = T(x^*)$ in

$$h(X) = \int f_X(x) \log \frac{1}{f_X(x)} dx = \int \underbrace{f_X(T(x^*)) T'(x^*)}_{f_{X^*}(x^*)} \log \frac{1}{f_X(T(x^*))} dx^*$$

■ in particular $h(aX) = h(X) + \log |a| \iff N(aX) = a^2 N(X)$; □

Ingredients: “Optimal Transport”

Lemma (Change of variable [Shannon'48])

For any continuous X, X^* , monotonic increasing transport $T(X^*) \sim X$,

$$h(X) = \boxed{h(T(X^*)) = h(X^*) + \mathbb{E} \log T'(X^*)}$$

Proof.

Proof: make the change of variable $x = T(x^*)$ in

$$h(X) = \int f_X(x) \log \frac{1}{f_X(x)} dx = \int \underbrace{f_X(T(x^*)) T'(x^*)}_{f_{X^*}(x^*)} \log \frac{1}{f_X(T(x^*))} dx^*$$

- in particular $h(aX) = h(X) + \log |a| \iff N(aX) = a^2 N(X)$; □
- more generally in nD : $h(T(X^*)) = h(X^*) + \mathbb{E} \log |\det T'(X^*)|$

A Proof that Shannon Missed

Proceed to prove the inequality $h(aX + bY) \geq h(aX^* + bY^*)$
where X^*, Y^* are indep. Gaussian s.t. $h(X^*) = h(X), h(Y) = h(Y^*)$

A Proof that Shannon Missed

Proceed to prove the inequality $h(aX + bY) \geq h(aX^* + bY^*)$
where X^*, Y^* are indep. Gaussian s.t. $h(X^*) = h(X) = h(Y) = h(Y^*)$

1. We may assume $h(X) = h(Y)$.

A Proof that Shannon Missed

Proceed to prove the inequality $h(aX + bY) \geq h(aX^* + bY^*)$
where X^*, Y^* are indep. Gaussian s.t. $h(X^*) = h(X) = h(Y) = h(Y^*)$

1. We may assume $h(X) = h(Y)$. Otherwise:
 - set $c = e^{-h(X)}$ and $d = e^{-h(Y)}$ so that $h(cX) = h(dY)$;
 - apply the above to cX and dY .

A Proof that Shannon Missed

Proceed to prove the inequality $h(aX + bY) \geq h(aX^* + bY^*)$
where X^*, Y^* are indep. Gaussian s.t. $h(X^*) = h(X) = h(Y) = h(Y^*)$

1. We may assume $h(X) = h(Y)$. Otherwise:
 - set $c = e^{-h(X)}$ and $d = e^{-h(Y)}$ so that $h(cX) = h(dY)$;
 - apply the above to cX and dY .

So w.l.o.g. X^*, Y^* are i.i.d. Gaussian.

A Proof that Shannon Missed

Proceed to prove the inequality $h(aX + bY) \geq h(aX^* + bY^*)$
where X^*, Y^* are indep. Gaussian s.t. $h(X^*) = h(X) = h(Y) = h(Y^*)$

1. We may assume $h(X) = h(Y)$. Otherwise:
 - set $c = e^{-h(X)}$ and $d = e^{-h(Y)}$ so that $h(cX) = h(dY)$;
 - apply the above to cX and dY .

So w.l.o.g. X^*, Y^* are i.i.d. Gaussian.

2. We may always normalize: $a^2 + b^2 = 1$.

A Proof that Shannon Missed

Proceed to prove the inequality $h(aX + bY) \geq h(aX^* + bY^*)$
where X^*, Y^* are indep. Gaussian s.t. $h(X^*) = h(X) = h(Y) = h(Y^*)$

1. We may assume $h(X) = h(Y)$. Otherwise:
 - set $c = e^{-h(X)}$ and $d = e^{-h(Y)}$ so that $h(cX) = h(dY)$;
 - apply the above to cX and dY .

So w.l.o.g. X^*, Y^* are i.i.d. Gaussian.

2. We may always normalize: $a^2 + b^2 = 1$. Otherwise:
 - divide a, b by $\Delta = \sqrt{a^2 + b^2}$;
 - the log Δ terms cancel.

A Proof that Shannon Missed

Proceed to prove the inequality $h(aX + bY) \geq h(aX^* + bY^*)$ where X^*, Y^* are indep. Gaussian s.t. $h(X^*) = h(X) = h(Y) = h(Y^*)$

1. We may assume $h(X) = h(Y)$. Otherwise:
 - set $c = e^{-h(X)}$ and $d = e^{-h(Y)}$ so that $h(cX) = h(dY)$;
 - apply the above to cX and dY .

So w.l.o.g. X^*, Y^* are i.i.d. Gaussian.

2. We may always normalize: $a^2 + b^2 = 1$. Otherwise:
 - divide a, b by $\Delta = \sqrt{a^2 + b^2}$;
 - the log Δ terms cancel.
3. Make the changes of variables $X = T(X^*), Y = U(Y^*)$:

A Proof that Shannon Missed

Proceed to prove the inequality $h(aX + bY) \geq h(aX^* + bY^*)$ where X^*, Y^* are indep. Gaussian s.t. $h(X^*) = h(X) = h(Y) = h(Y^*)$

1. We may assume $h(X) = h(Y)$. Otherwise:
 - set $c = e^{-h(X)}$ and $d = e^{-h(Y)}$ so that $h(cX) = h(dY)$;
 - apply the above to cX and dY .

So w.l.o.g. X^*, Y^* are i.i.d. Gaussian.

2. We may always normalize: $a^2 + b^2 = 1$. Otherwise:
 - divide a, b by $\Delta = \sqrt{a^2 + b^2}$;
 - the log Δ terms cancel.
3. Make the changes of variables $X = T(X^*), Y = U(Y^*)$:

One is led to prove $h(aT(X^*) + bU(Y^*)) \geq h(aX^* + bY^*)$

A Proof that Shannon Missed

Proceed to prove the inequality $h(aX + bY) \geq h(aX^* + bY^*)$ where X^*, Y^* are indep. Gaussian s.t. $h(X^*) = h(X) = h(Y) = h(Y^*)$

1. We may assume $h(X) = h(Y)$. Otherwise:
 - set $c = e^{-h(X)}$ and $d = e^{-h(Y)}$ so that $h(cX) = h(dY)$;
 - apply the above to cX and dY .

So w.l.o.g. X^*, Y^* are i.i.d. Gaussian.

2. We may always normalize: $a^2 + b^2 = 1$. Otherwise:
 - divide a, b by $\Delta = \sqrt{a^2 + b^2}$;
 - the log Δ terms cancel.

3. Make the changes of variables $X = T(X^*), Y = U(Y^*)$:

One is led to prove $h(aT(X^*) + bU(Y^*)) \geq h(aX^* + bY^*)$

4. Define $\tilde{X} = aX^* + bY^*$.

A Proof that Shannon Missed

Proceed to prove the inequality $h(aX + bY) \geq h(aX^* + bY^*)$
where X^*, Y^* are indep. Gaussian s.t. $h(X^*) = h(X) = h(Y) = h(Y^*)$

1. We may assume $h(X) = h(Y)$. Otherwise:
 - set $c = e^{-h(X)}$ and $d = e^{-h(Y)}$ so that $h(cX) = h(dY)$;
 - apply the above to cX and dY .

So w.l.o.g. X^*, Y^* are i.i.d. Gaussian.

2. We may always normalize: $a^2 + b^2 = 1$. Otherwise:
 - divide a, b by $\Delta = \sqrt{a^2 + b^2}$;
 - the log Δ terms cancel.

