ECE253/CSE208 Introduction to Information Theory

Lecture 13: Differential Entropy and Entropy Power Inequality (EPI)

Dr. Yu Zhang

ECE Department
University of California, Santa Cruz

- Chap 8 of *Elements of Information Theory (2nd Edition) by Thomas Cover & Joy Thomas*.

Continuous Sources

Consider a source model: $\{X_t \in \mathcal{X}, t \in \mathcal{T}\}$

- Discrete sources: Both $\mathcal{X}$ and $\mathcal{T}$ are discrete.
- Continuous sources:
  1. Discrete-time continuous sources: $\mathcal{X}$ is continuous; $\mathcal{T}$ is discrete.
  2. Waveform sources: Both $\mathcal{X}$ and $\mathcal{T}$ are continuous.

- So far we have studied information measures and their properties for discrete-time discrete-alphabet sources and systems (DMC).
- In this lecture, we focus on discrete-time continuous-alphabet (real-valued) sources.

Differential Entropy

Definition

The differential entropy $h(X)$ of a continuous random variable X with density $f(x)$ and support $\mathcal{S}$ is defined as

$$\boxed{h(X) = \mathrm{E}(-\log f(X)) = -\int_{\mathcal{S}} f(x) \log f(x) dx}$$

Example (Differential entropy can be negative)

Consider $X \sim$ Uniform$[0, a]$ for $a > 0$, its differential entropy is

$$h(X) = -\int_0^a \frac{1}{a} \log\left(\frac{1}{a}\right) dx = \log a \implies h(X) < 0 \text{ for } 0 < a < 1$$

Example

Consider a continuous RV X with pdf $f_X(x) = 2x, \forall x \in \mathcal{S}_X := [0, 1)$. Its differential entropy equals to

$$h(X) = -\int_0^1 2x \times \log(2x) dx = \frac{1}{2}x^2(\log e - 2\log(2x))\Big|_0^1 \approx -0.279 \text{ bits.}$$

Entropy of Normal Distribution

Example (Entropy of normal distribution)

Let $X \sim \mathcal{N}(0, \sigma^2)$ with the pdf $\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$. Then, we have

$$
\begin{aligned}
h(X) &= \mathrm{E}\left[\ln \frac{1}{\phi(X)}\right] && (1)\\
&= \mathrm{E}\left[\frac{X^2}{2\sigma^2} + \ln(\sigma\sqrt{2\pi})\right] && (2)\\
&= \frac{1}{2} + \frac{1}{2}\ln(2\pi\sigma^2) && (3)\\
&= \frac{1}{2}\ln(2\pi e\sigma^2) \text{ nats} && (4)\\
&= \frac{1}{2}\log(2\pi e\sigma^2) \text{ bits} && (5)
\end{aligned}
$$

AEP for Continuous Random Variables

Theorem

Let i.i.d. $X^n \sim f(x)$*. Then* $-\frac{1}{n}\log f(X^n) \xrightarrow{\text{i.p.}} \mathrm{E}(-\log f(X)) = h(X)$*.*

Definition (Typical set)

$A_\epsilon^{(n)} = \left\{x^n \in S^n : |-\frac{1}{n}\log f(x^n) - h(X)| \le \epsilon\right\}$, where $f(x^n) = \prod_{i=1}^n f(x_i)$.

Properties of the typical set.

1. $\Pr(A_\epsilon^{(n)}) > 1 - \epsilon$ for n sufficiently large
2. $\mathrm{Vol}(A_\epsilon^{(n)}) \le 2^{n(h(X)+\epsilon)}$ for all n
3. $\mathrm{Vol}(A_\epsilon^{(n)}) \ge (1-\epsilon)2^{n(h(X)-\epsilon)}$ for n sufficiently large,

where the volume of a set $A \subset \mathbb{R}^n$ is defined as $\mathrm{Vol}(A) = \int_A dx_1 dx_2 \cdots dx_n$.

