Matrix Entropy-Power Inequality via Normal Transport

International Conference on the Science of Electrical Engineering (ICSEE'2018)

Eilat, Israel, Dec. 13, 2018

Olivier Rioul & Ram Zamir

rioul@telecom-paristech.fr zamir@eng.tau.ac.il

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: $X \mapsto \mathbf{A}X$
- differential entropy: $h(X) = \int f(x) \log \frac{1}{f(x)} \, \mathrm{d}x$ if X has density f, otherwise $h(X) = -\infty$.
- 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: $X \mapsto \mathbf{A}X$
- differential entropy: $h(X) = \int f(x) \log \frac{1}{f(x)} \, \mathrm{d}x$ if X has density f, otherwise $h(X) = -\infty$.
- 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: $X \mapsto \mathbf{A}X$
- differential entropy: $h(X) = \int f(x) \log \frac{1}{f(x)} \, \mathrm{d}x$ if X has density f, otherwise $h(X) = -\infty$.
- consider $h(\mathbf{A}X)$:
  - assume it is nondegenerate: $h(\mathbf{A}X) > -\infty$
  - $\Longrightarrow$ $\mathbf{A}$ has full row rank
  - $\boxed{\mathbf{A}}$ is an $m \times n$ matrix with $m \leq n$ (or even $m < n$)

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: $X \mapsto \mathbf{A}X$
- differential entropy: $h(X) = \int f(x) \log \frac{1}{f(x)} \, \mathrm{d}x$ if X has density f, otherwise $h(X) = -\infty$.
- consider $h(\mathbf{A}X)$:
  - assume it is nondegenerate: $h(\mathbf{A}X) > -\infty$
  - $\Longrightarrow$ $\mathbf{A}$ has full row rank
  - $\boxed{\mathbf{A}}$ is an $m \times n$ matrix with $m \leq n$ (or even $m < n$)
- $\max h(\mathbf{A}X)$ *or* $\min h(\mathbf{A}X)$? (attained for Gaussian X)

Max/Min Entropy Principle

Let $\widetilde{X}$ be Gaussian with independent components $\widetilde{X}_i$ of same variances: $\mathrm{Var}(\widetilde{X}_i) = \mathrm{Var}(X_i)$.

Theorem (Maximum Entropy Principle)

$$h(\mathbf{A}X) \leq h(\mathbf{A}\widetilde{X}) \qquad \textit{with equality iff } X \textit{ is Gaussian}$$

Max/Min Entropy Principle

Let $\widetilde{X}$ be Gaussian with independent components $\widetilde{X}_i$ of same variances: $\mathrm{Var}(\widetilde{X}_i) = \mathrm{Var}(X_i)$.

Theorem (Maximum Entropy Principle)

$$h(\mathbf{A}X) \leq h(\mathbf{A}\widetilde{X}) \qquad \textit{with equality iff } X \textit{ is Gaussian}$$

- Proof: $h(\mathbf{A}\widetilde{X}) - h(\mathbf{A}X) = D_{\mathrm{KL}}(\mathbf{A}X \| \mathbf{A}\widetilde{X}) \geq 0$ □
- known in the 19th century (Gibbs' inequality) (components need not be independent)

Max/Min Entropy Principle

Let $\widetilde{X}$ be Gaussian with independent components $\widetilde{X}_i$ of same variances: $\mathrm{Var}(\widetilde{X}_i) = \mathrm{Var}(X_i)$.

Theorem (Maximum Entropy Principle)

$$h(\mathbf{A}X) \leq h(\mathbf{A}\widetilde{X}) \qquad \textit{with equality iff } X \textit{ is Gaussian}$$

- Proof: $h(\mathbf{A}\widetilde{X}) - h(\mathbf{A}X) = D_{\mathrm{KL}}(\mathbf{A}X \| \mathbf{A}\widetilde{X}) \geq 0$ □
- known in the 19th century (Gibbs' inequality) (components need not be independent)
- E. T. Jaynes, "Information theory and statistical mechanics," *Physical Review* 106(4):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 $\widetilde{X}$ be Gaussian with independent components $\widetilde{X}_i$ of same variances: $\mathrm{Var}(\widetilde{X}_i) = \mathrm{Var}(X_i)$.