3. Make the changes of variables $X = T(X^*), Y = U(Y^*)$:

One is led to prove $h(aT(X^*) + bU(Y^*)) \geq h(aX^* + bY^*)$

4. Define $\tilde{X} = aX^* + bY^*$. Complete the rotation: $\tilde{Y} = -bX^* + aY^*$
so that \tilde{X}, \tilde{Y} are i.i.d. Gaussian

A Proof that Shannon Missed

Proceed to prove the inequality $h(aX + bY) \geq h(aX^* + bY^*)$
where X^*, Y^* are indep. Gaussian s.t. $h(X^*) = h(X) = h(Y) = h(Y^*)$

1. We may assume $h(X) = h(Y)$. Otherwise:
 - set $c = e^{-h(X)}$ and $d = e^{-h(Y)}$ so that $h(cX) = h(dY)$;
 - apply the above to cX and dY .

So w.l.o.g. X^*, Y^* are i.i.d. Gaussian.

2. We may always normalize: $a^2 + b^2 = 1$. Otherwise:
 - divide a, b by $\Delta = \sqrt{a^2 + b^2}$;
 - the log Δ terms cancel.

3. Make the changes of variables $X = T(X^*), Y = U(Y^*)$:

One is led to prove $h(aT(X^*) + bU(Y^*)) \geq h(aX^* + bY^*)$

4. Define $\tilde{X} = aX^* + bY^*$. Complete the rotation: $\tilde{Y} = -bX^* + aY^*$
so that \tilde{X}, \tilde{Y} are i.i.d. Gaussian and $X^* = a\tilde{X} - b\tilde{Y}$, $Y^* = b\tilde{X} + a\tilde{Y}$

A Proof that Shannon Missed

One is led to prove $h(aT(X^*) + bU(Y^*)) \geq h(aX^* + bY^*)$

\tilde{X}, \tilde{Y} are i.i.d. Gaussian and $X^* = a\tilde{X} - b\tilde{Y}$, $Y^* = b\tilde{X} + a\tilde{Y}$.

A Proof that Shannon Missed

One is led to prove $h(aT(X^*) + bU(Y^*)) \geq h(aX^* + bY^*)$

\tilde{X}, \tilde{Y} are i.i.d. Gaussian and $X^* = a\tilde{X} - b\tilde{Y}$, $Y^* = b\tilde{X} + a\tilde{Y}$.

5. Since conditioning reduces entropy:

$$\begin{aligned} h(aT(X^*) + bU(Y^*)) &= h(aT(a\tilde{X} - b\tilde{Y}) + bU(b\tilde{X} + a\tilde{Y})) \\ &\geq h(aT(a\tilde{X} - b\tilde{Y}) + bU(b\tilde{X} + a\tilde{Y}) | \tilde{Y}) \end{aligned}$$

A Proof that Shannon Missed

One is led to prove $h(aT(X^*) + bU(Y^*)) \geq h(aX^* + bY^*)$

\tilde{X}, \tilde{Y} are i.i.d. Gaussian and $X^* = a\tilde{X} - b\tilde{Y}$, $Y^* = b\tilde{X} + a\tilde{Y}$.

5. Since conditioning reduces entropy:

$$\begin{aligned} h(aT(X^*) + bU(Y^*)) &= h(aT(a\tilde{X} - b\tilde{Y}) + bU(b\tilde{X} + a\tilde{Y})) \\ &\geq h(\underbrace{aT(a\tilde{X} - b\tilde{Y}) + bU(b\tilde{X} + a\tilde{Y})}_{T_{\tilde{Y}}(\tilde{X})} | \tilde{Y}) \end{aligned}$$

A Proof that Shannon Missed

One is led to prove $h(aT(X^*) + bU(Y^*)) \geq h(aX^* + bY^*)$

\tilde{X}, \tilde{Y} are i.i.d. Gaussian and $X^* = a\tilde{X} - b\tilde{Y}$, $Y^* = b\tilde{X} + a\tilde{Y}$.

5. Since conditioning reduces entropy:

$$\begin{aligned} h(aT(X^*) + bU(Y^*)) &= h(aT(a\tilde{X} - b\tilde{Y}) + bU(b\tilde{X} + a\tilde{Y})) \\ &\geq h(\underbrace{aT(a\tilde{X} - b\tilde{Y}) + bU(b\tilde{X} + a\tilde{Y})}_{T_{\tilde{Y}}(\tilde{X})} | \tilde{Y}) \end{aligned}$$

6. By the change of variable:

$$= h(\tilde{X} | \tilde{Y}) + \mathbb{E} \log T'_{\tilde{Y}}(\tilde{X})$$

A Proof that Shannon Missed

One is led to prove $h(aT(X^*) + bU(Y^*)) \geq h(aX^* + bY^*)$

\tilde{X}, \tilde{Y} are i.i.d. Gaussian and $X^* = a\tilde{X} - b\tilde{Y}$, $Y^* = b\tilde{X} + a\tilde{Y}$.

5. Since conditioning reduces entropy:

$$\begin{aligned} h(aT(X^*) + bU(Y^*)) &= h(aT(a\tilde{X} - b\tilde{Y}) + bU(b\tilde{X} + a\tilde{Y})) \\ &\geq h(\underbrace{aT(a\tilde{X} - b\tilde{Y}) + bU(b\tilde{X} + a\tilde{Y})}_{T_{\tilde{Y}}(\tilde{X})} | \tilde{Y}) \end{aligned}$$

6. By the change of variable:

$$= h(\tilde{X}) + \mathbb{E} \log T'_{\tilde{Y}}(\tilde{X})$$

A Proof that Shannon Missed

One is led to prove $h(aT(X^*) + bU(Y^*)) \geq h(aX^* + bY^*)$

\tilde{X}, \tilde{Y} are i.i.d. Gaussian and $X^* = a\tilde{X} - b\tilde{Y}$, $Y^* = b\tilde{X} + a\tilde{Y}$.

5. Since conditioning reduces entropy:

$$\begin{aligned} h(aT(X^*) + bU(Y^*)) &= h(aT(a\tilde{X} - b\tilde{Y}) + bU(b\tilde{X} + a\tilde{Y})) \\ &\geq h(\underbrace{aT(a\tilde{X} - b\tilde{Y}) + bU(b\tilde{X} + a\tilde{Y})}_{T_{\tilde{Y}}(\tilde{X})} | \tilde{Y}) \end{aligned}$$

6. By the change of variable:

$$\begin{aligned} &T_{\tilde{Y}}(\tilde{X}) \\ &= h(\tilde{X}) + \mathbb{E} \log T'_{\tilde{Y}}(\tilde{X}) \\ &= h(\tilde{X}) + \mathbb{E} \log (a^2 T'(a\tilde{X} - b\tilde{Y}) + b^2 U'(b\tilde{X} + a\tilde{Y})) \end{aligned}$$

A Proof that Shannon Missed

One is led to prove $h(aT(X^*) + bU(Y^*)) \geq h(aX^* + bY^*)$

\tilde{X}, \tilde{Y} are i.i.d. Gaussian and $X^* = a\tilde{X} - b\tilde{Y}$, $Y^* = b\tilde{X} + a\tilde{Y}$.

