2.1 Divergence: main inequality

Theorem 2.1 (Information Inequality)**.**

$$D(P\|Q) \ge 0\ ; \quad D(P\|Q) = 0 \quad \textit{iff } P = Q$$

Proof. Let $\varphi(x) \triangleq x\log x$, which is strictly convex, and use Jensen's Inequality:

$$D(P\|Q) = \sum_{\mathcal{X}} P(x)\log\frac{P(x)}{Q(x)} = \sum_{\mathcal{X}} Q(x)\varphi\left(\frac{P(x)}{Q(x)}\right) \ge \varphi\left(\sum_{\mathcal{X}} Q(x)\frac{P(x)}{Q(x)}\right) = \varphi(1) = 0$$

□

2.2 Conditional divergence

The main objects in our course are random variables. The main operation for creating new random variables, and also for defining relations between random variables, is that of a random transformation:

Definition 2.1. Conditional probability distribution (aka random transformation, transition probability kernel, Markov kernel, channel) $K(\cdot|\cdot)$ has two arguments: first argument is a measurable subset of $\mathcal{Y}$, second argument is an element of $\mathcal{X}$. It must satisfy:

1. For any $x \in \mathcal{X}$: $K(\,\cdot\,|x)$ is a probability measure on $\mathcal{Y}$
2. For any measurable A function $x \mapsto K(A|x)$ is measurable on $\mathcal{X}$.

In this case we will say that K acts from $\mathcal{X}$ to $\mathcal{Y}$. In fact, we will abuse notation and write $P_{Y|X}$ instead of K to suggest what spaces $\mathcal{X}$ and $\mathcal{Y}$ are[1]. Furthermore, if X and Y are connected by the random transformation $P_{Y|X}$ we will write $X \xrightarrow{P_{Y|X}} Y$.

Remark 2.1. (Very technical!) Unfortunately, condition 2 (standard for probability textbooks) will frequently not be sufficiently strong for this course. The main reason is that we want Radon-Nikodym derivatives such as $\frac{dP_{Y|X=x}}{dQ_Y}(y)$ to be jointly measurable in (x,y). See Section **??** for more.

Example:

1. deterministic system: $Y = f(X) \Leftrightarrow P_{Y|X=x} = \delta_{f(x)}$
2. decoupled system: $Y \perp\!\!\!\perp X \Leftrightarrow P_{Y|X=x} = P_Y$

[1] Another reason for writing $P_{Y|X}$ is that from any joint distribution $P_{X,Y}$ (on standard Borel spaces) one can extract a random transformation by conditioning on X.

3. additive noise (convolution): $Y = X + Z$ with $Z \perp\!\!\!\perp X \Leftrightarrow P_{Y|X=x} = P_{x+Z}$.

Multiplication:

$$\begin{tikzcd} {} \arrow[dr, "P_X"] & & \\ & X \arrow[r, "P_{Y|X}"] & Y \end{tikzcd}$$

to get $P_{XY} = P_X P_{Y|X}$:

$$P_{XY}(x,y) = P_{Y|X}(y|x)P_X(x)\,.$$

Composition (Marginalization): $P_Y = P_{Y|X} \circ P_X$, that is $P_{Y|X}$ acts on P_X to produce P_Y:

$$P_Y(y) = \sum_{x \in \mathcal{X}} P_{Y|X}(y|x)P_X(x)\,.$$

Will also write $P_X \xrightarrow{P_{Y|X}} P_Y$.

Definition 2.2 (Conditional divergence)**.**

$$\begin{aligned} D(P_{Y|X}\|Q_{Y|X}|P_X) &= \mathbb{E}_{x\sim P_X}[D(P_{Y|X=x}\|Q_{Y|X=x})] && (2.1)\\ &= \sum_{x\in\mathcal{X}} P_X(x)D(P_{Y|X=x}\|Q_{Y|X=x})\,. && (2.2)\end{aligned}$$

Note: $H(X|Y) = \log|\mathcal{A}| - D(P_{X|Y}\|U_X|P_Y)$, where U_X is is uniform distribution on $\mathcal{X}$.