Theorem (Maximum Entropy Principle)

$$h(\mathbf{A}X) \leq h(\mathbf{A}\widetilde{X}) \qquad \textit{with equality iff } X \textit{ is Gaussian}$$

Max/Min Entropy Principle

Let $\widetilde{X}$ be Gaussian with independent components $\widetilde{X}_i$ of same variances: $\mathrm{Var}(\widetilde{X}_i) = \mathrm{Var}(X_i)$.

Theorem (Maximum Entropy Principle)

$$h(\mathbf{A}X) \leq h(\mathbf{A}\widetilde{X}) \qquad \textit{with equality iff } X \textit{ is Gaussian}$$

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

Theorem (Minimun Entropy Principle)

$$h(\mathbf{A}X) \geq h(\mathbf{A}X^*) \qquad \textit{with equality iff } X \textit{ is Gaussian} \ldots \textit{or} \ldots$$

Max/Min Entropy Principle

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

Theorem (Minimun Entropy Principle)

$$h(\mathbf{A}X) \geq h(\mathbf{A}X^*) \qquad \textit{with equality if } X \textit{ is Gaussian...}$$

Max/Min Entropy Principle

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

Theorem (Minimun Entropy Principle)

$$h(\mathbf{A}X) \geq h(\mathbf{A}X^*) \qquad \textit{with equality if } X \textit{ is Gaussian...}$$

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

Max/Min Entropy Principle

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

Theorem (Minimun Entropy Principle)

$$h(\mathbf{A}X) \geq h(\mathbf{A}X^*) \qquad \textit{with equality if } X \textit{ is Gaussian...}$$

... 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, 565-–608, 1981.
- R. Zamir & M. Feder, "A generalization of the entropy power inequality," *IEEE Trans. IT*, 39(5):1723, 1993.
- Application to deconvolution / blind separation where the equality condition is essential

Purpose of this Presentation

Previous Proofs

- Original proof by double induction over (m, n) [Zamir Feder, 1993]
- Advanced proofs integrate (over a continuous path of additive Gaussian perturbation) either:
  - Fisher's information using de Bruijn's identity [Zamir Feder, 1993]
  - or minimum mean-squared error using the I-MMSE relation [Guo Shamai Verdú, 2006]

Purpose of this Presentation

Previous Proofs

- Original proof by double induction over (m, n) [Zamir Feder, 1993]
- Advanced proofs integrate (over a continuous path of additive Gaussian perturbation) either:
  - Fisher's information using de Bruijn's identity [Zamir Feder, 1993]
  - or minimum mean-squared error using the I-MMSE relation [Guo Shamai Verdú, 2006]
- The equality case has not yet been settled in general (as a necessary condition)

Purpose of this Presentation

Previous Proofs

- Original proof by double induction over (m, n) [Zamir Feder, 1993]
- Advanced proofs integrate (over a continuous path of additive Gaussian perturbation) either:
  - Fisher's information using de Bruijn's identity [Zamir Feder, 1993]
  - or minimum mean-squared error using the I-MMSE relation [Guo Shamai Verdú, 2006]
- The equality case has not yet been settled in general (as a necessary condition)

Purpose of this presentation

- The aim here is to:

Purpose of this Presentation

Previous Proofs

- Original proof by double induction over (m, n) [Zamir Feder, 1993]
- Advanced proofs integrate (over a continuous path of additive Gaussian perturbation) either:
  - Fisher's information using de Bruijn's identity [Zamir Feder, 1993]
  - or minimum mean-squared error using the I-MMSE relation [Guo Shamai Verdú, 2006]
- The equality case has not yet been settled in general (as a necessary condition)

Purpose of this presentation

- The aim here is to:
  - provide a simple "transportation" proof;

Purpose of this Presentation

Previous Proofs

- Original proof by double induction over (m, n) [Zamir Feder, 1993]
- Advanced proofs integrate (over a continuous path of additive Gaussian perturbation) either:
  - Fisher's information using de Bruijn's identity [Zamir Feder, 1993]
  - or minimum mean-squared error using the I-MMSE relation [Guo Shamai Verdú, 2006]
- The equality case has not yet been settled in general (as a necessary condition)

Purpose of this presentation

- The aim here is to:
  - provide a simple "transportation" proof;
  - settle the equality case.

