A PROOF OF ENTROPY POWER INEQUALITY: LECTURE NOTES (WINTER SCHOOL 2017)

CHANDRA NAIR

1. Preliminaries

Let X be a random variable taking finite (or countably infinite) possible values with a probability mass function (p.m.f.) given by $p_X(x)$. Then entropy of X is defined as

$$H(p) := -\sum_x p_X(x) \log p_X(x) = -\operatorname{E}(\log p_X).$$

Entropy is a *measure of information* revealed upon knowing the realization of X. If the base of the logarithm is 2, then entropy is measured as bits; if it is natural logarithm, the unit is called nat. Entropy is a concave, non-negative function of the p.m.f. $p_X(x)$.

Remark: Due to an abuse of notation dating back many years in information theory, usually we express entropy as $H(X)$, though it is really a function of the p.m.f. rather than the realization of the random variable.

Mathematicians usually work with a related quantity called *relative entropy*. We say that a p.m.f. $p_X(x)$ is *absolutely continuous* with respect to another p.m.f. $q_X(x)$ if $q_X(x) = 0$ implies $p_X(x) = 0$, usually denoted as $p \ll q$. (This is a notion that extends naturally to arbitrary random variables). When $p \ll q$, then the relative entropy of $p_X(x)$ w.r.t. $q_X(x)$ is defined as

$$H(p|q) := \sum_x p_X(x) \log \frac{p_X(x)}{q_X(x)} = \operatorname{E}\left(\log \frac{p(X)}{q(X)}\right).$$

Jensen's inequality states that if $f(\cdot)$ is a concave function, then $\operatorname{E}(f(X)) \leq f(\operatorname{E}(X))$. Since $\log(x)$ is concave, we have

$$-H(p|q) = \operatorname{E}\left(\log \frac{q(X)}{p(X)}\right) \leq \log\left(\operatorname{E}\left(\frac{q(X)}{p(X)}\right)\right) = \log 1 = 0,$$

implying non-negativity of relative entropy. Further note that equality holds *if and only if* $q = p$.

Given a joint distribution $p_{X,Y}(x,y)$ the relative entropy between the joint distribution and product marginals $q_{X,Y} = p_X p_Y$ is called as *mutual information*, $I(X;Y)$. Thus

$$I(X;Y) := H(p_{X,Y}|p_X p_Y) = \sum_{x,y} p_{X,Y}(x,y) \log \frac{p_{X,Y}(x,y)}{p_X(x) p_Y(y)}.$$

This is *a measure of information that one random variable provides about another variable*. It is clearly a symmetric quantity. By the non-negativity of relative

Date: January 17, 2017.

entropy $I(X;Y) \geq 0$ and further equality holds *if and only if* X and Y are independent.

Exercise 1

(*a*) If a random variable X takes values in a finite set, say $\{1, 2, ..., m\}$, then show that $0 \leq H(X) \leq \log m$. (Hint: Consider a uniform distribution on the set, and take the relative entropy of p_X with respect to the uniform measure.)
(*b*) If X takes values in $\mathbb{N}$ and $\mathrm{E}(X) = \lambda$, determine the distribution that maximizes the entropy $H(p)$.

We sometimes consider conditional entropy, which is defined as follows:

$$H(X|Y) = -\sum_{x,y} p(x,y) \log p_{X|Y}(x|y) = -\mathrm{E}(\log p_{X|Y}).$$

Similarly define conditional mutual information according to

$$I(X;Y|Z) = \sum_{x,y,Z} p(x,y,z) \log \frac{p_{X,Y|Z}(x,y|z)}{p_{X|Z}(x|z) p_{Y|Z}(y|z)} = \mathrm{E}_Z \left(H(p_{X,Y|Z} | p_{X|Z} p_{Y|Z}) \right).$$

Note that $I(X;Y|Z) = 0$ if and only if X and Y are conditionally independent of Z.

Exercise 2

(*a*) $H(X,Y) = H(X) + H(Y|X)$.
(*b*) $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$.
(*c*) $I(X;Y,Z) = I(X;Z) + I(X;Y|Z)$.