Theorem 2.2 (Properties of Divergence)**.**

1. $D(P_{Y|X}\|Q_{Y|X}|P_X) = D(P_X P_{Y|X}\|P_X Q_{Y|X})$
2. *(Simple chain rule)* $D(P_{XY}\|Q_{XY}) = D(P_{Y|X}\|Q_{Y|X}|P_X) + D(P_X\|Q_X)$
3. *(Monotonicity)* $D(P_{XY}\|Q_{XY}) \ge D(P_Y\|Q_Y)$
4. *(Full chain rule)*

$$D(P_{X_1\cdots X_n}\|Q_{X_1\cdots X_n}) = \sum_{i=1}^n D(P_{X_i|X^{i-1}}\|Q_{X_i|X^{i-1}}|P_{X^{i-1}})$$

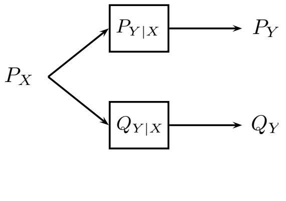
In the special case of $Q_{X^n} = \prod_i Q_{X_i}$ *we have*

$$D(P_{X_1\cdots X_n}\|Q_{X_1}\cdots Q_{X_n}) = D(P_{X_1\cdots X_n}\|P_{X_1}\cdots P_{X_n}) + \sum D(P_{X_i}\|Q_{X_i})$$

5. *(**Conditioning increases divergence**) Let* $P_{Y|X}$ *and* $Q_{Y|X}$ *be two kernels, let* $P_Y = P_{Y|X}\circ P_X$ *and* $Q_Y = Q_{Y|X}\circ P_X$*. Then*

$$\begin{aligned} D(P_Y\|Q_Y) &\le D(P_{Y|X}\|Q_{Y|X}|P_X)\\ &\quad \textit{equality iff } D(P_{X|Y}\|Q_{X|Y}|P_Y) = 0\end{aligned}$$

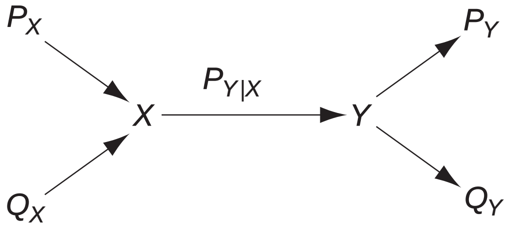
Pictorially:



6. *(**Data-processing for divergences**) Let $P_Y = P_{Y|X} \circ P_X$*

$$\left.\begin{array}{rcl} P_Y & = & \int P_{Y|X}(\,\cdot\,|x) dP_X \\ Q_Y & = & \int P_{Y|X}(\,\cdot\,|x) dQ_X \end{array}\right\} \Longrightarrow D(P_Y\|Q_Y) \le D(P_X\|Q_X) \tag{2.3}$$

Pictorially:



$$\Longrightarrow D(P_X\|Q_X) \ge D(P_Y\|Q_Y)$$

Proof. We only illustrate these results for the case of finite alphabets. General case follows by doing a careful analysis of Radon-Nikodym derivatives, introduction of regular branches of conditional probability etc. For certain cases (e.g. separable metric spaces), however, we can simply discretize alphabets and take granularity of discretization to 0. This method will become clearer in Lecture 4, once we understand continuity of D.