Theorem (cf. section 3.3 in the book)

$A_\epsilon^{(n)}$ *is the smallest volume set w.p. at least* $1-\epsilon$*, to first order in the exponent.*

Proof of Typical Set Properties

Property 1 is a direct result from the AEP theorem. Property 2 and 3 are due to the lower and upper bounds of $f(x^n)$. The details are given as follows

$$\begin{aligned}
1 &= \int_{S^n} f(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n \\
&\geq \int_{A_\epsilon^{(n)}} f(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n \\
&\geq \int_{A_\epsilon^{(n)}} 2^{-n(h(X)+\epsilon)}\, dx_1\, dx_2 \cdots dx_n \\
&= 2^{-n(h(X)+\epsilon)} \int_{A_\epsilon^{(n)}} dx_1\, dx_2 \cdots dx_n \\
&= 2^{-n(h(X)+\epsilon)}\ \mathrm{Vol}\left(A_\epsilon^{(n)}\right).
\end{aligned}$$

Proof of Typical Set Properties (cont'd)

$$
\begin{aligned}
1-\epsilon &\le \int_{A_\epsilon^{(n)}} f(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n \\
&\le \int_{A_\epsilon^{(n)}} 2^{-n(h(X)-\epsilon)}\, dx_1\, dx_2 \cdots dx_n \\
&= 2^{-n(h(X)-\epsilon)} \int_{A_\epsilon^{(n)}} dx_1\, dx_2 \cdots dx_n \\
&= 2^{-n(h(X)-\epsilon)}\ \mathrm{Vol}\left(A_\epsilon^{(n)}\right),
\end{aligned}
$$

Implications of Differential Entropy

1. The volume of the smallest set that contains most of the probability $\approx 2^{nh} \implies$ The corresponding side length is $(2^{nh})^{\frac{1}{n}} = 2^h$.
2. $h(X)$ is the logarithm of the *equivalent side length* of the smallest set that contains most of the probability: X with low entropy is confined to a small effective volume, and widely dispersed if $h(X)$ is big.
3. $h(X)$ can be negative, but $2^{nh(X)}$ is always positive.
4. $h(X)$ is related to $\mathrm{Vol}(A_\epsilon^{(n)})$ while $I(\theta)$ is related to the surface area of $A_\epsilon^{(n)}$, where $I(\theta)$ is the Fisher information[1]: A way of measuring the amount of info that an observable random variable X carries about an unknown parameter θ of a distribution that models X. Formally, it is the variance of the score.

$$\mathcal{I}(\theta) = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X;\theta)\right)^2 \middle| \theta\right] = \int_{\mathbb{R}} \left(\frac{\partial}{\partial\theta}\log f(x;\theta)\right)^2 f(x;\theta)\,dx$$

[1] See details in Sections 11.10 and 17.8 of Cover's book.

Relationship with $H(X)$

Theorem (On average $h(X) + n$ bits are required to describe X to n-bit accuracy.)

Consider a continuous RV $X \sim f(x)$ and its quantized version $X^\Delta = x_i$ for $i\Delta \leq X < (i+1)\Delta$, where $f(x_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} f(x)\, dx$. If X is Riemann integrable, then $H(X^\Delta) + \log \Delta \xrightarrow{\Delta \to 0} h(X)$. Thus, the entropy of an n-bit quantization of a continuous RV X is approximately $h(X) + n$.

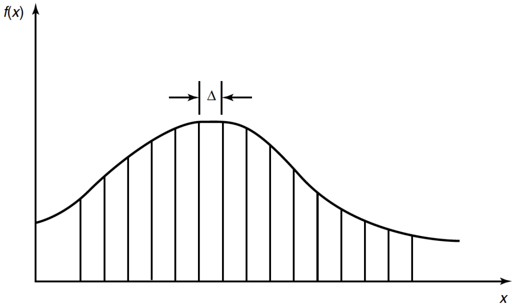

Figure: Quantization of a continuous random variable X.

Proof of Quantization Error

Note that we have

$$p_i \triangleq \Pr(X^\Delta = x_i) = \int_{i\Delta}^{(i+1)\Delta} f(x)\,dx = f(x_i)\Delta.$$

Thus, the entropy of the quantized version is

$$\begin{aligned}
H(X^\Delta) &= -\sum_{-\infty}^{\infty} p_i \log p_i \\
&= -\sum_{-\infty}^{\infty} f(x_i)\Delta \log(f(x_i)\Delta) \\
&= -\sum_{-\infty}^{\infty} \Delta f(x_i) \log f(x_i) - \sum_{-\infty}^{\infty} f(x_i)\Delta \log(\Delta) \\
&= -\sum_{-\infty}^{\infty} \Delta f(x_i) \log f(x_i) - \log(\Delta) \\
&\implies \boxed{\lim_{\Delta \to 0} H(X^\Delta) + \log \Delta = h(X)}.
\end{aligned}$$

If $\Delta = \frac{1}{2^n}$ (n-bit quantization), then $H(X^\Delta) = h(X) + n$.