5. Since conditioning reduces entropy:

$$\begin{aligned} h(aT(X^*) + bU(Y^*)) &= h(aT(a\tilde{X} - b\tilde{Y}) + bU(b\tilde{X} + a\tilde{Y})) \\ &\geq h(\underbrace{aT(a\tilde{X} - b\tilde{Y}) + bU(b\tilde{X} + a\tilde{Y})}_{T_{\tilde{Y}}(\tilde{X})} | \tilde{Y}) \end{aligned}$$

6. By the change of variable:

$$\begin{aligned} &T_{\tilde{Y}}(\tilde{X}) \\ &= h(\tilde{X}) + \mathbb{E} \log T'_{\tilde{Y}}(\tilde{X}) \\ &= h(\tilde{X}) + \mathbb{E} \log (a^2 T'(a\tilde{X} - b\tilde{Y}) + b^2 U'(b\tilde{X} + a\tilde{Y})) \\ &= h(aX^* + bY^*) + \mathbb{E} \log (a^2 T'(X^*) + b^2 U'(Y^*)) \end{aligned}$$

A Proof that Shannon Missed

One is led to prove $h(aT(X^*) + bU(Y^*)) \geq h(aX^* + bY^*)$

\tilde{X}, \tilde{Y} are i.i.d. Gaussian and $X^* = a\tilde{X} - b\tilde{Y}$, $Y^* = b\tilde{X} + a\tilde{Y}$.

5. Since conditioning reduces entropy:

$$\begin{aligned} h(aT(X^*) + bU(Y^*)) &= h(aT(a\tilde{X} - b\tilde{Y}) + bU(b\tilde{X} + a\tilde{Y})) \\ &\geq h(\underbrace{aT(a\tilde{X} - b\tilde{Y}) + bU(b\tilde{X} + a\tilde{Y})}_{T_{\tilde{Y}}(\tilde{X})} | \tilde{Y}) \end{aligned}$$

6. By the change of variable:

$$\begin{aligned} &= h(\tilde{X}) + \mathbb{E} \log T'_{\tilde{Y}}(\tilde{X}) \\ &= h(\tilde{X}) + \mathbb{E} \log (a^2 T'(a\tilde{X} - b\tilde{Y}) + b^2 U'(b\tilde{X} + a\tilde{Y})) \\ &= h(aX^* + bY^*) + \mathbb{E} \log (a^2 T'(X^*) + b^2 U'(Y^*)) \end{aligned}$$

7. By concavity of the log:

$$\geq h(aX^* + bY^*) + a^2 \mathbb{E} \log T'(X^*) + b^2 \mathbb{E} \log U'(Y^*)$$

A Proof that Shannon Missed

One is led to prove $h(aT(X^*) + bU(Y^*)) \geq h(aX^* + bY^*)$

\tilde{X}, \tilde{Y} are i.i.d. Gaussian and $X^* = a\tilde{X} - b\tilde{Y}$, $Y^* = b\tilde{X} + a\tilde{Y}$.

5. Since conditioning reduces entropy:

$$\begin{aligned} h(aT(X^*) + bU(Y^*)) &= h(aT(a\tilde{X} - b\tilde{Y}) + bU(b\tilde{X} + a\tilde{Y})) \\ &\geq h(\underbrace{aT(a\tilde{X} - b\tilde{Y}) + bU(b\tilde{X} + a\tilde{Y})}_{T_{\tilde{Y}}(\tilde{X})} | \tilde{Y}) \end{aligned}$$

6. By the change of variable:

$$\begin{aligned} &T_{\tilde{Y}}(\tilde{X}) \\ &= h(\tilde{X}) + \mathbb{E} \log T'_{\tilde{Y}}(\tilde{X}) \\ &= h(\tilde{X}) + \mathbb{E} \log (a^2 T'(a\tilde{X} - b\tilde{Y}) + b^2 U'(b\tilde{X} + a\tilde{Y})) \\ &= h(aX^* + bY^*) + \mathbb{E} \log (a^2 T'(X^*) + b^2 U'(Y^*)) \end{aligned}$$

7. By concavity of the log:

$$\begin{aligned} &\geq h(aX^* + bY^*) + a^2 \underbrace{\mathbb{E} \log T'(X^*)}_{h(X) - h(X^*) = 0} + b^2 \underbrace{\mathbb{E} \log U'(Y^*)}_{h(Y) - h(Y^*) = 0} \end{aligned}$$

A Proof that Shannon Missed

One is led to prove $h(aT(X^*) + bU(Y^*)) \geq h(aX^* + bY^*)$

\tilde{X}, \tilde{Y} are i.i.d. Gaussian and $X^* = a\tilde{X} - b\tilde{Y}$, $Y^* = b\tilde{X} + a\tilde{Y}$.

5. Since conditioning reduces entropy:

$$\begin{aligned} h(aT(X^*) + bU(Y^*)) &= h(aT(a\tilde{X} - b\tilde{Y}) + bU(b\tilde{X} + a\tilde{Y})) \\ &\geq h(\underbrace{aT(a\tilde{X} - b\tilde{Y}) + bU(b\tilde{X} + a\tilde{Y})}_{T_{\tilde{Y}}(\tilde{X})} | \tilde{Y}) \end{aligned}$$

6. By the change of variable:

$$\begin{aligned} &= h(\tilde{X}) + \mathbb{E} \log T'_{\tilde{Y}}(\tilde{X}) \\ &= h(\tilde{X}) + \mathbb{E} \log (a^2 T'(a\tilde{X} - b\tilde{Y}) + b^2 U'(b\tilde{X} + a\tilde{Y})) \\ &= h(aX^* + bY^*) + \mathbb{E} \log (a^2 T'(X^*) + b^2 U'(Y^*)) \end{aligned}$$

7. By concavity of the log:

$$\begin{aligned} &\geq h(aX^* + bY^*) + a^2 \mathbb{E} \log T'(X^*) + b^2 \mathbb{E} \log U'(Y^*) \\ &\geq h(aX^* + bY^*) \quad \square \end{aligned}$$

Equality Case

For nonzero a, b :

- in log concavity inequality:

$$\mathbb{E} \log(a^2 T'(X^*) + b^2 U'(Y^*)) = a^2 \mathbb{E} \log T'(X^*) + b^2 \mathbb{E} \log U'(Y^*)$$

$$\implies T'(X^*) = U'(X^*) = c > 0 \text{ constant a.e.}$$

Equality Case

For nonzero a, b :

- in log concavity inequality:

$$\mathbb{E} \log(a^2 T'(X^*) + b^2 U'(Y^*)) = a^2 \mathbb{E} \log T'(X^*) + b^2 \mathbb{E} \log U'(Y^*)$$

$$\implies T'(X^*) = U'(X^*) = c > 0 \text{ constant a.e.}$$

$$\implies T, U \text{ are linear: } X = T(X^*) = cX^*, Y = U(Y^*) = cY^* \text{ Gaussian.}$$

Equality Case

For nonzero a, b :

- in log concavity inequality:

$$\mathbb{E} \log(a^2 T'(X^*) + b^2 U'(Y^*)) = a^2 \mathbb{E} \log T'(X^*) + b^2 \mathbb{E} \log U'(Y^*)$$

$\implies T'(X^*) = U'(X^*) = c > 0$ constant a.e.

$\implies T, U$ are linear: $X = T(X^*) = cX^*, Y = U(Y^*) = cY^*$ Gaussian.

$\implies c = 1$ since $h(X) = h(X^*), h(Y) = h(Y^*)$.

