Rutgers University
The State University Of New Jersey
College of Engineering
Department of Electrical and Computer Engineering

330:542 | Information Theory and Coding | Spring 1996

Problem Set 2 Solutions

1. C&T problem 2.12

$$\begin{aligned} I(X;Y;Z) &= I(X;Y) - I(X;Y|Z) \\ &= I(X;Y) - I(X;Y,Z) + I(X;Z) \\ &= I(X;Y) - H(X) - H(Y,Z) + H(X,Y,Z) + I(X;Z) \\ &= I(X;Y) - H(X) - H(Y) - H(Z) + I(Y;Z) + H(X,Y,Z) + I(X;Z) \\ &= I(X;Y) + I(Y;Z) + I(X;Z) + H(X,Y,Z) - H(X) - H(Y) - H(Z) \end{aligned}$$

To show symmetry we can use identities like $I(X;Y) = H(X) + H(Y) - H(X,Y)$ As for negativity, just choose random variables X and Y independent but NOT conditionally independent; i.e., $p_{XY}(x,y) = p_X(x)p_Y(y)$ but $p_{XY|Z}(x,y|z) \neq p_X(x|z)p_Y(y|z)$ to get $I(X;Y) = 0$ and $I(X;Y|Z)$ nonzero.

2. C&T problem 2.15

   (a) Let $\rho(X,Y) = H(X|Y) + H(Y|X)$. Then since conditional entropy ≥ 0 we have $\rho(X,Y) \geq 0$. Obviously $\rho(X,Y) = \rho(Y,X)$.
   Now note that $H(X|Y)$ is zero iff X is a function of Y. Likewise $H(Y|X)$ is zero iff Y is a function of X. Therefore, $\rho(X,Y) = 0$ iff $X = g(Y)$ and $Y = f(X)$ for all X, Y pairs. This corresponds to an invertible (one-to-one) mapping.
   As for the triangle inequality, note that

$$\begin{aligned} H(X|Y) + H(Y|X) &\geq H(X|Y,Z) + H(Y|Z) \\ &= H(X,Y|Z) \\ &= H(X|Z) + H(Y|X,Z) \\ &\geq H(X|Z) \end{aligned}$$

   Thus, $\rho(X,Y) + \rho(Y,Z) \geq H(X|Z) + H(Z|X)$ This inequality is strict unless $X \to Y \to Z$ forms a Markov chain and Y is a function of X and Z.
   In total we then have, if X and Y are "equivalent" $\rho()$ is a metric (distance measure). (hooray!)

   (b) $H(X|Y) = H(X) - I(X;Y)$ gives $\rho(X,Y) = H(X) + H(Y) - 2(IX;Y)$ The second relation follows from the $H(X,Y) = H(X) + H(Y) - I(X;Y)$. The third follows after substituting $I(X;Y) = H(X) + H(Y) - H(X;Y)$.

3. C&T problem 2.29

   (a) Using the chain rule for conditional entropy we have $H(X,Y|Z) = H(X|Z) + H(Y|X,Z) \geq H(X|Z)$ with equality iff $H(Y|X,Z) = 0$ (i.e., Y function of X and/or Z).

   (b) Chain rule yields $I(X,Y;Z) = I(X;Z) + I(Y;Z|X) \geq I(X;Z)$ with equality iff $I(Y;Z|X) = 0$.

(c) Chain rule for entropy then definition of conditional mutual info yields

$$H(X,Y,Z)-H(X,Y) = H(Z|X,Y) = H(Z|X)-I(Y;Z|X) \le H(Z|X) = H(X,Z)-H(X)$$

(d) Chain rule: $I(X;Z|Y) + I(Z;Y) = I(X,Y;Z) = I(Z;Y|X) + I(X;Z)$ and therefore $I(X;Z|Y) = I(Z;Y|X) - I(Z;Y) + I(X;Z)$.