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. $h(X^*) = h(X)$, $h(Y) = h(Y^*)$,

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

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. $h(X^*) = h(X)$, $h(Y) = h(Y^*)$,

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

Definition (Entropy Power [Shannon'48])

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

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

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

$$\boxed{N(X) = \exp\big(2h(X)\big)/2\pi e}$$

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. $h(X^*) = h(X)$, $h(Y) = h(Y^*)$,

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

Definition (Entropy Power [Shannon'48])

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

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

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

$$\boxed{N(X) = \exp\big(2h(X)\big)/2\pi e}$$

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. $h(X^*) = h(X)$, $h(Y) = h(Y^*)$,

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

Definition (Entropy Power [Shannon'48])

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

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

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

$$\boxed{N(X) = \exp\big(2h(X)\big)/2\pi e} \qquad N(X^*) = \mathrm{Var}(X^*)$$

Simplest Nontrivial Case: $(m,n) = (1,2)$

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

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

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

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

Definition (Entropy Power [Shannon'48])

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

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

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

$$\boxed{N(X) = \exp\big(2h(X)\big)/2\pi e} \qquad N(X^*) = \mathrm{Var}(X^*)$$

Simplest Nontrivial Case: $(m,n) = (1,2)$

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

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

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

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

Definition (Entropy Power [Shannon'48])

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

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

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

$$\boxed{N(X) = \exp\big(2h(X)\big)/2\pi e} \qquad N(X^*) = \mathrm{Var}(X^*)$$

Simplest Nontrivial Case: $(m,n) = (1,2)$

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

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

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

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

Definition (Entropy Power [Shannon'48])

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

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

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

$$\boxed{N(X) = \exp\big(2h(X)\big)/2\pi e} \qquad N(X^*) = \mathrm{Var}(X^*)$$

Simplest Nontrivial Case: $(m,n) = (1,2)$

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

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

For any two independent X, Y,

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

Definition (Entropy Power [Shannon'48])

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

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

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

$$\boxed{N(X) = \exp\big(2h(X)\big)/2\pi e} \qquad N(X^*) = \mathrm{Var}(X^*)$$

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,

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

Definition (Entropy Power [Shannon'48])

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

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

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

$$\boxed{N(X) = \exp\big(2h(X)\big)/2\pi e} \qquad N(X^*) = \mathrm{Var}(X^*)$$

The Entropy-Power Inequality (EPI)

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* $\overline{N}_1$ *and* $\overline{N}_2$. *Then the entropy power of the sum,* $\overline{N}_3$, *is bounded by*

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

White Gaussian noise has the peculiar property that it can absorb any other noise or signal ensemble which may be added to it with a resultant entropy power approximately equal to the sum of the white noise power and the signal power (measured from the average signal value, which is normally zero), provided the signal power is small, in a certain sense, compared to the noise.

The EPI has a Long History

1948 Stated and “proved” by Shannon in his seminal paper

The EPI has a Long History

1948 Stated and “proved” by Shannon in his seminal paper

1959 Stam’s proof using Fisher information

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

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 O. Rioul, “Yet another proof of the entropy power inequality,” *IEEE Trans IT* 63(6):3595–3599, 2017 using normal transport

A Simple Change of Variables

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.

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

A Simple Change of Variables

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.

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

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.

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

A Simple Change of Variables: Entropy

Lemma (Change of variable [Shannon'48])

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

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

Proof.

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

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

□

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

A Proof that Shannon Missed. . .

Proceed to prove the inequality $\boxed{h(aX + bY) \geq h(aX^* + bY^*)}$

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 $\boxed{h(aX + bY) \geq h(aX^* + bY^*)}$

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 $\boxed{h(aX + bY) \geq h(aX^* + bY^*)}$

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 $\boxed{h(aX + bY) \geq h(aX^* + bY^*)}$

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. $\boxed{X^*, Y^* \text{ are i.i.d.}}$ Gaussian.

A Proof that Shannon Missed. . .

Proceed to prove the inequality $\boxed{h(aX + bY) \geq h(aX^* + bY^*)}$

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. $\boxed{X^*, Y^* \text{ are i.i.d.}}$ Gaussian.
2. We may always normalize: $\boxed{a^2 + b^2 = 1}$.

A Proof that Shannon Missed. . .

Proceed to prove the inequality $\boxed{h(aX + bY) \geq h(aX^* + bY^*)}$

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. $\boxed{X^*, Y^* \text{ are i.i.d.}}$ Gaussian.
2. We may always normalize: $\boxed{a^2 + b^2 = 1}$. Otherwise:
   - divide a, b by $\Delta = \sqrt{a^2 + b^2}$;
   - the $\log \Delta$ terms cancel.