1.1. **Data-processing inequality.** If we process data, we are bound to lose information. This is captured by data-processing inequality. Let $X \to Y \to Z$ be a Markov chain, i.e.$p(z|y,x) = p(z|y)$. In other words Z is some random transformation of Y, conditionally independent of X given Y. In other words, X and Z are conditionally independent of Y. Hence $I(X;Z|Y) = 0$. Thus

$$I(X;Y) = I(X;Y) + I(X;Z|Y) = I(X;Y,Z) = I(X;Z) + I(X;Y|Z) \geq I(X;Z).$$

Exercise 3

Let X_1 and X_2 be independent and identically distributed random variables (say taking values in $\mathbb{N}$ though this is immaterial). Let U be any random variable such that $U \to (X_1 + X_2) \to X_1$ is Markov. Then show that $I(U; X_1 + X_2) \geq 2I(U; X_1)$.

2. Entropy Power Inequality

Let X be a continuous random variable with a density $f_X(x)$. The differential entropy of X is defined as (when the integral is well-defined)

$$h(X) := \int_{-\infty}^{\infty} -f(x) \log f(x) dx = \mathrm{E}(-\log f(X)).$$

Note that notation is abused to denote $h(\mu_X)$ as $h(X)$.

For any two independent real-valued random variables X and Y we have (assume logarithms are to base 2)

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

In general if $\mathbf{X}$ and $\mathbf{Y}$ are d-dimensional independent random vectors then

$$2^{\frac{2}{d}h(\mathbf{X}+\mathbf{Y})} \geq 2^{\frac{2}{d}h(\mathbf{X})} + 2^{\frac{2}{d}h(\mathbf{Y})}.$$

This has many applications in information theory. It also implies *Minkowski's* inequality in geometry below.

Exercise 4

(1) Let the density of $\mathbf{X}$ be non-zero in a set A (of finite volume, sat $v(A)$). Show that $h(X) \leq \log v(A)$, and that equality is achieved when $\mathbf{X}$ is uniform on A.
(2) Show that $h(B\mathbf{X}) = h(\mathbf{X}) + \log|B|$.
(3) Prove the following inequality as a corollary of the entropy power inequality. Let A and B be two sets in $\mathbb{R}^d$, then

$$v(A+B)^{\frac{2}{d}} \geq v(A)^{\frac{2}{d}} + v(B)^{\frac{2}{d}}.$$

Here $A + B = \{\mathbf{z} : \mathbf{z} = \mathbf{x} + \mathbf{y}, \text{ for some } \mathbf{x} \in A, \mathbf{y} \in B\}$.

Gaussian random variables: These random variables occur naturally as the limit of various operations (for instance, the central limit theorem). The density of a d-dimensional Gaussian random variable with mean $\mathbf{m}$ and covariance K is given by

$$\mu_G(\mathbf{x}) = \frac{1}{(2\pi|K|)^{d/2}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T K^{-1}(\mathbf{x}-\mathbf{m})}.$$

The differential entropy of a the above Gaussian vector is given by

$$\begin{aligned} h(\mu_G) &= -\int \mu_G(\mathbf{x}) \log \mu_G(\mathbf{x}) d\mathbf{x} = \frac{d}{2}\log(2\pi|K|) + \frac{\log e}{2}\,\mathrm{E}\left((\mathbf{x}-\mathbf{m})^T K^{-1}(\mathbf{x}-\mathbf{m})\right) \\ &= \frac{d}{2}\log(2\pi|K|) + \frac{\log e}{2}\,\mathrm{E}\, tr\left(K^{-1}(\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^T\right) = \frac{d}{2}\log(2\pi|K|) + \frac{d\log e}{2} \\ &= \frac{d}{2}\log(2\pi e|K|). \end{aligned}$$

Gaussian random variables enjoy may properties:

- linear combinations of jointly Gaussian variables are Gaussian
- Sums of independent Gaussians is a Gaussian with mean and covariance given by sum of the means and sum of the covariances.
- If two Gaussian random variables are uncorrelated, then they are independent.

Exercise 5

(1) Let $\mathbf{X}$ be a random variable with density μ, that satisfy a covariance constraint $\mathrm{E}((\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^T) \preceq K$, then show that

$$h(\mu) \leq \frac{d}{2}\log(2\pi e|K|).$$

(2) If $\mathbf{X}$ and $\mathbf{Y}$ are Gaussians with proportional covariances, then equality holds in entropy power inequality.