4. Convexity and all that:

(a) Convexity of $g(z)$ implies $g(\lambda z_1 + (1-\lambda)z_2) \le \lambda g(z_1) + (1-\lambda)g(z_2)$. In our case $g(z) = f(r(z), s(z))$ so that we need to show that

$$f(r(\lambda z_1 + (1-\lambda)z_2), s(\lambda z_1 + (1-\lambda)z_2)) \overset{?}{\le} \lambda f(r(z_1), s(z_1)) + (1-\lambda) f(r(z_2), s(z_2)).$$

Well invoke the linearity of r and s to get $r(\lambda z_1 + (1-\lambda)z_2) = \lambda r(z_1) + (1-\lambda)r(z_2)$ and $s(\lambda z_1 + (1-\lambda)z_2) = \lambda s(z_1) + (1-\lambda)s(z_2)$ and the left side of the inequality becomes $f(r(\lambda z_1 + (1-\lambda)z_2), s(\lambda z_1 + (1-\lambda)z_2))$ QED.

(b) $f(x,y)$ nondecreasing implies $f(x+\delta, y) \ge f(x,y)$ and $f(x, y+\delta) \ge f(x,y)$ for $\delta > 0$. We have $r(\lambda z_1 + (1-\lambda)z_2) \le \lambda r(z_1) + (1-\lambda)r(z_2)$ and the same for $s()$. This yields

$$\begin{aligned} f(r(\lambda z_1 + (1-\lambda)z_2), s(\lambda z_1 + (1-\lambda)z_2)) &\le f(\lambda r(z_1) + (1-\lambda)r(z_2), \lambda s(z_1) + (1-\lambda)s(z_2)) \\ &\le \lambda f(r(z_1), s(z_1)) + (1-\lambda) f(r(z_2), s(z_2)) \end{aligned}$$

(c) Same as (b) only inequalities reversed.

(d) Convexity means that for any two points $\underline{x}_1$ and $\underline{x}_2$ we have

$$f(\lambda \underline{x}_1 (1-\lambda)\underline{x}_2) \le \lambda f(\underline{x}_1) + (1-\lambda) f(\underline{x}_2)$$

where $\underline{x}_1$ and $\underline{x}_2$ are multidimensional vectors. This implies that along any LINE $\underline{x} = \lambda \underline{x}_1 + (1-\lambda)\underline{x}_2$ we must have $f(\lambda \underline{x}_1 (1-\lambda)\underline{x}_2)$ convex in the scalar λ (with $0 \le \lambda \le 1$). In order for $f()$ to be convex in λ we must have $\partial^2 f/\partial \lambda^2 \ge 0$. We then have

$$\frac{\partial f}{\partial \lambda} = \sum_{i=1}^{n} (x_{1,i} - x_{2,i}) \frac{\partial f}{\partial x_i}$$

and

$$\frac{\partial^2 f}{\partial \lambda^2} = \sum_{i,j=1}^{n} (x_{1,i} - x_{2,i})(x_{1,j} - x_{2,j}) \frac{\partial^2 f}{\partial x_i x_j} = \sum_{i=1}^{n} (x_{1,i} - x_{2,i}) \sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_i x_j} (x_{1,j} - x_{2,j})$$

for all $\underline{x}_1, \underline{x}_2 \in \Re^n$.

The linear algebraists among you will recognize this sum as a quadratic form $\mathbf{x}^T \mathbf{F} \mathbf{x}$ where $\mathbf{x} \in \Re^n$ and $\mathbf{F}$ is the matrix of cross partials. Thus, the condition for convexity in terms of the partials is that $(\underline{x}_1 - \underline{x}_2)^T \mathbf{F} (\underline{x}_1 - \underline{x}_2) \ge 0$ for all $\underline{x}_1, \underline{x}_2 \in \Re^n$. In other words the matrix $\mathbf{F}$ must be positive semidefinite everywhere.

This is a difficult condition to check in general.

However, here's a simple example. Let $f(x,y) = x^2 + y^2 - axy$ where $a > 0$. The partial matrix is everywhere

$$\mathbf{F} = \begin{bmatrix} 2 & a \\ a & 2 \end{bmatrix}$$

The characteristic equation is $(2-s)^2 - a^2$ so that the eigenvalues are $s = 2 \pm a$. Thus, if $a < -2$ we will have negative roots and the matrix will not be positive semidefinite.

This can be checked by noting that along the line $x = y$ the surface will be concave if $a < -2$; i.e., $f(x, y = x) = (2 + a)x^2 = (2 + a)y^2$, which is concave for $a < -2$.
The other moral of this problem is that NON-NEGATIVE SECOND PARTIALS DO NOT GUARANTEE CONVEXITY!!!!!!.