1. $\mathbb{E}_{x\sim P_X}[D(P_{Y|X=x}\|Q_{Y|X=x})] = \mathbb{E}_{(X,Y)\sim P_X P_{Y|X}}\left[\log \frac{P_{Y|X}}{Q_{Y|X}}\frac{P_X}{P_X}\right]$

2. Disintegration: $\mathbb{E}_{(X,Y)}\left[\log \frac{P_{XY}}{Q_{XY}}\right] = \mathbb{E}_{(X,Y)}\left[\log \frac{P_{Y|X}}{Q_{Y|X}} + \log \frac{P_X}{Q_X}\right]$

3. Apply 2. with X and Y interchanged and use $D(\cdot\|\cdot) \ge 0$.

4. Telescoping $P_{X^n} = \prod_{i=1}^n P_{X_i|X^{i-1}}$ and $Q_{X^n} = \prod_{i=1}^n Q_{X_i|X^{i-1}}$.

5. Inequality follows from monotonicity. To get conditions for equality, notice that by the chain rule for D:

$$\begin{aligned} D(P_{XY}\|Q_{XY}) &= D(P_{Y|X}\|Q_{Y|X}|P_X) + \underbrace{D(P_X\|P_X)}_{=0} \\ &= D(P_{X|Y}\|Q_{X|Y}|P_Y) + D(P_Y\|Q_Y) \end{aligned}$$

   and hence we get the claimed result from positivity of D.

6. This again follows from monotonicity. □

Corollary 2.1.

$$\begin{array}{rcl} D(P_{X_1\cdots X_n}\|Q_{X_1}\cdots Q_{X_n}) & \ge & \sum D(P_{X_i}\|Q_{X_i}) \quad \textit{or} \\ & = & \textit{iff } P_{X^n} = \prod_{j=1}^n P_{X_j} \end{array}$$

Note: In general we can have $D(P_{XY}\|Q_{XY}) \lessgtr D(P_X\|Q_X) + D(P_Y\|Q_Y)$. For example, if $X = Y$ under P and Q, then $D(P_{XY}\|D(Q_{XY}) = D(P_X\|Q_X) < 2D(P_X\|Q_X)$. Conversely, if $P_X = Q_X$ and $P_Y = Q_Y$ but $P_{XY} \ne Q_{XY}$ we have $D(P_{XY}\|Q_{XY}) > 0 = D(P_X\|Q_X) + D(P_Y\|Q_Y)$.

Corollary 2.2. *$Y = f(X) \Rightarrow D(P_Y\|Q_Y) \le D(P_X\|Q_X)$, with equality if f is 1-1.*

Note: $D(P_Y \| Q_Y) = D(P_X \| Q_X) \not\Rightarrow f$ is 1-1. Example: P_X = Gaussian, Q_X = Laplace, $Y = |X|$.

Corollary 2.3 (Large deviations estimate)**.** *For any subset $E \subset \mathcal{X}$ we have*

$$d(P_X[E] \| Q_X[E]) \le D(P_X \| Q_X)$$

Proof. Consider $Y = \mathbf{1}_{\{X \in E\}}$. □

2.3 Mutual information

Definition 2.3 (Mutual information)**.**

$$I(X;Y) = D(P_{XY} \| P_X P_Y)$$

Note:

- Intuition: $I(X;Y)$ measures the dependence between X and Y, or, the information about X (resp. Y) provided by Y (resp. X)
- Defined by Shannon (in a different form), in this form by Fano.
- Note: not restricted to discrete.
- $I(X;Y)$ is a functional of the joint distribution P_{XY}, or equivalently, the pair $(P_X, P_{Y|X})$.

Theorem 2.3 (Properties of I)**.**

1. $I(X;Y) = D(P_{XY} \| P_X P_Y) = D(P_{Y|X} \| P_Y | P_X) = D(P_{X|Y} \| P_X | P_Y)$
2. *Symmetry:* $I(X;Y) = I(Y;X)$
3. *Positivity:* $I(X;Y) \ge 0$; $I(X;Y) = 0$ *iff* $X \perp\!\!\!\perp Y$
4. $I(f(X);Y) \le I(X;Y)$; f *one-to-one* $\Rightarrow I(f(X);Y) = I(X;Y)$
5. *"More data $\Rightarrow$ More info":* $I(X_1, X_2; Z) \ge I(X_1; Z)$

Proof.