Equality Case

For nonzero a, b :

- in log concavity inequality:

$$\mathbb{E} \log(a^2 T'(X^*) + b^2 U'(Y^*)) = a^2 \mathbb{E} \log T'(X^*) + b^2 \mathbb{E} \log U'(Y^*)$$

$$\implies T'(X^*) = U'(X^*) = c > 0 \text{ constant a.e.}$$

$$\implies T, U \text{ are linear: } X = T(X^*) = cX^*, Y = U(Y^*) = cY^* \text{ Gaussian.}$$

$$\implies c = 1 \text{ since } h(X) = h(X^*), h(Y) = h(Y^*).$$

- in information inequality:

$$h(aT(a\tilde{X} - b\tilde{Y}) + bU(b\tilde{X} + a\tilde{Y})) = h(aT(a\tilde{X} - b\tilde{Y}) + bU(b\tilde{X} + a\tilde{Y}) | \tilde{Y})$$

comes for free since $a(a\tilde{X} - b\tilde{Y}) + b(b\tilde{X} + a\tilde{Y}) = \tilde{X}$ is indep of \tilde{Y} .



Outline

Introduction

What is Entropy?

Max/Min Entropy Principles

Equivalence to the Entropy Power Inequality

A Proof that Shannon Missed

Generalization to Linear Transformations

Shannon vs. Rényi

A Proof that Shannon Missed (Revisited)

Generalization to Rényi Entropies





Generalization to Linear Transformations

Proceed to prove $h(\mathbf{AX}) \geq h(\mathbf{AX}^*)$.

Generalization to Linear Transformations

Proceed to prove $h(\mathbf{AX}) \geq h(\mathbf{AX}^*)$.

- We may assume all X_i have the same entropy: Otherwise, introduce $c_i = e^{-h(X_i)}$ and apply the result to the $c_i X_i$.

Generalization to Linear Transformations

Proceed to prove $h(\mathbf{AX}) \geq h(\mathbf{AX}^*)$.

- We may assume all X_i have the same entropy: Otherwise, introduce $c_i = e^{-h(X_i)}$ and apply the result to the $c_i X_i$.
- Since $h(X_i^*) = h(X_i)$, all X_i^* have the same variance, hence are i.i.d.

Generalization to Linear Transformations

Proceed to prove $h(\mathbf{AX}) \geq h(\mathbf{AX}^*)$.

- We may assume all X_i have the same entropy: Otherwise, introduce $c_i = e^{-h(X_i)}$ and apply the result to the $c_i X_i$.
- Since $h(X_i^*) = h(X_i)$, all X_i^* have the same variance, hence are i.i.d.
- We may assume that \mathbf{A} has rank $= m \leq n$ (otherwise the result is trivial): $h(\mathbf{AX}) = h(\mathbf{AX}^*) = -\infty$.

Generalization to Linear Transformations

Proceed to prove $h(\mathbf{AX}) \geq h(\mathbf{AX}^*)$.

- We may assume all X_i have the same entropy: Otherwise, introduce $c_i = e^{-h(X_i)}$ and apply the result to the $c_i X_i$.
- Since $h(X_i^*) = h(X_i)$, all X_i^* have the same variance, hence are i.i.d.
- We may assume that \mathbf{A} has rank $= m \leq n$ (otherwise the result is trivial): $h(\mathbf{AX}) = h(\mathbf{AX}^*) = -\infty$.
- The difference $h(\mathbf{AX}) - h(\mathbf{AX}^*)$ is invariant by elementary row operations. By the Gram-Schmidt procedure, we may assume that the rows of \mathbf{A} are orthonormal: $\mathbf{AA}^t = \mathbf{I}$.

Generalization to Linear Transformations

Proceed to prove $h(\mathbf{AX}) \geq h(\mathbf{AX}^*)$.

- We may assume all X_i have the same entropy: Otherwise, introduce $c_i = e^{-h(X_i)}$ and apply the result to the $c_i X_i$.
- Since $h(X_i^*) = h(X_i)$, all X_i^* have the same variance, hence are i.i.d.
- We may assume that \mathbf{A} has rank $= m \leq n$ (otherwise the result is trivial): $h(\mathbf{AX}) = h(\mathbf{AX}^*) = -\infty$.
- The difference $h(\mathbf{AX}) - h(\mathbf{AX}^*)$ is invariant by elementary row operations. By the Gram-Schmidt procedure, we may assume that the rows of \mathbf{A} are orthonormal: $\mathbf{AA}^t = \mathbf{I}$.
- Extend \mathbf{A} to an orthogonal matrix $\mathbf{A}' = \begin{pmatrix} \mathbf{A} \\ \mathbf{A}^c \end{pmatrix}$

Generalization to Linear Transformations

- Then let $\tilde{X} = \mathbf{A}X^*$ et $\tilde{X}^c = \mathbf{A}^c X^*$ so that $\tilde{X}' = \begin{pmatrix} \tilde{X} \\ \tilde{X}^c \end{pmatrix} = \mathbf{A}'X^*$ has i.i.d. components. Inverting yields $X^* = \mathbf{A}'^t \tilde{X}'$.

Generalization to Linear Transformations

- Then let $\tilde{X} = \mathbf{A}X^*$ et $\tilde{X}^c = \mathbf{A}^c X^*$ so that $\tilde{X}' = \begin{pmatrix} \tilde{X} \\ \tilde{X}^c \end{pmatrix} = \mathbf{A}'X^*$ has i.i.d. components. Inverting yields $X^* = \mathbf{A}'^t \tilde{X}'$.
- By the changes of variables $X_i = T_i(X_i^*)$, since conditioning reduces entropy:

$$\begin{aligned} h(\mathbf{A}X) &= h(\mathbf{A}\mathbf{T}(X^*)) \\ &= h(\mathbf{A}\mathbf{T}(\mathbf{A}'^t \tilde{X}')) \\ &\geq h(\mathbf{A}\mathbf{T}(\mathbf{A}'^t \tilde{X}') | \tilde{X}^c) \end{aligned}$$

Generalization to Linear Transformations

- Then let $\tilde{X} = \mathbf{A}X^*$ et $\tilde{X}^c = \mathbf{A}^c X^*$ so that $\tilde{X}' = \begin{pmatrix} \tilde{X} \\ \tilde{X}^c \end{pmatrix} = \mathbf{A}'X^*$ has i.i.d. components. Inverting yields $X^* = \mathbf{A}'^t \tilde{X}'$.
- By the changes of variables $X_i = T_i(X_i^*)$, since conditioning reduces entropy:

$$\begin{aligned} h(\mathbf{A}X) &= h(\mathbf{A}\mathbf{T}(X^*)) \\ &= h(\mathbf{A}\mathbf{T}(\mathbf{A}'^t \tilde{X}')) \\ &\geq h(\mathbf{A}\mathbf{T}(\mathbf{A}'^t \tilde{X}') | \tilde{X}^c) \end{aligned}$$

- But the Jacobian matrix of

$\mathbf{T}_{\tilde{X}^c}(\tilde{X}) = \mathbf{A}\mathbf{T}(\mathbf{A}'^t \tilde{X}') = \mathbf{A}\mathbf{T}(\mathbf{A}'^t \tilde{X} + (\mathbf{A}^c)^t \tilde{X}^c)$ for fixed \tilde{X}^c is

$\mathbf{T}'_{\tilde{X}^c}(\tilde{X}) = \mathbf{A}\mathbf{T}'(\mathbf{A}'^t \tilde{X}')\mathbf{A}'^t = \mathbf{A}\mathbf{T}'(X^*)\mathbf{A}'^t$ where $\mathbf{T}'(X^*) = \text{diag}(T'_i(X_i^*))$

Generalization to Linear Transformations

- The change of variables in the entropy yields

$$\begin{aligned}h(\mathbf{A}X) &\geq h(\mathbf{A} \mathbf{T}(\mathbf{A}'^t \tilde{X}') | \tilde{X}^c) \\ &= h(\tilde{X} | \tilde{X}^c) + \mathbb{E} \log \det(\mathbf{A} \mathbf{T}'(X^*) \mathbf{A}'^t)\end{aligned}$$