2.1. **Proof of the entropy power inequality.**

Proposition 1. *The following two statements are equivalent:*

(*i*) $2^{\frac{2}{d}h(\mathbf{X}+\mathbf{Y})} \geq 2^{\frac{2}{d}h(\mathbf{X})} + 2^{\frac{2}{d}h(\mathbf{Y})}$. *holds for all continuous and independent X, Y;*

(*ii*) $h(\sqrt{\lambda}\mathbf{X} + \sqrt{1-\lambda}\mathbf{Y}) \geq \lambda h(\mathbf{X}) + (1-\lambda)h(\mathbf{Y})$ *holds for all continuous and independent X, Y and $\lambda \in [0,1]$.*

Proof. $(i) \implies (ii)$: From (i) we have

$$\begin{aligned} 2^{\frac{2}{d}h(\sqrt{\lambda}\mathbf{X}+\sqrt{1-\lambda}\mathbf{Y})} &\geq 2^{\frac{2}{d}h(\sqrt{\lambda}\mathbf{X})} + 2^{\frac{2}{d}h(\sqrt{1-\lambda}\mathbf{Y})} \\ &= \lambda 2^{\frac{2}{d}h(\mathbf{X})} + (1-\lambda)2^{\frac{2}{d}h(\mathbf{Y})} \\ &\geq 2^{\frac{2}{d}(\lambda h(\mathbf{X})+(1-\lambda)h(\mathbf{Y}))} \qquad \text{(convexity of } 2^x\text{)}. \end{aligned}$$

$(ii) \implies (i)$: Let

$$\lambda = \frac{2^{\frac{2}{d}h(\mathbf{X})}}{2^{\frac{2}{d}h(\mathbf{X})} + 2^{\frac{2}{d}h(\mathbf{Y})}}.$$

Note that (ii) implies

$$\begin{aligned} h(\mathbf{X}+\mathbf{Y}) &\geq \lambda h(\frac{\mathbf{X}}{\sqrt{\lambda}}) + (1-\lambda)h(\frac{\mathbf{Y}}{\sqrt{1-\lambda}}) \\ &= \lambda h(\mathbf{X}) - \frac{d\lambda}{2}\log(\lambda) + (1-\lambda)h(\mathbf{Y}) - \frac{d(1-\lambda)}{2}\log(1-\lambda) \\ &= \frac{d}{2}\log\left(2^{\frac{2}{d}h(\mathbf{X})} + 2^{\frac{2}{d}h(\mathbf{Y})}\right). \end{aligned}$$

□

Hence we will prove the equivalent inequality that

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

holds for all continuous and independent random variables $\mathbf{X}$ and $\mathbf{Y}$ with finite differential entropies. In fact, this inequality is dimension independent.

Idea of the proof. : We know that equality holds when $\mathbf{X}, \mathbf{Y} \sim \mathcal{N}(0, I)$.

We create a *path in the space of distributions* defined by

$$(\mathbf{X}_t, \mathbf{Y}_t) \stackrel{d}{=} (\sqrt{t}\mathbf{X} + \sqrt{1-t}\mathbf{Z}_1, \sqrt{t}\mathbf{Y} + \sqrt{1-t}\mathbf{Z}_2),$$

where $\mathbf{Z}_1, \mathbf{Z}_2$ are two independent Gaussian variables distributed as $\mathcal{N}(0, I)$. Define a function from $[0,1] \mapsto \mathbb{R}$ according to:

$$f(t) := h(\sqrt{\lambda}\mathbf{X}_t + \sqrt{1-\lambda}\mathbf{Y}_t) - \lambda h(\mathbf{X}_t) - (1-\lambda)h(\mathbf{Y}_t).$$

We know $f(0) = 0$, and we would like to show $f(1) \geq 0$. This is accomplished by showing $f'(t) \geq 0$.

2.1.1. *Derivative of differential entropy along Gaussian perturbation.* Define[1]

$$J(\mathbf{X}) = \frac{d}{ds} h(\mathbf{X} + \sqrt{s}\mathbf{Z})|_{s\to 0+}$$

where $\mathbf{Z} \sim N(0, I)$ is independent of $\mathbf{X}$.