1. $I(X;Y) = \mathbb{E} \log \frac{P_{XY}}{P_X P_Y} = \mathbb{E} \log \frac{P_{Y|X}}{P_Y} = \mathbb{E} \log \frac{P_{X|Y}}{P_X}$.
2. Apply data-processing inequality twice to the map $(x,y) \to (y,x)$ to get $D(P_{X,Y} \| P_X P_Y) = D(P_{Y,X} \| P_Y P_X)$.
3. By definition.
4. We will use the data-processing property of mutual information (to be proved shortly, see Theorem 2.5). Consider the chain of data processing: $(x,y) \mapsto (f(x), y) \mapsto (f^{-1}(f(x)), y)$. Then
   $I(X;Y) \ge I(f(X);Y) \ge I(f^{-1}(f(X));Y) = I(X;Y)$
5. Consider $f(X_1, X_2) = X_1$. □

Theorem 2.4 (I v.s. H)**.**

1. *$I(X;X) = \begin{cases} H(X) & X \text{ discrete} \\ +\infty & \text{otherwise} \end{cases}$*

2. *If X, Y discrete then*
$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$
If only X discrete then
$$I(X;Y) = H(X) - H(X|Y)$$

3. *If X, Y are real-valued vectors, have joint pdf and all three differential entropies are finite then*
$$I(X;Y) = h(X) + h(Y) - h(X,Y)$$
If X has marginal pdf p_X and conditional pdf $p_{X|Y}(x|y)$ then
$$I(X;Y) = h(X) - h(X|Y)\,.$$

4. *If X or Y are discrete then $I(X;Y) \le \min(H(X), H(Y))$, with equality iff $H(X|Y) = 0$ or $H(Y|X) = 0$, i.e., one is a deterministic function of the other.*

Proof. 1. By definition, $I(X;X) = D(P_{X|X}\|P_X|P_X) = \mathbb{E}_{x\sim X} D(\delta_x\|P_X)$. If P_X is discrete, then $D(\delta_x\|P_X) = \log\frac{1}{P_X(x)}$ and $I(X;X) = H(X)$. If P_X is not discrete, then let $\mathcal{A} = \{x : P_X(x) > 0\}$ denote the set of atoms of P_X. Let $\Delta = \{(x,x) : x \notin \mathcal{A}\} \subset \mathcal{X}\times\mathcal{X}$. Then $P_{X,X}(\Delta) = P_X(\mathcal{A}^c) > 0$ but since
$$(P_X \times P_X)(E) \triangleq \int_{\mathcal{X}} P_X(dx_1)\int_{\mathcal{X}} P_X(dx_2) 1\{(x_1,x_2)\in E\}$$
we have by taking $E = \Delta$ that $(P_X \times P_X)(\Delta) = 0$. Thus $P_{X,X} \not\ll P_X \times P_X$ and thus
$$I(X;X) = D(P_{X,X}\|P_X P_X) = +\infty\,.$$

2. $\mathbb{E}\log\frac{P_{XY}}{P_X P_Y} = \mathbb{E}\left[\log\frac{1}{P_X} + \log\frac{1}{P_Y} - \log\frac{1}{P_{XY}}\right]$. □

3. Similarly, when $P_{X,Y}$ and $P_X P_Y$ have densities p_{XY} and $p_X p_Y$ we have
$$D(P_{XY}\|P_X P_Y) \triangleq \mathbb{E}\left[\log\frac{p_{XY}}{p_X p_Y}\right] = h(X) + h(Y) - h(X,Y)$$

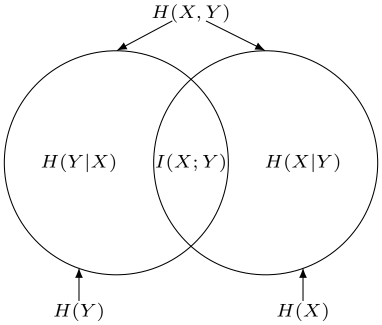
4. Follows from 2.