Generalization to Linear Transformations

- The change of variables in the entropy yields

$$\begin{aligned}h(\mathbf{A}X) &\geq h(\mathbf{A} \mathbf{T}(\mathbf{A}'^t \tilde{X}') | \tilde{X}^c) \\ &= h(\tilde{X} | \tilde{X}^c) + \mathbb{E} \log \det(\mathbf{A} \mathbf{T}'(X^*) \mathbf{A}^t)\end{aligned}$$

- By the concavity of the logarithm:

$$\log \det(\mathbf{A} \mathbf{T}'(X^*) \mathbf{A}^t) \geq \text{tr}(\mathbf{A} \cdot \log \mathbf{T}'(X^*) \cdot \mathbf{A}^t)$$

thus

$$h(\mathbf{A}X) \geq h(\tilde{X} | \tilde{X}^c) + \text{tr}(\mathbf{A} \cdot \mathbb{E} \log \mathbf{T}'(\tilde{X}) \cdot \mathbf{A}^t)$$

Generalization to Linear Transformations

- The change of variables in the entropy yields

$$\begin{aligned}h(\mathbf{A}\mathbf{X}) &\geq h(\mathbf{A}\mathbf{T}(\mathbf{A}^t \tilde{\mathbf{X}}') | \tilde{\mathbf{X}}^c) \\ &= h(\tilde{\mathbf{X}} | \tilde{\mathbf{X}}^c) + \mathbb{E} \log \det(\mathbf{A}\mathbf{T}'(\mathbf{X}^*)\mathbf{A}^t)\end{aligned}$$

- By the concavity of the logarithm:

$$\log \det(\mathbf{A}\mathbf{T}'(\mathbf{X}^*)\mathbf{A}^t) \geq \text{tr}(\mathbf{A} \cdot \log \mathbf{T}'(\mathbf{X}^*) \cdot \mathbf{A}^t)$$

thus

$$h(\mathbf{A}\mathbf{X}) \geq h(\tilde{\mathbf{X}} | \tilde{\mathbf{X}}^c) + \text{tr}(\mathbf{A} \cdot \mathbb{E} \log \mathbf{T}'(\tilde{\mathbf{X}}) \cdot \mathbf{A}^t)$$

- But $h(\tilde{\mathbf{X}} | \tilde{\mathbf{X}}^c) = h(\tilde{\mathbf{X}}) = h(\mathbf{A}\mathbf{X}^*)$ and
 $\mathbb{E} \log T'_i(\tilde{X}_i) = h(T_i(\tilde{X}_i)) - h(\tilde{X}_i) = h(X_i) - h(\tilde{X}_i) = 0$; so

$$h(\mathbf{A}\mathbf{X}) \geq h(\mathbf{A}\mathbf{X}^*)$$



Generalization to Linear Transformations

- The change of variables in the entropy yields

$$\begin{aligned}h(\mathbf{A}X) &\geq h(\mathbf{A} \mathbf{T}(\mathbf{A}^t \tilde{X}') | \tilde{X}^c) \\ &= h(\tilde{X} | \tilde{X}^c) + \mathbb{E} \log \det(\mathbf{A} \mathbf{T}'(X^*) \mathbf{A}^t)\end{aligned}$$

- By the concavity of the logarithm:

$$\log \det(\mathbf{A} \mathbf{T}'(X^*) \mathbf{A}^t) \geq \text{tr}(\mathbf{A} \cdot \log \mathbf{T}'(X^*) \cdot \mathbf{A}^t)$$

thus

$$h(\mathbf{A}X) \geq h(\tilde{X} | \tilde{X}^c) + \text{tr}(\mathbf{A} \cdot \mathbb{E} \log \mathbf{T}'(\tilde{X}) \cdot \mathbf{A}^t)$$

- But $h(\tilde{X} | \tilde{X}^c) = h(\tilde{X}) = h(\mathbf{A}X^*)$ and
 $\mathbb{E} \log T'_i(\tilde{X}_i) = h(T_i(\tilde{X}_i)) - h(\tilde{X}_i) = h(X_i) - h(\tilde{X}_i) = 0$; so

$$h(\mathbf{A}X) \geq h(\mathbf{A}X^*)$$

□

- Equality iff either \mathbf{A} is trivial or $T'_i(X_i) = \text{Cst.}$, hence X is Gaussian



Outline

Introduction

What is Entropy?

Max/Min Entropy Principles

Equivalence to the Entropy Power Inequality

A Proof that Shannon Missed

Generalization to Linear Transformations

Shannon vs. Rényi

A Proof that Shannon Missed (Revisited)

Generalization to Rényi Entropies



Shannon's and Rényi's Entropies

$$h(X) = \int f \log \frac{1}{f}$$

Shannon's and Rényi's Entropies

$$h(X) = \int f \log \frac{1}{f} = h_1(X)$$

$$h_r(X) = \frac{1}{1-r} \log \int f^r$$

"A mathematician is a device for turning coffee into theorems"



Alfred Rényi (1921–1970)



Lieb's Restatement of the EPI

Theorem (Lieb, 1978)

$$e^{2h(X+Y)} \geq e^{2h(X)} + e^{2h(Y)}$$

\iff for any $0 < \lambda < 1$

$$h(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) \geq \lambda h(X) + (1-\lambda)h(Y)$$

Proof.

\implies : $X = \sqrt{\lambda}X'$, $Y = \sqrt{1-\lambda}Y'$, take the log (concavity of the log)

\impliedby $X = X'/\sqrt{\lambda}$, $Y = Y'/\sqrt{1-\lambda}$, take the exp, assuming λ such that $h(X) = h(Y)$, r.h.s. is $(e^{2h(X)})^\lambda (e^{2h(Y)})^{1-\lambda} = \lambda e^{2h(X)} + (1-\lambda)e^{2h(Y)}$. \square

Lieb's Restatement of the EPI

Theorem (A generalization:)

$$e^{2h_r(X+Y)} \geq c \cdot (e^{2h_r(X)} + e^{2h_r(Y)})$$

\iff for any $0 < \lambda < 1$

$$h_r(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) \geq \lambda h_r(X) + (1-\lambda)h_r(Y) + \log \sqrt{c}$$

Same proof!

\implies : $X = \sqrt{\lambda}X'$, $Y = \sqrt{1-\lambda}Y'$, take the log (concavity of the log)

\impliedby $X = X'/\sqrt{\lambda}$, $Y = Y'/\sqrt{1-\lambda}$, take the exp, assuming λ such that $h_r(X) = h_r(Y)$, r.h.s. $(e^{2h_r(X)})^\lambda (e^{2h_r(Y)})^{1-\lambda} = \lambda e^{2h_r(X)} + (1-\lambda)e^{2h_r(Y)} \square$

Lieb's Restatement of the EPI

Theorem (A generalization:)

$$e^{2h_r(X+Y)} \geq c \cdot (e^{2h_r(X)} + e^{2h_r(Y)})$$

\iff for any $0 < \lambda < 1$

$$h_r(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) \geq \lambda h_r(X) + (1-\lambda)h_r(Y) + n \log \sqrt{c}$$

Same proof! (with c independent of the dimension).