(e) Once again we must use first principles. First notice that for a "linear" function as defined in this problem $\underline{q}(\underline{z}) = A\underline{z} + \underline{c}$ where A is a matrix and $\underline{c}$ is a constant vector. Notice that the constant ruins the usual definition of linearity. However,

$$\begin{aligned} \underline{q}(\lambda \underline{z}_1 + (1 - \lambda \underline{z}_2)) &= \lambda A \underline{z}_1 + \lambda \underline{c} + (1 - \lambda) A \underline{z}_2 + (1 - \lambda)\underline{c} \\ &= \lambda A \underline{z}_1 + (1 - \lambda)\underline{z}_2 + \underline{c} \\ &= \lambda \underline{q}(\underline{z}_1) + (1 - \lambda)\underline{q}(\underline{z}_2) \end{aligned}$$

So if we're given

$$f(\underline{x}_1 + (1 - \lambda)\underline{x}_2, \underline{y}_1 + (1 - \lambda)\underline{y}_2) \leq \lambda f(\underline{x}_1, \underline{y}_1) + (1 - \lambda) f(\underline{x}_2, \underline{y}_2).$$

then

$$\begin{aligned} f(r(\lambda \underline{z}_1 + (1 - \lambda)\underline{z}_2), s(\lambda \underline{z}_1 + (1 - \lambda)\underline{z}_2)) &= f(\lambda r(\underline{z}_1) + (1 - \lambda) r(\underline{z}_2), \lambda s(\underline{z}_1) + (1 - \lambda) s(\underline{z}_2)) \\ &\leq \lambda f(r(\underline{z}_1), s(\underline{z}_1)) + (1 - \lambda) f(r(\underline{z}_2), s(\underline{z}_2)) \\ &= \lambda f(\underline{z}_1) + (1 - \lambda) f(\underline{z}_2) \end{aligned}$$

so that $f()$ is also convex in $\underline{z}$.

5. (a) $\sum t_{ij} = 1$ and $t_{ij} > 0$ so it's a joint probability function. c_j and r_i are the marginals.

$$\begin{aligned} \Pi(T) &= \textstyle\sum_{ij} t_{ij} \log \frac{1}{t_{ij}} + \sum_{ij} t_{ij} \log \frac{1}{t_{ij}} + \sum_{ij} t_{ij} \log c_j + \sum_{ij} t_{ij} \log r_i \\ &= 2 \textstyle\sum_{ij} t_{ij} \log \frac{1}{t_{ij}} + \sum_{ij} t_{ij} \log c_j + \sum_{ij} t_{ij} \log r_i \\ &= 2H(XY) - H(Y) - H(X) \\ &= (H(XY) - H(X)) + (H(XY) - H(Y)) \end{aligned}$$

and the terms in parenthese are both greater than or equal to zero. So, $\Pi(T) \geq 0$ and $\Pi(T) = 0$ iff the mapping from X to Y is deterministic (fixed routing).

(b) $\Pi(T) = H(XY) - H(X) + H(XY) - H(Y)$ and since $H(XY) \leq H(X) + H(Y)$ we have $\Pi(T) \leq H(Y) + H(X) \leq \log M + \log N$. If X and Y independent and uniform, the bound is acheived; i.e., $t_{ij} = 1/MN$.

(c) Define the marginals, $U = \sum_{jk} d_{ijk}$, $V = \sum_{ik} d_{ijk}$, $W = \sum_{ij} d_{ijk}$. Then

$$\begin{aligned} \Pi(T) &= 2H(UW) - H(U) - H(W) \\ \Pi(T) &= 2H(UV) - H(U) - H(V) \\ \Pi(T) &= 2H(VW) - H(V) - H(W) \end{aligned}$$

So, is $\Pi(X) + \Pi(Y)$ less than or greater than $\Pi(T)$?

$$2H(UV) + 2H(VW) - H(U) - 2H(V) - H(W) \geq 2H(UW) - H(U) - H(W)$$

Simplifying,

$$H(UV) + H(VW) - H(V) \geq H(UW)$$

Now note that $H(UV) + H(VW) \geq H(UVW + H(V)$ since $H(UVW) - H(UV) = H(W|UV) \leq H(W|V) = H(WV) - H(V)$ by conditioning. Rearranging yields $H(UVW) - H(V) \leq H(WV) + H(UV)$. Thus,

$$H(UVW) \geq H(UW)$$

thereby proving the theorem.