Corollary 2.4 (Conditioning reduces entropy)**.** *X discrete: $H(X|Y) \le H(X)$, with equality iff $X \perp\!\!\!\perp Y$.*
Intuition: *The amount of entropy reduction = mutual information*

Example: $X = U \,\texttt{OR}\, Y$, where $U, Y \overset{\text{i.i.d.}}{\sim} \text{Bern}(\frac{1}{2})$. Then $X \sim \text{Bern}(\frac{3}{4})$ and $H(X) = h(\frac{1}{4}) < 1\,\texttt{bits} = H(X|Y=0)$, i.e., conditioning on $Y = 0$ increases entropy. But *on average*, $H(X|Y) = \mathbb{P}[Y=0]\,H(X|Y=0) + \mathbb{P}[Y=1]\,H(X|Y=1) = \frac{1}{2}\,\texttt{bits} < H(X)$, by the strong concavity of $h(\cdot)$.

Note: Information, entropy and Venn diagrams:

1. The following Venn diagram illustrates the relationship between entropy, conditional entropy, joint entropy, and mutual information.



2. If you do the same for 3 variables, you will discover that the triple intersection corresponds to

$$H(X_1) + H(X_2) + H(X_3) - H(X_1, X_2) - H(X_2, X_3) - H(X_1, X_3) + H(X_1, X_2, X_3) \tag{2.4}$$

   which is sometimes denoted $I(X;Y;Z)$. It can be both positive and negative (why?).

3. In general, one can treat random variables as sets (so that r.v. X_i corresponds to set E_i and (X_1, X_2) corresponds to $E_1 \cup E_2$). Then we can define a unique signed measure μ on the finite algebra generated by these sets so that every information quantity is found by replacing

$$I/H \to \mu \quad ; \to \cap \quad , \to \cup \quad | \to \setminus .$$

   As an example, we have

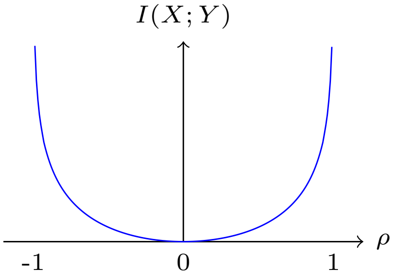
$$H(X_1|X_2, X_3) = \mu(E_1 \setminus (E_2 \cup E_3)), \tag{2.5}$$
$$I(X_1, X_2; X_3|X_4) = \mu(((E_1 \cup E_2) \cap E_3) \setminus E_4). \tag{2.6}$$

   By inclusion-exclusion, quantity (2.4) corresponds to $\mu(E_1 \cap E_2 \cap E_3)$, which explains why μ is not necessarily a positive measure.

Example: *Bivariate Gaussian.* X, Y — jointly Gaussian

$$I(X;Y) = \frac{1}{2} \log \frac{1}{1 - \rho_{XY}^2}$$

where $\rho_{XY} \triangleq \frac{\mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]}{\sigma_X \sigma_Y} \in [-1, 1]$ is the correlation coefficient.



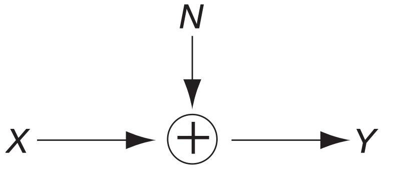
Proof. WLOG, by shifting and scaling if necessary, we can assume $\mathbb{E}X = \mathbb{E}Y = 0$ and $\mathbb{E}X^2 = \mathbb{E}Y^2 = 1$. Then $\rho = \mathbb{E}XY$. By joint Gaussianity, $Y = \rho X + Z$ for some $Z \sim \mathcal{N}(0, 1 - \rho^2) \perp\!\!\!\perp X$. Then using the divergence formula for Gaussians (1.16), we get

$$\begin{aligned} I(X;Y) &= D(P_{Y|X} \| P_Y | P_X) \\ &= \mathbb{E} D(\mathcal{N}(\rho X, 1 - \rho^2) \| \mathcal{N}(0, 1)) \\ &= \mathbb{E}\left[\frac{1}{2} \log \frac{1}{1 - \rho^2} + \frac{\log e}{2} \left((\rho X)^2 + 1 - \rho^2 - 1\right)\right] \\ &= \frac{1}{2} \log \frac{1}{1 - \rho^2} \end{aligned}$$