\implies : $X = \sqrt{\lambda}X'$, $Y = \sqrt{1-\lambda}Y'$, take the log (concavity of the log)

\impliedby $X = X'/\sqrt{\lambda}$, $Y = Y'/\sqrt{1-\lambda}$, take the exp, assuming λ such that $h_r(X) = h_r(Y)$, r.h.s. $(e^{2h_r(X)})^\lambda (e^{2h_r(Y)})^{1-\lambda} = \lambda e^{2h_r(X)} + (1-\lambda)e^{2h_r(Y)} \square$

Restatement for More than Two Variables

N independent variables X_1, X_2, \dots, X_N .

Theorem

$$e^{2h_r(\sum_i X_i)} \geq c \cdot \sum_i e^{2h_r(X_i)}$$

\iff for any convex combination ($\sum_i \lambda_i = 1$)

$$h_r\left(\sum_i \sqrt{\lambda_i} X_i\right) \geq \sum_i \lambda_i h_r(X_i) + \frac{n}{2} \log c$$

Same proof.

Variation for More than Two Variables

N independent variables X_1, X_2, \dots, X_N .

Theorem

$$e^{2\alpha h_r(\sum_i X_i)} \geq \sum_i e^{2\alpha h_r(X_i)}$$

\Leftrightarrow for any convex combination ($\sum_i \lambda_i = 1$)

$$h_r\left(\sum_i \sqrt{\lambda_i} X_i\right) \geq \sum_i \lambda_i h_r(X_i) + \frac{n}{2} (1/\alpha - 1) H(\lambda)$$

Same proof.



Outline

Introduction

What is Entropy?

Max/Min Entropy Principles

Equivalence to the Entropy Power Inequality

A Proof that Shannon Missed

Generalization to Linear Transformations

Shannon vs. Rényi

A Proof that Shannon Missed (Revisited)

Generalization to Rényi Entropies



The Proof that Shannon Missed (Again)

Take any $X \perp\!\!\!\perp Y$ and X^*, Y^* i.i.d. Gaussian. Set $X = T(X^*)$ and $Y = U(Y^*)$. Then

$$\begin{aligned} & h(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) - \lambda h(X) - (1-\lambda)h(Y) \\ &= h(\sqrt{\lambda}T(X^*) + \sqrt{1-\lambda}U(Y^*)) - \lambda h(T(X^*)) - (1-\lambda)h(U(Y^*)) \end{aligned}$$

The Proof that Shannon Missed (Again)

Take any $X \perp\!\!\!\perp Y$ and X^*, Y^* i.i.d. Gaussian. Set $X = T(X^*)$ and $Y = U(Y^*)$. Then

$$\begin{aligned} & h(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) - \lambda h(X) - (1-\lambda)h(Y) \\ &= h(\sqrt{\lambda}T(X^*) + \sqrt{1-\lambda}U(Y^*)) - \lambda \underbrace{h(T(X^*))}_{h(X^*) + \mathbb{E} \log T'(X^*)} - (1-\lambda) \underbrace{h(U(Y^*))}_{h(Y^*) + \mathbb{E} \log U'(Y^*)} \end{aligned}$$

Compare this to

$$h(\underbrace{\sqrt{\lambda}X^* + \sqrt{1-\lambda}Y^*}_{\tilde{X}}) - \lambda h(X^*) + (1-\lambda)h(Y^*)$$

The Proof that Shannon Missed (Again)

Take any $X \perp Y$ and X^*, Y^* i.i.d. Gaussian. Set $X = T(X^*)$ and $Y = U(Y^*)$. Then

$$\begin{aligned} & h(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) - \lambda h(X) - (1-\lambda)h(Y) \\ &= h(\sqrt{\lambda}T(X^*) + \sqrt{1-\lambda}U(Y^*)) - \lambda \underbrace{h(T(X^*))}_{h(X^*) + \mathbb{E} \log T'(X^*)} - (1-\lambda) \underbrace{h(U(Y^*))}_{h(Y^*) + \mathbb{E} \log U'(Y^*)} \end{aligned}$$

Compare this to

$$\begin{aligned} & h(\underbrace{\sqrt{\lambda}X^* + \sqrt{1-\lambda}Y^*}_{\tilde{X}}) - \lambda h(X^*) + (1-\lambda)h(Y^*) \\ & \begin{cases} \tilde{X} = \sqrt{\lambda}X^* + \sqrt{1-\lambda}Y^* \\ \tilde{Y} = -\sqrt{1-\lambda}X^* + \sqrt{\lambda}Y^* \end{cases} \quad \begin{cases} X^* = \sqrt{\lambda}\tilde{X} - \sqrt{1-\lambda}\tilde{Y} \\ Y^* = \sqrt{1-\lambda}\tilde{X} + \sqrt{\lambda}\tilde{Y} \end{cases} \end{aligned}$$

The Proof that Shannon Missed (Again)

Take any $X \perp\!\!\!\perp Y$ and X^*, Y^* i.i.d. Gaussian. Set $X = T(X^*)$ and $Y = U(Y^*)$. Then

$$\begin{aligned} & h(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) - \lambda h(X) - (1-\lambda)h(Y) \\ &= h(\sqrt{\lambda}T(X^*) + \sqrt{1-\lambda}U(Y^*)) - \lambda \underbrace{h(T(X^*))}_{h(X^*) + \mathbb{E} \log T'(X^*)} - (1-\lambda) \underbrace{h(U(Y^*))}_{h(Y^*) + \mathbb{E} \log U'(Y^*)} \end{aligned}$$

Compare this to

$$\begin{aligned} & h(\sqrt{\lambda}X^* + \sqrt{1-\lambda}Y^*) - \lambda h(X^*) + (1-\lambda)h(Y^*) \\ & \begin{cases} \tilde{X} = \sqrt{\lambda}X^* + \sqrt{1-\lambda}Y^* \\ \tilde{Y} = -\sqrt{1-\lambda}X^* + \sqrt{\lambda}Y^* \end{cases} \quad \begin{cases} X^* = \sqrt{\lambda}\tilde{X} - \sqrt{1-\lambda}\tilde{Y} \\ Y^* = \sqrt{1-\lambda}\tilde{X} + \sqrt{\lambda}\tilde{Y} \end{cases} \end{aligned}$$

Then $\sqrt{\lambda}T(X^*) + \sqrt{1-\lambda}U(Y^*)$ becomes a function of $\tilde{X}, \tilde{Y} \dots$

A Proof that Shannon Missed (Cont'd)

$$\begin{aligned}h(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) &= h(\sqrt{\lambda}T(X^*) + \sqrt{1-\lambda}U(Y^*)) \\&= h(\sqrt{\lambda}T(\sqrt{\lambda}\tilde{X} - \sqrt{1-\lambda}\tilde{Y}) + \sqrt{1-\lambda}U(\sqrt{1-\lambda}\tilde{X} + \sqrt{\lambda}\tilde{Y})) \\&\geq h(\sqrt{\lambda}T(\sqrt{\lambda}\tilde{X} - \sqrt{1-\lambda}\tilde{Y}) + \sqrt{1-\lambda}U(\sqrt{1-\lambda}\tilde{X} + \sqrt{\lambda}\tilde{Y})|\tilde{Y}) \\&= h(\tilde{X}|\tilde{Y}) + \mathbb{E} \log(\lambda T'(\sqrt{\lambda}\tilde{X} - \sqrt{1-\lambda}\tilde{Y}) + (1-\lambda)U'(\sqrt{1-\lambda}\tilde{X} + \sqrt{\lambda}\tilde{Y})) \\&= h(\tilde{X}) + \mathbb{E} \log(\lambda T'(X^*) + (1-\lambda)U'(Y^*)) \\&\geq h(\sqrt{\lambda}X^* + \sqrt{1-\lambda}Y^*) + \lambda \mathbb{E} \log T'(X^*) + (1-\lambda) \mathbb{E} \log U'(Y^*)\end{aligned}$$