Joint, Conditional, Relative Entropy, and Mutual Information

Definition

$$h(X^n) = -\int f(x^n) \log f(x^n)\, dx^n \tag{6}$$

$$h(X|Y) = -\int f(x,y) \log f(x|y)\, dxdy = h(X,Y) - h(Y) \tag{7}$$

$$D(f||g) = \int f \log \frac{f}{g} \tag{8}$$

$$\begin{aligned} I(X;Y) &= D\left(f(x,y)||f(x)f(y)\right) &(9)\\ &= h(X) - h(X|Y) = h(Y) - h(Y|X) = h(X) + h(Y) - h(X,Y) &(10)\\ &= \lim_{\Delta\to 0} I(X^\Delta; Y^\Delta) &(11)\\ &= \sup_{\mathcal{P},\mathcal{Q}} I([X]_\mathcal{P}; [Y]_\mathcal{Q}) &(12)\end{aligned}$$

whee the 'supremum' is taken over all finite partitions $\mathcal{P}$ and $\mathcal{Q}$. The quantization of X by $\mathcal{P}$ is the discrete RV defined by $\Pr([X]_\mathcal{P} = i) = \Pr(X \in P_i) = \int_{P_i} f(x)\, dx$, where the disjoint sets P_i's form a partition of the range of X such that $\cup_i P_i = \mathcal{X}$.

Entropy of Gaussian Distribution

Theorem (Entropy of multivariate normal distribution)

$$\boxed{h(\mathcal{N}_n(\boldsymbol{\mu}, \mathbf{K})) = \frac{1}{2}\log(2\pi e)^n |\mathbf{K}| \text{ bits}}$$

Proof: Note that the pdf is $\phi(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n |\mathbf{K}|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{K}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$

$$\begin{aligned}
h(\mathcal{N}_n(\boldsymbol{\mu}, \mathbf{K})) &= \mathrm{E}\left[\ln \frac{1}{\phi(\mathbf{x})}\right] \\
&= \mathrm{E}\left[\frac{1}{2}\mathrm{Tr}\left((\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{K}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) + \frac{1}{2}\ln\left[(2\pi)^n|\mathbf{K}|\right]\right] \\
&= \frac{1}{2}\mathrm{Tr}\left(\mathrm{E}\left[\mathbf{K}^{-1}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top\right]\right) + \frac{1}{2}\ln\left[(2\pi)^n|\mathbf{K}|\right] \\
&= \frac{1}{2}\mathrm{Tr}\left(\mathbf{K}^{-1}\mathbf{K}\right) + \frac{1}{2}\ln\left[(2\pi)^n|\mathbf{K}|\right] \\
&= \frac{1}{2}\ln\left[(2\pi e)^n|\mathbf{K}|\right] \text{ nats} \\
&= \frac{1}{2}\log\left[(2\pi e)^n|\mathbf{K}|\right] \text{ bits}
\end{aligned}$$

Invariance of Quadratic Term Evaluation

From the previous derivation of $h(\mathcal{N}_n(\boldsymbol{\mu}, \mathbf{K}))$, we observe that

$$\boxed{\int \phi(\mathbf{x}) \ln \phi(\mathbf{x})\, d\mathbf{x} = \int f(\mathbf{x}) \ln \phi(\mathbf{x})\, d\mathbf{x}}$$

for any density function $f(\mathbf{x})$ with the same covariance matrix $\mathbf{K}$.

This is due to the fact that

- $\int \phi(\mathbf{x})\, d\mathbf{x} = \int f(\mathbf{x})\, d\mathbf{x} = 1$
- The quadratic term $\mathrm{Tr}(\mathrm{E}(\cdot))$ has the same value for $\phi(\mathbf{x})$ and $f(\mathbf{x})$

Later we will use this useful result to prove that "Gaussian maximizes the entropy for a fixed covariance."