□

Note: Similar to the role of mutual information, the correlation coefficient also measures the dependency between random variables which are real-valued (more generally, on an inner-product space) in certain sense. However, mutual information is invariant to bijections and more general: it can be defined not just for numerical random variables, but also for apples and oranges.

Example: *Additive white Gaussian noise (AWGN) channel.* $X \perp\!\!\!\perp N$ — independent Gaussian



$$I(X;X+N) = \tfrac{1}{2}\log\Big(1+\underbrace{\tfrac{\sigma_X^2}{\sigma_N^2}}_{\text{signal-to-noise ratio (SNR)}}\Big)$$

Example: *Gaussian vectors.* $\mathbf{X}\in\mathbb{R}^m, \mathbf{Y}\in\mathbb{R}^n$ — jointly Gaussian

$$I(\mathbf{X};\mathbf{Y}) = \frac{1}{2}\log\frac{\det\Sigma_{\mathbf{X}}\det\Sigma_{\mathbf{Y}}}{\det\Sigma_{[\mathbf{X},\mathbf{Y}]}}$$

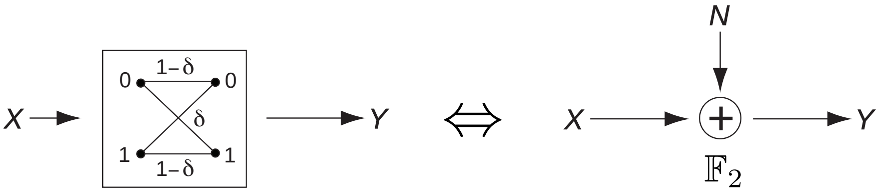
where $\Sigma_{\mathbf{X}} \triangleq \mathbb{E}\left[(\mathbf{X}-\mathbb{E}\mathbf{X})(\mathbf{X}-\mathbb{E}\mathbf{X})'\right]$ denotes the covariance matrix of $\mathbf{X}\in\mathbb{R}^m$, and $\Sigma_{[\mathbf{X},\mathbf{Y}]}$ denotes the the covariance matrix of the random vector $[\mathbf{X},\mathbf{Y}]\in\mathbb{R}^{m+n}$.

In the special case of additive noise: $\mathbf{Y}=\mathbf{X}+\mathbf{N}$ for $\mathbf{N}\perp\!\!\!\perp\mathbf{X}$, we have

$$I(\mathbf{X};\mathbf{X}+\mathbf{N}) = \frac{1}{2}\log\frac{\det(\Sigma_{\mathbf{X}}+\Sigma_{\mathbf{N}})}{\det\Sigma_{\mathbf{N}}}$$

since $\det\Sigma_{[\mathbf{X},\mathbf{X}+\mathbf{N}]} = \det\left(\begin{smallmatrix}\Sigma_{\mathbf{X}} & \Sigma_{\mathbf{X}}\\ \Sigma_{\mathbf{X}} & \Sigma_{\mathbf{X}}+\Sigma_{\mathbf{N}}\end{smallmatrix}\right) \overset{\text{why?}}{=} \det\Sigma_{\mathbf{X}}\det\Sigma_{\mathbf{N}}$.