A Proof that Shannon Missed (Cont'd)

$$\begin{aligned}h(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) &= h(\sqrt{\lambda}T(X^*) + \sqrt{1-\lambda}U(Y^*)) \\&= h(\sqrt{\lambda}T(\sqrt{\lambda}\tilde{X} - \sqrt{1-\lambda}\tilde{Y}) + \sqrt{1-\lambda}U(\sqrt{1-\lambda}\tilde{X} + \sqrt{\lambda}\tilde{Y})) \\&\geq h(\sqrt{\lambda}T(\sqrt{\lambda}\tilde{X} - \sqrt{1-\lambda}\tilde{Y}) + \sqrt{1-\lambda}U(\sqrt{1-\lambda}\tilde{X} + \sqrt{\lambda}\tilde{Y})|\tilde{Y}) \\&= h(\tilde{X}|\tilde{Y}) + \mathbb{E} \log(\lambda T'(\sqrt{\lambda}\tilde{X} - \sqrt{1-\lambda}\tilde{Y}) + (1-\lambda)U'(\sqrt{1-\lambda}\tilde{X} + \sqrt{\lambda}\tilde{Y})) \\&= h(\tilde{X}) + \mathbb{E} \log(\lambda T'(X^*) + (1-\lambda)U'(Y^*)) \\&\geq h(\sqrt{\lambda}X^* + \sqrt{1-\lambda}Y^*) + \lambda \mathbb{E} \log T'(X^*) + (1-\lambda) \mathbb{E} \log U'(Y^*)\end{aligned}$$

Then subtract

$$\lambda h(X) + (1-\lambda)h(Y) = \lambda h(X^*) + (1-\lambda)h(Y^*) + \lambda \mathbb{E} \log T'(X^*) + (1-\lambda) \mathbb{E} \log U'(Y^*):$$

$$\begin{aligned}h(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) - \lambda h(X) - (1-\lambda)h(Y) \\&\geq h(\sqrt{\lambda}X^* + \sqrt{1-\lambda}Y^*) - \lambda h(X^*) + (1-\lambda)h(Y^*) = 0\end{aligned}$$

A Proof that Shannon Missed (Cont'd)

$$\begin{aligned}h(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) &= h(\sqrt{\lambda}T(X^*) + \sqrt{1-\lambda}U(Y^*)) \\&= h(\sqrt{\lambda}T(\sqrt{\lambda}\tilde{X} - \sqrt{1-\lambda}\tilde{Y}) + \sqrt{1-\lambda}U(\sqrt{1-\lambda}\tilde{X} + \sqrt{\lambda}\tilde{Y})) \\&\geq h(\sqrt{\lambda}T(\sqrt{\lambda}\tilde{X} - \sqrt{1-\lambda}\tilde{Y}) + \sqrt{1-\lambda}U(\sqrt{1-\lambda}\tilde{X} + \sqrt{\lambda}\tilde{Y})|\tilde{Y}) \\&= h(\tilde{X}|\tilde{Y}) + \mathbb{E} \log(\lambda T'(\sqrt{\lambda}\tilde{X} - \sqrt{1-\lambda}\tilde{Y}) + (1-\lambda)U'(\sqrt{1-\lambda}\tilde{X} + \sqrt{\lambda}\tilde{Y})) \\&= h(\tilde{X}) + \mathbb{E} \log(\lambda T'(X^*) + (1-\lambda)U'(Y^*)) \\&\geq h(\sqrt{\lambda}X^* + \sqrt{1-\lambda}Y^*) + \lambda \mathbb{E} \log T'(X^*) + (1-\lambda) \mathbb{E} \log U'(Y^*)\end{aligned}$$

Then subtract

$$\lambda h(X) + (1-\lambda)h(Y) = \lambda h(X^*) + (1-\lambda)h(Y^*) + \lambda \mathbb{E} \log T'(X^*) + (1-\lambda) \mathbb{E} \log U'(Y^*):$$

$$\begin{aligned}h(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) - \lambda h(X) - (1-\lambda)h(Y) \\&\geq h(\sqrt{\lambda}X^* + \sqrt{1-\lambda}Y^*) - \lambda h(X^*) + (1-\lambda)h(Y^*) = 0\end{aligned}$$

Equality case: $T' = U' = \text{Cst}$ hence $X \propto X^*$, $Y \propto Y^*$.



Outline

Introduction

What is Entropy?

Max/Min Entropy Principles

Equivalence to the Entropy Power Inequality

A Proof that Shannon Missed

Generalization to Linear Transformations

Shannon vs. Rényi

A Proof that Shannon Missed (Revisited)

Generalization to Rényi Entropies



Conclusion for Rényi's Entropy

$$\begin{aligned} h_r(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) &- \lambda h_p(X) - (1-\lambda)h_q(Y) \\ &\geq h_r(\sqrt{\lambda}X^* + \sqrt{1-\lambda}Y^*) - \lambda h_p(X^*) - (1-\lambda)h_q(Y^*) \\ &= \frac{r}{2(r-1)} \left(\frac{\log r}{r} - \frac{\log p}{p} - \frac{\log q}{q} \right) \end{aligned}$$

where $\frac{1}{p} + \frac{1}{q} = 1 + \frac{1}{r}$ (Young's triple with rate λ), i.e., where Hölder conjugates satisfy $\frac{1}{r'} = \underbrace{\frac{1}{p'}}_{\frac{\lambda}{r}} + \underbrace{\frac{1}{q'}}_{\frac{1-\lambda}{r}}$.

Equality case: $T' = U' = \text{Cst}$ hence $X \propto X^*$, $Y \propto Y^*$.

Conclusion for Rényi's Entropy

$$\begin{aligned} h_r(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) - \lambda h_p(X) - (1-\lambda)h_q(Y) \\ \geq h_r(\sqrt{\lambda}X^* + \sqrt{1-\lambda}Y^*) - \lambda h_p(X^*) - (1-\lambda)h_q(Y^*) \\ = \frac{r}{2(r-1)} \left(\frac{\log r}{r} - \frac{\log p}{p} - \frac{\log q}{q} \right) \end{aligned}$$

where $\frac{1}{p} + \frac{1}{q} = 1 + \frac{1}{r}$ (Young's triple with rate λ), i.e., where Hölder conjugates satisfy $\frac{1}{r'} = \underbrace{\frac{1}{p'}}_{\frac{\lambda}{r}} + \underbrace{\frac{1}{q'}}_{\frac{1-\lambda}{r}}$.

Equality case: $T' = U' = \text{Cst}$ hence $X \propto X^*$, $Y \propto Y^*$.

- the natural generalization of the EPI for Rényi entropies

Conclusion for Rényi's Entropy

$$\begin{aligned} h_r(\sqrt{\lambda}X + \sqrt{1-\lambda}Y) &- \lambda h_p(X) - (1-\lambda)h_q(Y) \\ &\geq h_r(\sqrt{\lambda}X^* + \sqrt{1-\lambda}Y^*) - \lambda h_p(X^*) - (1-\lambda)h_q(Y^*) \\ &= \frac{r}{2(r-1)} \left(\frac{\log r}{r} - \frac{\log p}{p} - \frac{\log q}{q} \right) \end{aligned}$$

where $\frac{1}{p} + \frac{1}{q} = 1 + \frac{1}{r}$ (Young's triple with rate λ), i.e., where Hölder conjugates satisfy $\frac{1}{r'} = \underbrace{\frac{1}{p'}}_{\frac{\lambda}{r}} + \underbrace{\frac{1}{q'}}_{\frac{1-\lambda}{r}}$.