Entropy of Gaussian Distribution (Cont'd)

Example

Let $(X, Y) \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$, where $\mathbf{K} = \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix} \implies h(X) = h(Y) = \frac{1}{2}\log(2\pi e)\sigma^2$

$$h(X, Y) = \frac{1}{2}\log(2\pi e)^2|\mathbf{K}| = \frac{1}{2}\log\left[(2\pi e)^2\sigma^4(1-\rho^2)\right] \implies$$

$$I(X; Y) = h(X) + h(Y) - h(X, Y) = -\frac{1}{2}\log(1-\rho^2) \implies$$

$$\begin{cases} X \perp\!\!\!\perp Y \ (\rho = 0) & \Leftrightarrow I(X; Y) = 0 \\ X \parallel Y \ (\rho = \pm 1) & \Leftrightarrow I(X; Y) = \infty \end{cases}$$

Properties of Differential Entropy and KL Divergence

Similar to the discrete case, we have

- $D(f||g) \geq 0 \implies I(X;Y) \geq 0$, $h(X|Y) \leq h(X)$.
- $D(f||g)$ is finite only if $\mathbf{S}_f \subseteq \mathbf{S}_g$.
- Independence bound: $h(X^n) = \sum_{i=1}^{n} h(X_i|X^{i-1}) \leq \sum_i h(X_i)$.
- **Differential entropy is translation invariant:**

  $h(X+c) = h(X)$ for any constant $c \in \mathbb{R}$, and $h(X+Y|Y) = h(X|Y)$.

  This result can be generalized to n-tuple case:

$$\boxed{h(X^n + Y^n|Y^n) = h(X^n|Y^n)}.$$

Differential Entropy under Invertible Transformation

Different from the discrete case, **differential entropy is generally non-invariant under invertible mapping.**

- *Linear mapping:* For a continuous random vector $\mathbf{x} \in \mathbb{R}^n$ and invertible matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, it holds that

$$\boxed{h(\mathbf{A}\mathbf{x}) = h(\mathbf{x}) + \log |\det(\mathbf{A})|.}$$

  For one dimension: $h(aX) = h(X) + \log |a|$, which can be proved by the property $f_Y(y) = \frac{1}{|a|} f_X(\frac{y}{a})$ for $Y = aX$.

- *Nonlinear mapping:* For an invertible transformation $g : \mathbf{x} \to \mathbf{y}$, it holds that

$$\boxed{h(\mathbf{y}) = h(\mathbf{x}) + \int_{\mathbb{R}^n} f_X(\mathbf{x}) \log |\det(\mathbf{J}_g(\mathbf{x}))| \, dx,}$$

  where $\mathbf{J}_g(\mathbf{x})$ is the Jacobian matrix of the vector-valued function g.

Maximum Entropy

Among all random variables with a given variance, the Gaussian has the highest entropy, and thus the hardest to describe.

Theorem (Normal distribution maximizes the entropy for a given covariance)

Let random vector $\mathbf{x} \in \mathbb{R}^n$ *have zero mean and covariance* $\mathbf{K}$. *Then,*

$$\max_{\mathrm{E}(\mathbf{x}\mathbf{x}^\top)=\mathbf{K}} h(\mathbf{x}) = \frac{1}{2}\log(2\pi e)^n|\mathbf{K}|,$$

where the maximum is attained iif $\mathbf{x} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{K})$.

Proof: Let $f(\mathbf{x})$ be any density function with covariance $\mathbf{K}$, and $g(\mathbf{x}) \sim \mathcal{N}_n(\mathbf{0}, \mathbf{K})$. Then, we have

$$0 \leq D(f||g) = -h(f) - \int f \log g = -h(f) - \int g \log g = -h(f) + h(g),$$

where the second equality is due to the fact f and g have the same covariance $\mathbf{K}$.

Minimum Estimation Error

Theorem (Estimation error)

For any random variable X and estimator $\hat{X}$, we have $\mathrm{E}(X-\hat{X})^2 \geq \frac{1}{2\pi e} e^{2h(X)}$ with equality iif X is Gaussian with mean $\hat{X}$.

Proof:

$$
\begin{aligned}
\mathrm{E}(X-\hat{X})^2 &\geq \min_{\hat{X}} \mathrm{E}(X-\hat{X})^2 && (13)\\
&= \mathrm{E}(X-\mathrm{E}(X))^2 && (14)\\
&= \mathrm{Var}(X) && (15)\\
&\geq \frac{1}{2\pi e} e^{2h(X)} && (16)
\end{aligned}
$$

where (14) is because $\mathrm{E}(X)$ is the best estimator for X, and (16) follows from the fact that the normal distribution has the maximum entropy for a given variance.