A Proof that Shannon Missed. . .

Proceed to prove the inequality $\boxed{h(aX + bY) \geq h(aX^* + bY^*)}$ 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. $\boxed{X^*, Y^* \text{ are i.i.d.}}$ Gaussian.
2. We may always normalize: $\boxed{a^2 + b^2 = 1}$. Otherwise:
   - divide a, b by $\Delta = \sqrt{a^2 + b^2}$;
   - the $\log \Delta$ terms cancel.
3. Make the changes of variables $X = T(X^*)$, $Y = U(Y^*)$:

A Proof that Shannon Missed. . .

Proceed to prove the inequality $\boxed{h(aX + bY) \geq h(aX^* + bY^*)}$

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

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

A Proof that Shannon Missed. . .

Proceed to prove the inequality $\boxed{h(aX + bY) \geq h(aX^* + bY^*)}$

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

   One is led to prove $\boxed{h(aT(X^*) + bU(Y^*)) \geq h(aX^* + bY^*)}$
4. Define $\widetilde{X} = aX^* + bY^*$.

A Proof that Shannon Missed. . .

Proceed to prove the inequality $\boxed{h(aX + bY) \geq h(aX^* + bY^*)}$
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. $\boxed{X^*, Y^* \text{ are i.i.d.}}$ Gaussian.
2. We may always normalize: $\boxed{a^2 + b^2 = 1}$. Otherwise:
   - divide a, b by $\Delta = \sqrt{a^2 + b^2}$;
   - the $\log \Delta$ terms cancel.
3. Make the changes of variables $X = T(X^*)$, $Y = U(Y^*)$:
   One is led to prove $\boxed{h(aT(X^*) + bU(Y^*)) \geq h(aX^* + bY^*)}$
4. Define $\widetilde{X} = aX^* + bY^*$. Complete the rotation: $\widetilde{Y} = -bX^* + aY^*$
   so that $\boxed{\widetilde{X}, \widetilde{Y} \text{ are i.i.d.}}$ Gaussian

A Proof that Shannon Missed. . .

Proceed to prove the inequality $\boxed{h(aX + bY) \geq h(aX^* + bY^*)}$

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

   One is led to prove $\boxed{h(aT(X^*) + bU(Y^*)) \geq h(aX^* + bY^*)}$
4. Define $\widetilde{X} = aX^* + bY^*$. Complete the rotation: $\widetilde{Y} = -bX^* + aY^*$

   so that $\boxed{\widetilde{X}, \widetilde{Y} \text{ are i.i.d.}}$ Gaussian and $\boxed{X^* = a\widetilde{X} - b\widetilde{Y}}$, $\boxed{Y^* = b\widetilde{X} + a\widetilde{Y}}$.

A Proof that Shannon Missed

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

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

A Proof that Shannon Missed

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

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

5. Since conditioning reduces entropy:

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

A Proof that Shannon Missed

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

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

5. Since conditioning reduces entropy:

$$h(aT(X^*) + bU(Y^*)) = h(aT(a\widetilde{X} - b\widetilde{Y}) + bU(b\widetilde{X} + a\widetilde{Y}))$$
$$\geq h(\underbrace{aT(a\widetilde{X} - b\widetilde{Y}) + bU(b\widetilde{X} + a\widetilde{Y})}_{T_{\widetilde{Y}}(\widetilde{X})} | \widetilde{Y})$$

A Proof that Shannon Missed

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

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

5. Since conditioning reduces entropy:

$$h(aT(X^*) + bU(Y^*)) = h(aT(a\widetilde{X} - b\widetilde{Y}) + bU(b\widetilde{X} + a\widetilde{Y}))$$
$$\geq h(\underbrace{aT(a\widetilde{X} - b\widetilde{Y}) + bU(b\widetilde{X} + a\widetilde{Y})}_{T_{\widetilde{Y}}(\widetilde{X})} | \widetilde{Y})$$

6. By the change of variable:

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

A Proof that Shannon Missed

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

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

5. Since conditioning reduces entropy:

$$h(aT(X^*) + bU(Y^*)) = h(aT(a\widetilde{X} - b\widetilde{Y}) + bU(b\widetilde{X} + a\widetilde{Y}))$$
$$\geq h(\underbrace{aT(a\widetilde{X} - b\widetilde{Y}) + bU(b\widetilde{X} + a\widetilde{Y})}_{T_{\widetilde{Y}}(\widetilde{X})} | \widetilde{Y})$$