Lemma 1. *Let $\mathbf{X}$ is a continuous random variable with a density and $\mathbf{Z} \sim \mathcal{N}(0, I)$ be independent of $\mathbf{X}$. We show that $J(\cdot)$ satisfies the following:*

(i) $J(\mathbf{X} + \sqrt{s}\mathbf{Z}) = \frac{d}{ds} h(\mathbf{X} + \sqrt{s}\mathbf{Z})$, *when* $s > 0$.
(ii) $J(a\mathbf{X}) = \frac{1}{a^2} J(\mathbf{X})$.
(iii) $\frac{d}{dt} h(\sqrt{t}\mathbf{X} + \sqrt{1-t}\mathbf{Z}) = -\frac{1}{t} J(\sqrt{t}\mathbf{X} + \sqrt{1-t}\mathbf{Z}) + \frac{d}{2t\ln 2}$.

Proof. (i): Let $\mathbf{Z}_1 \sim \mathcal{N}(0, I)$ be independent of $\mathbf{X}$ and $\mathbf{Z}$.

$$\begin{aligned}
\frac{d}{ds} h(\mathbf{X} + \sqrt{s}\mathbf{Z}) &= \lim_{\delta\to 0} \frac{h(\mathbf{X} + \sqrt{s+\delta}\mathbf{Z}) - h(\mathbf{X} + \sqrt{s}\mathbf{Z})}{\delta} \\
&= \lim_{\delta\to 0} \frac{h(\mathbf{X} + \sqrt{s}\mathbf{Z} + \sqrt{\delta}\mathbf{Z}_1) - h(\mathbf{X} + \sqrt{s}\mathbf{Z})}{\delta} \\
&= \frac{d}{dt} h(\mathbf{X} + \sqrt{s}\mathbf{Z} + \sqrt{t}\mathbf{Z}_1)\Big|_{t=0} \\
&= J(\mathbf{X} + \sqrt{s}\mathbf{Z}) \qquad \square
\end{aligned}$$

(ii): W.l.o.g. $a > 0$ and let $u = \frac{s}{a^2}$. Observe that

$$\begin{aligned}
J(a\mathbf{X}) &= \frac{d}{ds} h(a\mathbf{X} + \sqrt{s}\mathbf{Z})|_{s=0} = \frac{d}{ds}\big(h(\mathbf{X} + \frac{1}{a}\sqrt{s}\mathbf{Z}) + d\log_2 a\big)|_{s=0} \\
&= \frac{1}{a^2}\frac{d}{du} h(\mathbf{X} + \sqrt{u}\mathbf{Z})|_{s=0} = \frac{1}{a^2} J(\mathbf{X}). \quad \square
\end{aligned}$$

(iii): Let $s = \frac{1-t}{t}$. Note that $\frac{ds}{dt} = -\frac{1}{t^2}$. Observe that

$$\begin{aligned}
\frac{d}{dt} h(\mathbf{X}\sqrt{t} + \sqrt{1-t}\mathbf{Z}) &= \frac{d}{dt}\big(h(\mathbf{X} + \sqrt{\frac{1-t}{t}}\mathbf{Z}) + d/2\log_2 t\big) \\
&= -\frac{1}{t^2}\frac{d}{ds} h(\mathbf{X} + \sqrt{s}\mathbf{Z}) + \frac{d}{2t\ln 2} \\
&\overset{(a)}{=} -\frac{1}{t^2} J(\mathbf{X} + \sqrt{s}\mathbf{Z}) + \frac{d}{2t\ln 2} \\
&= -\frac{1}{t^2} J(\mathbf{X} + \sqrt{\frac{1-t}{t}}\mathbf{Z}) + \frac{d}{2t\ln 2} \\
&\overset{(a)}{=} -\frac{1}{t} J(\sqrt{t}\mathbf{X} + \sqrt{1-t}\mathbf{Z}) + \frac{d}{2t\ln 2},
\end{aligned}$$

where (a) follows from part (i) and (b) follows from part (ii).

We apply the results of the above Lemma to obtain the following:

$$\begin{aligned}
f'(t) &= \frac{d}{dt}\Big(h(\sqrt{\lambda}\mathbf{X}_t + \sqrt{1-\lambda}\mathbf{Y}_t) - \lambda h(\mathbf{X}_t) - (1-\lambda)h(\mathbf{Y}_t)\Big) \\
&= -\frac{1}{t}\Big(J(\sqrt{\lambda}\mathbf{X}_t + \sqrt{1-\lambda}\mathbf{Y}_t) - \lambda J(\mathbf{X}_t) - (1-\lambda)J(\mathbf{Y}_t)\Big)
\end{aligned}$$

[1] A scaled version of $J(X)$ is called Fisher information.