Equality case: $T' = U' = \text{Cst}$ hence $X \propto X^*$, $Y \propto Y^*$.

- the natural generalization of the EPI for Rényi entropies
- turns out to be equivalent to **strong Young's inequality and its reverse** [Dembo, Cover, Thomas, 1991] [Barthe, 1998]

Conclusion (Rényi's Entropy)

For $N \geq 2$ variables:

$h_r(\sum_i \sqrt{\lambda_i} X_i) - \sum_i \lambda_i h_{r_i}(X_i)$ is minimum for X_i i.i.d. Gaussian:

$$h_r\left(\sum_i \sqrt{\lambda_i} X_i\right) \geq \sum_i \lambda_i h_{r_i}(X_i) + \frac{r'}{2} \left(\frac{\log r}{r} - \sum_i \frac{\log r_i}{r_i} \right)$$

where Hölder conjugates satisfy $\frac{1}{r'} = \sum_i \frac{1}{r'_i}$ where $\frac{1}{r'_i} = \frac{\lambda_i}{r'}$

Conclusion (Rényi's Entropy)

For $N \geq 2$ variables:

$h_r(\sum_i \sqrt{\lambda_i} X_i) - \sum_i \lambda_i h_{r_i}(X_i)$ is minimum for X_i i.i.d. Gaussian:

$$h_r\left(\sum_i \sqrt{\lambda_i} X_i\right) \geq \sum_i \lambda_i h_{r_i}(X_i) + \frac{r'}{2} \left(\frac{\log r}{r} - \sum_i \frac{\log r_i}{r_i} \right)$$

where Hölder conjugates satisfy $\frac{1}{r'} = \sum_i \frac{1}{r'_i}$ where $\frac{1}{r'_i} = \frac{\lambda_i}{r'}$

In particular for $r > 1$, then $r \geq r_i$, $h_r(X) \leq h_{r_i}(X)$

For any for any convex combination ($\sum_i \lambda_i = 1$), choosing $r'_i = r'/\lambda_i$:

$$h_r\left(\sum_i \sqrt{\lambda_i} X_i\right) \geq \sum_i \lambda_i h_{r_i}(X_i) + \frac{r'}{2} \left(\frac{\log r}{r} - \sum_i \frac{\log r_i}{r_i} \right)$$

Back to the Rényi EPI

Theorem (Rényi Entropy Power Inequality)

$$e^{2h_r(\sum_i X_i)} \geq c \cdot \sum_i e^{2h_r(X_i)}$$

\Leftrightarrow for any convex combination ($\sum_i \lambda_i = 1$)

$$h_r\left(\sum_i \sqrt{\lambda_i} X_i\right) \geq \sum_i \lambda_i h_r(X_i) + \log \sqrt{c}$$

Back to the Rényi EPI

Theorem (Rényi Entropy Power Inequality)

$$e^{2h_r(\sum_i X_i)} \geq c \cdot \sum_i e^{2h_r(X_i)}$$

\iff for any convex combination ($\sum_i \lambda_i = 1$)

$$h_r\left(\sum_i \sqrt{\lambda_i} X_i\right) \geq \sum_i \lambda_i h_r(X_i) + \log \sqrt{c}$$

We have found (for $r > 1$):

$$\log \sqrt{c} = \min \left\{ \frac{r'}{2} \left(\frac{\log r}{r} - \sum_i \frac{\log r_i}{r_i} \right) \text{ s.t. } \sum_i \frac{1}{r_i} = n - \frac{1}{r'} \right\}$$

Optimal Constant

$$\log c = \min \left\{ r' \left(\frac{\log r}{r} - \sum_i \frac{\log r_i}{r_i} \right) \text{ s.t. } \sum_i \frac{1}{r_i} = n - \frac{1}{r'} \right\}$$

Optimal Constant

$$\log c = \min \left\{ r' \left(\frac{\log r}{r} - \sum_i \frac{\log r_i}{r_i} \right) \text{ s.t. } \sum_i \frac{1}{r_i} = n - \frac{1}{r'} \right\}$$

But by the log-sum inequality:

$$\sum_i x_i \log \frac{x_i}{y_i} \geq \sum_i x_i \log \frac{\sum_i x_i}{\sum_i y_i} \quad \text{with equality iff } x_i \propto y_i$$

Optimal Constant

$$\log c = \min \left\{ r' \left(\frac{\log r}{r} - \sum_i \frac{\log r_i}{r_i} \right) \text{ s.t. } \sum_i \frac{1}{r_i} = n - \frac{1}{r'} \right\}$$

But by the log-sum inequality:

$$\sum_i x_i \log \frac{x_i}{y_i} \geq \sum_i x_i \log \frac{\sum_i x_i}{\sum_i y_i} \quad \text{with equality iff } x_i \propto y_i$$

$$\sum_i \frac{\log r_i}{r_i} = - \sum_i \frac{1}{r_i} \log \frac{1}{r_i} \leq - \sum_i \frac{1}{r_i} \log \frac{\sum_i \frac{1}{r_i}}{N} = - \left(N - \frac{1}{r'} \right) \log \frac{N - \frac{1}{r'}}{N}$$

with equality iff r_i are equal.

Optimal Constant

$$\log c = \min \left\{ r' \left(\frac{\log r}{r} - \sum_i \frac{\log r_i}{r_i} \right) \text{ s.t. } \sum_i \frac{1}{r_i} = n - \frac{1}{r'} \right\}$$

But by the log-sum inequality:

$$\sum_i x_i \log \frac{x_i}{y_i} \geq \sum_i x_i \log \frac{\sum_i x_i}{\sum_i y_i} \quad \text{with equality iff } x_i \propto y_i$$

$$\sum_i \frac{\log r_i}{r_i} = - \sum_i \frac{1}{r_i} \log \frac{1}{r_i} \leq - \sum_i \frac{1}{r_i} \log \frac{\sum_i \frac{1}{r_i}}{N} = - \left(N - \frac{1}{r'} \right) \log \frac{N - \frac{1}{r'}}{N}$$

with equality iff r_i are equal.

This gives

$$\log c = \frac{\log r}{r-1} + (Nr' - 1) \log \left(1 - \frac{1}{Nr'} \right)$$

Optimal Constant

$$\log c = \min \left\{ r' \left(\frac{\log r}{r} - \sum_i \frac{\log r_i}{r_i} \right) \text{ s.t. } \sum_i \frac{1}{r_i} = n - \frac{1}{r'} \right\}$$

But by the log-sum inequality:

$$\sum_i x_i \log \frac{x_i}{y_i} \geq \sum_i x_i \log \frac{\sum_i x_i}{\sum_i y_i} \quad \text{with equality iff } x_i \propto y_i$$

$$\sum_i \frac{\log r_i}{r_i} = - \sum_i \frac{1}{r_i} \log \frac{1}{r_i} \leq - \sum_i \frac{1}{r_i} \log \frac{\sum_i \frac{1}{r_i}}{N} = - \left(N - \frac{1}{r'} \right) \log \frac{N - \frac{1}{r'}}{N}$$

with equality iff r_i are equal.

This gives

$$c = r^{\frac{1}{r-1}} \left(1 - \frac{1}{Nr'} \right)^{Nr'-1}$$

which was found by [Ram&Sason,2016] as an improvement of [Bobkov&Chistyakov,2015] (for which $c = r^{\frac{1}{r-1}}/e$)

Thank you !

Questions?