6. By the change of variable:

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

A Proof that Shannon Missed

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

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

5. Since conditioning reduces entropy:

$$h(aT(X^*)+bU(Y^*)) = h(aT(a\widetilde{X}-b\widetilde{Y})+bU(b\widetilde{X}+a\widetilde{Y}))$$
$$\geq h(\underbrace{aT(a\widetilde{X}-b\widetilde{Y})+bU(b\widetilde{X}+a\widetilde{Y})}_{T_{\widetilde{Y}}(\widetilde{X})}|\widetilde{Y})$$

6. By the change of variable:

$$= h(\widetilde{X}) + \mathbb{E}\log T'_{\widetilde{Y}}(\widetilde{X})$$
$$= h(\widetilde{X}) + \mathbb{E}\log\left(a^2T'(a\widetilde{X}-b\widetilde{Y}) + b^2U'(b\widetilde{X}+a\widetilde{Y})\right)$$

A Proof that Shannon Missed

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

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

5. Since conditioning reduces entropy:

$$h(aT(X^*)+bU(Y^*)) = h(aT(a\widetilde{X}-b\widetilde{Y})+bU(b\widetilde{X}+a\widetilde{Y}))$$
$$\geq h(\underbrace{aT(a\widetilde{X}-b\widetilde{Y})+bU(b\widetilde{X}+a\widetilde{Y})}_{T_{\widetilde{Y}}(\widetilde{X})}|\widetilde{Y})$$

6. By the change of variable:

$$= h(\widetilde{X}) + \mathbb{E}\log T'_{\widetilde{Y}}(\widetilde{X})$$
$$= h(\widetilde{X}) + \mathbb{E}\log\left(a^2T'(a\widetilde{X}-b\widetilde{Y})+b^2U'(b\widetilde{X}+a\widetilde{Y})\right)$$
$$= h(aX^*+bY^*) + \mathbb{E}\log\left(a^2T'(X^*)+b^2U'(Y^*)\right)$$

A Proof that Shannon Missed

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

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

5. Since conditioning reduces entropy:

$$h(aT(X^*)+bU(Y^*)) = h(aT(a\widetilde{X}-b\widetilde{Y})+bU(b\widetilde{X}+a\widetilde{Y}))$$
$$\geq h(\underbrace{aT(a\widetilde{X}-b\widetilde{Y})+bU(b\widetilde{X}+a\widetilde{Y})}_{T_{\widetilde{Y}}(\widetilde{X})}|\widetilde{Y})$$

6. By the change of variable:

$$= h(\widetilde{X}) + \mathbb{E}\log T'_{\widetilde{Y}}(\widetilde{X})$$
$$= h(\widetilde{X}) + \mathbb{E}\log\left(a^2T'(a\widetilde{X}-b\widetilde{Y})+b^2U'(b\widetilde{X}+a\widetilde{Y})\right)$$
$$= h(aX^*+bY^*) + \mathbb{E}\log\left(a^2T'(X^*)+b^2U'(Y^*)\right)$$

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 $\boxed{h(aT(X^*)+bU(Y^*)) \geq h(aX^*+bY^*)}$

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

5. Since conditioning reduces entropy:

$$h(aT(X^*)+bU(Y^*)) = h(aT(a\widetilde{X}-b\widetilde{Y})+bU(b\widetilde{X}+a\widetilde{Y}))$$
$$\geq h(\underbrace{aT(a\widetilde{X}-b\widetilde{Y})+bU(b\widetilde{X}+a\widetilde{Y})}_{T_{\widetilde{Y}}(\widetilde{X})}|\widetilde{Y})$$

6. By the change of variable:

$$= h(\widetilde{X}) + \mathbb{E}\log T'_{\widetilde{Y}}(\widetilde{X})$$
$$= h(\widetilde{X}) + \mathbb{E}\log\left(a^2T'(a\widetilde{X}-b\widetilde{Y})+b^2U'(b\widetilde{X}+a\widetilde{Y})\right)$$
$$= h(aX^*+bY^*) + \mathbb{E}\log\left(a^2T'(X^*)+b^2U'(Y^*)\right)$$