This, if we show that when $\mathbf{X}$ and $\mathbf{Y}$ be independent continuous random variables and for $\lambda \in (0,1)$ we have

$$J(\sqrt{\lambda}\mathbf{X} + \sqrt{1-\lambda}\mathbf{Y}) \leq \lambda J(\mathbf{X}) + (1-\lambda)J(\mathbf{Y}),$$

then we are done. (Looks similar to before, but turns out to have a rather simple proof.)

Proposition 2. *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be independent continuous random variables. Let* $\lambda \in (0,1)$*. We have*

$$J(\sqrt{\lambda}\mathbf{X} + \sqrt{1-\lambda}\mathbf{Y}) \leq \lambda J(\mathbf{X}) + (1-\lambda)J(\mathbf{Y}).$$

Proof. Let $\mathbf{Z} \sim \mathcal{N}(0, I)$ be independent of $\mathbf{X}, \mathbf{Y}$. Define $\mathbf{X}_\tau = \mathbf{X} + \sqrt{\lambda\tau}\mathbf{Z}, \mathbf{Y}_\tau = \mathbf{Y} + \sqrt{(1-\lambda)\tau}\mathbf{Z}$. Note that we have the following two Markov chains:

- $\mathbf{Z} \to (\mathbf{X}_\tau, \mathbf{Y}_\tau) \to \sqrt{\lambda}\mathbf{X}_\tau + \sqrt{1-\lambda}\mathbf{Y}_\tau$,
- $\mathbf{X}_\tau \to \mathbf{Z} \to \mathbf{Y}_\tau$.

By data processing inequality, we have

$$\begin{aligned} I(\mathbf{Z}; \sqrt{\lambda}\mathbf{X}_\tau + \sqrt{1-\lambda}\mathbf{Y}_\tau) &\leq I(\mathbf{Z}; \mathbf{X}_\tau, \mathbf{Y}_\tau) \\ &\leq I(\mathbf{Z}; \mathbf{X}_\tau) + I(\mathbf{X}_\tau, \mathbf{Z}; \mathbf{Y}_\tau) \\ &= I(\mathbf{Z}; \mathbf{X}_\tau) + I(\mathbf{Z}; \mathbf{Y}_\tau). \end{aligned}$$

Define

$$g(\tau) = I(\mathbf{Z}; \mathbf{X}_\tau) + I(\mathbf{Z}; \mathbf{Y}_\tau) - I(\mathbf{Z}; \sqrt{\lambda}\mathbf{X}_\tau + \sqrt{1-\lambda}\mathbf{Y}_\tau)).$$

Note that $g(0) = 0$ and $g(\tau) \geq 0$ for $\tau \geq 0$. Hence $g'(0) \geq 0$.

Observe that

$$\frac{d}{d\tau} I(\mathbf{Z}; \mathbf{X}_\tau) = \frac{d}{d\tau}(h(\mathbf{X}_\tau) - h(\mathbf{X})) = \lambda J(\mathbf{X} + \sqrt{\lambda\tau}\mathbf{Z}).$$

In a similar fashion

$$\frac{d}{d\tau} I(\mathbf{Z}; \mathbf{Y}_\tau) = (1-\lambda)J(\mathbf{X} + \sqrt{(1-\lambda)\tau}\mathbf{Z}),$$

$$\frac{d}{d\tau} I(\mathbf{Z}; \sqrt{\lambda}\mathbf{X}_\tau + \sqrt{1-\lambda}\mathbf{Y}_\tau) = \frac{d}{d\tau} I(\mathbf{Z}; \sqrt{\lambda}\mathbf{X} + \sqrt{1-\lambda}\mathbf{Y} + \sqrt{\tau}\mathbf{Z}) = J(\sqrt{\lambda}\mathbf{X} + \sqrt{1-\lambda}\mathbf{Y}).$$

Substituting the above into $g'(0) \geq 0$ yields the proposition. $\square$