Minimum Estimation Error with Side Information

Theorem (Estimation error with side info)

Given side info Y and estimator $\hat{X}(Y)$, it follows that $\mathrm{E}(X - \hat{X}(Y))^2 \geq \frac{1}{2\pi e} e^{2h(X|Y)}$.

Proof:

- We have

$$\mathrm{E}[(X - \hat{X})^2 | Y = y] \geq \mathrm{Var}(X | Y = y) \geq \frac{1}{2\pi e} e^{2h(X|Y=y)},$$

  where the second inequality follows from the fact that entropy of X conditioned on $Y = y$ is upper bounded by the entropy of Gaussian RV with the same variance.
- Take expectation (over Y) of both sides and apply Jensen's inequality yields the stated result.

Entropy Power

Definition

The entropy power of a random vector $\mathbf{x} \in \mathbb{R}^d$ with a density is defined as

$$N(\mathbf{x}) = \frac{1}{2\pi e} e^{\frac{2}{d} h(\mathbf{x})}.$$

- $h(a\mathbf{x}) = h(\mathbf{x}) + d \log |a|$ and $N(a\mathbf{x}) = a^2 N(\mathbf{x})$ for any constant $a \in \mathbb{R}$.
- For $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{K_x})$, $N(\mathbf{x}) = |\mathbf{K_x}|^{\frac{1}{d}}$ is the geometric mean of all eigenvalues of $\mathbf{K_x}$.
- $0 < N(\mathbf{x}) \leq |\mathbf{K_x}|^{\frac{1}{d}}$, where $\mathbf{K_x}$ is the covariance matrix of $\mathbf{x}$.
- $|\mathbf{K_x}|$ is referred to as generalized variance while $N(\mathbf{x})$ is the effective variance.
- Entropy power can be interpreted as a positive bounded measure of 'Gaussianity'.

Entropy Power Inequality (EPI)

Theorem (EPI: Entropy power is super-additive)

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ *be independent random vectors. Then,*

$$N(\mathbf{x}+\mathbf{y}) \geq N(\mathbf{x}) + N(\mathbf{y}), \text{ or equivalently, } e^{\frac{2}{d}h(\mathbf{x}+\mathbf{y})} \geq e^{\frac{2}{d}h(\mathbf{x})} + e^{\frac{2}{d}h(\mathbf{y})}.$$

Moreover, equality holds iif $\mathbf{x}$ *and* $\mathbf{y}$ *are multivariate Gaussian with proportional covariances (*$\mathbf{K}_{\mathbf{y}} = c\mathbf{K}_{\mathbf{x}}$ *for some constant* $c > 0$*).*

EPI Equivalent Statements[2]

Theorem (EPI: Sum of Gaussian RVs has the smallest entropy)

For any two independent random vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, we have

$$h(\mathbf{x} + \mathbf{y}) \geq h(\tilde{\mathbf{x}} + \tilde{\mathbf{y}}),$$

where $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$ are two independent Gaussian with proportional covariances, chosen so that $h(\tilde{\mathbf{x}}) = h(\mathbf{x})$ and $h(\tilde{\mathbf{y}}) = h(\mathbf{y})$.

Proof: $N(\tilde{\mathbf{x}} + \tilde{\mathbf{y}}) = N(\tilde{\mathbf{x}}) + N(\tilde{\mathbf{y}}) = N(\mathbf{x}) + N(\mathbf{y}) \leq N(\mathbf{x} + \mathbf{y})$.

Theorem (Concavity of entropy under the covariance preserving transformation)

For any $\lambda \in [0, 1]$, we have

$$h(\sqrt{\lambda}\mathbf{x} + \sqrt{1-\lambda}\mathbf{y}) \geq \lambda h(\mathbf{x}) + (1-\lambda) h(\mathbf{y}) \tag{17}$$

[2] A. Dembo, T. Cover and J. Thomas, "Information theoretic inequalities," in *IEEE Transactions on Information Theory*, 1991.