7. By concavity of the log:

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

A Proof that Shannon Missed

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

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

5. Since conditioning reduces entropy:

$$h(aT(X^*)+bU(Y^*)) = h(aT(a\widetilde{X}-b\widetilde{Y})+bU(b\widetilde{X}+a\widetilde{Y}))$$
$$\geq h(\underbrace{aT(a\widetilde{X}-b\widetilde{Y})+bU(b\widetilde{X}+a\widetilde{Y})}_{T_{\widetilde{Y}}(\widetilde{X})}|\widetilde{Y})$$

6. By the change of variable:

$$= h(\widetilde{X}) + \mathbb{E}\log T'_{\widetilde{Y}}(\widetilde{X})$$
$$= h(\widetilde{X}) + \mathbb{E}\log\left(a^2T'(a\widetilde{X}-b\widetilde{Y})+b^2U'(b\widetilde{X}+a\widetilde{Y})\right)$$
$$= h(aX^*+bY^*) + \mathbb{E}\log\left(a^2T'(X^*)+b^2U'(Y^*)\right)$$

7. By concavity of the log:

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

Equality Case

For nonzero a, b:

- in log concavity inequality:

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

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

Equality Case

For nonzero a, b:

- in log concavity inequality:

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

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

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

Equality Case

For nonzero a, b:

- in log concavity inequality:

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

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

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

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

Equality Case

For nonzero a, b:

- in log concavity inequality:

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

$\Longrightarrow$ $T'(X^*) = U'(X^*) = c > 0$ constant a.e.
$\Longrightarrow$ T, U are linear: $X = T(X^*) = cX^*$, $Y = U(Y^*) = cY^*$ Gaussian.
$\Longrightarrow$ $c = 1$ since $h(X) = h(X^*)$, $h(Y) = h(Y^*)$.

- in information inequality:

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

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

Extension to Linear Transformations

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

Extension to Linear Transformations

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

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

Extension to Linear Transformations

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

- 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.

Extension to Linear Transformations

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

- 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): $h(\mathbf{A}X) = h(\mathbf{A}X^*) = -\infty$.

Extension to Linear Transformations

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

- 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): $h(\mathbf{A}X) = h(\mathbf{A}X^*) = -\infty$.
- The difference $h(\mathbf{A}X) - h(\mathbf{A}X^*)$ is invariant by elementary row operations. By the Gram-Schmidt procedure, we may assume that the rows of $\mathbf{A}$ are orthonormal: $\mathbf{A}\mathbf{A}^t = \mathbf{I}$.

Extension to Linear Transformations

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

- 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): $h(\mathbf{A}X) = h(\mathbf{A}X^*) = -\infty$.
- The difference $h(\mathbf{A}X) - h(\mathbf{A}X^*)$ is invariant by elementary row operations. By the Gram-Schmidt procedure, we may assume that the rows of $\mathbf{A}$ are orthonormal: $\mathbf{A}\mathbf{A}^t = \mathbf{I}$.
- Extend $\mathbf{A}$ to an orthogonal matrix $\mathbf{A}' = \left(\begin{array}{c} \mathbf{A} \\ \hline \mathbf{A}^C \end{array}\right)$

Extension to Linear Transformations

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

Extension to Linear Transformations

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

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

Extension to Linear Transformations

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

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

- But the Jacobian matrix of $\mathbf{T}_{\widetilde{X}^c}(\widetilde{X}) = \mathbf{A}\,\mathbf{T}(\mathbf{A}'^t \, \widetilde{X}') = \mathbf{A}\,\mathbf{T}(\mathbf{A}^t \widetilde{X} + (\mathbf{A}^c)^t \widetilde{X}^c)$ for fixed $\widetilde{X}^c$ is $\mathbf{T}'_{\widetilde{X}^c}(\widetilde{X}) = \mathbf{A}\mathbf{T}'(\mathbf{A}'^t \, \widetilde{X}')\mathbf{A}^t = \mathbf{A}\mathbf{T}'(X^*)\mathbf{A}^t$ where $\mathbf{T}'(X^*) = \mathrm{diag}(T_i'(X_i^*))$

Extension to Linear Transformations

- The change of variables in the entropy yields

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

Extension to Linear Transformations

- The change of variables in the entropy yields

$$h(\mathbf{A}X) \geq h(\mathbf{A}\,\mathbf{T}(\mathbf{A}'^{t}\,\widetilde{X}')\,|\,\widetilde{X}^{c})$$
$$= h(\widetilde{X}|\widetilde{X}^{c}) + \mathbb{E}\log\det(\mathbf{A}\mathbf{T}'(X^{*})\mathbf{A}^{t})$$

- By the concavity of the logarithm: [Zamir-Feder's Lemma, 1993]

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

thus

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

Extension to Linear Transformations

- The change of variables in the entropy yields

$$h(\mathbf{A}X) \geq h(\mathbf{A}\,\mathbf{T}(\mathbf{A}'^{t}\,\widetilde{X}')\,|\,\widetilde{X}^{c})$$
$$= h(\widetilde{X}|\widetilde{X}^{c}) + \mathbb{E}\log\det(\mathbf{A}\mathbf{T}'(X^{*})\mathbf{A}^{t})$$

- By the concavity of the logarithm: [Zamir-Feder's Lemma, 1993]

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

thus

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

- But $h(\widetilde{X}|\widetilde{X}^{c}) = h(\widetilde{X}) = h(\mathbf{A}X^{*})$ and $\mathbb{E}\log T_i'(\widetilde{X}_i) = h(T_i(\widetilde{X}_i)) - h(\widetilde{X}_i) = h(X_i) - h(\widetilde{X}_i) = 0$; so

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

Extension to Linear Transformations

- The change of variables in the entropy yields

$$h(\mathbf{A}X) \geq h(\mathbf{A}\,\mathbf{T}(\mathbf{A}'^{t}\,\widetilde{X}')\,|\,\widetilde{X}^c)$$
$$= h(\widetilde{X}|\widetilde{X}^c) + \mathbb{E}\log\det(\mathbf{A}\mathbf{T}'(X^*)\mathbf{A}^t)$$

- By the concavity of the logarithm: [Zamir-Feder's Lemma, 1993]

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

thus

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

- But $h(\widetilde{X}|\widetilde{X}^c) = h(\widetilde{X}) = h(\mathbf{A}X^*)$ and
$\mathbb{E}\log T_i'(\widetilde{X}_i) = h(T_i(\widetilde{X}_i)) - h(\widetilde{X}_i) = h(X_i) - h(\widetilde{X}_i) = 0$; so

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

- Equality case?

Equality Case

Definitions

A component X_j of X is

- present in the linear mixture $\mathbf{A}X$ if $\mathbf{A}X$ depends on X_j (the jth column of $\mathbf{A}$ is non zero).
- recoverable from the the linear mixture $\mathbf{A}X$ if there exists a linear transformation (line vector ℓ) such that $X_j = \ell(\mathbf{A}X)$ a.e. (i.e., $\exists \ell$ s.t. $\ell\mathbf{A} = (0 \cdots 0 \underbrace{1}_{j} 0 \cdots 0)$).

Equality Case

Definitions

A component X_j of X is

- present in the linear mixture $\mathbf{A}X$ if $\mathbf{A}X$ depends on X_j (the jth column of $\mathbf{A}$ is non zero).
- recoverable from the the linear mixture $\mathbf{A}X$ if there exists a linear transformation (line vector ℓ) such that $X_j = \ell(\mathbf{A}X)$ a.e. (i.e., $\exists \ell$ s.t. $\ell\mathbf{A} = (0 \cdots 0 \underbrace{1}_{j} 0 \cdots 0)$).

Theorem

Equality $h(\mathbf{A}X) = h(\mathbf{A}X^*)$ holds iff all unrecoverable components present in the mixture are Gaussian.

Equality Case

Definitions

A component X_j of X is

- present in the linear mixture $\mathbf{A}X$ if $\mathbf{A}X$ depends on X_j (the jth column of $\mathbf{A}$ is non zero).
- recoverable from the the linear mixture $\mathbf{A}X$ if there exists a linear transformation (line vector ℓ) such that $X_j = \ell(\mathbf{A}X)$ a.e. (i.e., $\exists \ell$ s.t. $\ell\mathbf{A} = (0 \cdots 0 \underbrace{1}_{j} 0 \cdots 0)$).

Theorem

Equality $h(\mathbf{A}X) = h(\mathbf{A}X^*)$ holds iff all unrecoverable components present in the mixture are Gaussian.

Matrix Entropy-Power Inequality via Normal Transport

International Conference on the Science of Electrical Engineering (ICSEE'2018)

Eilat, Israel, Dec. 13, 2018

Thank you for your attention!

IEEE Israel

Olivier Rioul & Ram Zamir

rioul@telecom-paristech.fr zamir@eng.tau.ac.il