EPI Equivalent Statements (cont'd)[3]

Theorem (Equivalent EPIs)

For finitely many independent random vectors $\{\mathbf{x}_i\}_i$ *with finite differential entropies, and real-valued coefficients* $\{a_i\}_i$*, the following inequalities are equivalent*

$$N\left(\sum_i a_i \mathbf{x}_i\right) \geq \sum_i a_i^2 N\left(\mathbf{x}_i\right), \tag{18}$$

$$h\left(\sum_i a_i \mathbf{x}_i\right) \geq h\left(\sum_i a_i \tilde{\mathbf{x}}_i\right), \tag{19}$$

$$h\left(\sum_i a_i \mathbf{x}_i\right) \geq \sum_i a_i^2 h\left(\mathbf{x}_i\right) \quad \text{with} \quad \sum_i a_i^2 = 1, \tag{20}$$

where $\{\tilde{\mathbf{x}}_i\}$ *are independent Gaussian random vectors with proportional convariances and corresponding entropies* $h(\tilde{\mathbf{x}}_i) = h(\mathbf{x}_i)$*.*

[3] `https://arxiv.org/abs/0704.1751`

Proofs and Applications of EPI

Proofs: There are many techniques used for various proofs of EPI; e.g.,

- DPI, Sato's inequality, Fisher info inequality (FII)
- De Bruijn's identity (Thm 17.7.2), mutual info inequality (MII)
- Integration over a path of continuous Gaussian perturbation.

Applications:

1. Bounding channel capacity and rate-distortion regions.
2. Blind source separation.
3. Providing easy proofs of some inequalities.
4. Strengthening CLT (Andrew Barron-1986): Gaussianity increases on summing.
5. $\cdots$

Inequalities

Theorem (Minkowski's Inequality)

For any two nonnegative definite matrices $\mathbf{K}_1$ *and* $\mathbf{K}_2$*, we have*

$$|\mathbf{K}_1 + \mathbf{K}_2|^{\frac{1}{n}} \geq |\mathbf{K}_1|^{\frac{1}{n}} + |\mathbf{K}_2|^{\frac{1}{n}},$$

with equality iff $\mathbf{K}_1 = c\mathbf{K}_2$ *for some constant* c*.*

Proof: Let $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_i)$ be independent for $i = 1, 2$. Thus, $\mathbf{x}_1 + \mathbf{x}_2 \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_1 + \mathbf{K}_2)$,

$$\underbrace{|\mathbf{K}_1 + \mathbf{K}_2|^{\frac{1}{n}}}_{N(\mathbf{x}_1+\mathbf{x}_2)} \geq \underbrace{|\mathbf{K}_1|^{\frac{1}{n}} + |\mathbf{K}_2|^{\frac{1}{n}}}_{N(\mathbf{x}_1)+N(\mathbf{x}_2)}$$

is a direct result of EPI.

Theorem (Ky Fan's Inequality)

$\log|\mathbf{K}|$ *is concave in* $\mathbb{S}^d_{++}$*.*

Proof: Let $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_i)$ be independent for $i = 1, 2$. Thus, for any $\lambda \in [0, 1]$, $\sqrt{\lambda}\mathbf{x}_1 + \sqrt{1-\lambda}\mathbf{x}_2 \sim \mathcal{N}(\mathbf{0}, \lambda\mathbf{K}_1 + (1-\lambda)\mathbf{K}_2)$. Then, (17) becomes

$$\log|\lambda\mathbf{K}_1 + (1-\lambda)\mathbf{K}_2| \geq \lambda \log|\mathbf{K}_1| + (1-\lambda)\log|\mathbf{K}_2|.$$

Entropic Central Limit Theorem

Let $X_1, X_2, \cdots$ be i.i.d. RVs with mean μ and variance σ^2. Let

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu)$$

denote the normalized sum of the first n terms.

- Classical CLT: $Z_n \xrightarrow{d} \mathcal{N}(0, \sigma^2)$.
- Entropic CLT: $h(Z_n) \to h(\mathcal{N}(0, \sigma^2)) = \frac{1}{2} \log(2\pi e \sigma^2)$. Furthermore, if $\{X_i\}$ are non-Gaussian, then the sequence $\{h(Z_n)\}$ is strictly increasing:

$$h(X_1) = h(Z_1) < h(Z_2) < \cdots < h(Z_n) < \frac{1}{2} \log(2\pi e \sigma^2).$$

Thank You!

Email: `<zhangy@ucsc.edu>`

Homepage: `https://people.ucsc.edu/~yzhan419/`