Example: *Binary symmetric channel (BSC).*



$$\begin{aligned} X\ &\sim \text{Bern}\Big(\frac{1}{2}\Big),\ N\sim\text{Bern}(\delta)\\ Y &= X+N\\ I(X;Y) &= \log 2 - h(\delta)\end{aligned}$$

Example: *Addition over finite groups.* X is uniform on G and independent of Z. Then

$$I(X;X+Z) = \log|G| - H(Z)$$

Proof. Show that $X+Z$ is uniform on G regardless of Z. □

2.4 Conditional mutual information and conditional independence

Definition 2.4 (Conditional mutual information)**.**

$$\begin{aligned} I(X;Y|Z) &= D(P_{XY|Z} \| P_{X|Z}P_{Y|Z}|P_Z) && (2.7) \\ &= \mathbb{E}_{z\sim P_Z}[I(X;Y|Z=z)]\,. && (2.8) \end{aligned}$$

where the product of two random transformations is $(P_{X|Z=z}P_{Y|Z=z})(x,y) \triangleq P_{X|Z}(x|z)P_{Y|Z}(y|z)$, under which X and Y are independent conditioned on Z.

Note: $I(X;Y|Z)$ is a functional of P_{XYZ}.

Remark 2.2 (Conditional independence)**.** A family of distributions can be represented by a directed acyclic graph. A simple example is a Markov chain (line graph), which represents distributions that factor as $\{P_{XYZ} : P_{XYZ} = P_X P_{Y|X} P_{Z|Y}\}$.

$$\text{Cond. indep. notation} \left\{ \begin{array}{lll} X \to Y \to Z & \Leftrightarrow & P_{XZ|Y} = P_{X|Y} \cdot P_{Z|Y} \\ & \Leftrightarrow & P_{Z|XY} = P_{Z|Y} \\ & \Leftrightarrow & P_{XYZ} = P_X \cdot P_{Y|X} \cdot P_{Z|Y} \\ & \Leftrightarrow & X, Y, Z \text{ form a Markov chain} \\ & \Leftrightarrow & X \perp\!\!\!\perp Z|Y \\ & \Leftrightarrow & P_{XYZ} = P_Y \cdot P_{X|Y} \cdot P_{Z|Y} \\ & \Leftrightarrow & Z \to Y \to X \end{array} \right.$$

Theorem 2.5 (Further properties of Mutual Information)**.**

1. *$I(X;Z|Y) \geq 0$, with equality iff $X \to Y \to Z$*

2. *(Kolmogorov identity or small chain rule)*

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

3. *(**Data Processing**) If $X \to Y \to Z$, then*

   a) $I(X;Z) \leq I(X;Y)$

   b) $I(X;Y|Z) \leq I(X;Y)$

4. *(Full chain rule)*

$$I(X^n;Y) = \sum_{k=1}^{n} I(X_k;Y|X^{k-1})$$

Proof. 1. By definition and Theorem 2.3.3.

2.

$$\frac{P_{XYZ}}{P_{XY}P_Z} = \frac{P_{XZ}}{P_X P_Z} \cdot \frac{P_{Y|XZ}}{P_{Y|X}}$$

3. Apply Kolmogorov identity to $I(Y,Z;X)$:

$$\begin{aligned}I(Y,Z;X) &= I(X;Y) + \underbrace{I(X;Z|Y)}_{=0} \\ &= I(X;Z) + I(X;Y|Z)\end{aligned}$$

4. Recursive application of Kolmogorov identity. □

Example: 1-to-1 function $\Rightarrow I(X;Y) = I(X;f(Y))$
Note: In general, $I(X;Y|Z) \gtrless I(X;Y)$. Examples:

a) ">": Conditioning does not always decrease M.I. To find counterexamples when X,Y,Z do not form a Markov chain, notice that there is only one directed acyclic graph non-isomorphic to $X \to Y \to Z$, namely $X \to Y \leftarrow Z$. Then a counterexample is

$$\begin{aligned} X, Z \overset{\text{i.i.d.}}{\sim} \text{Bern}(\frac{1}{2}) & \qquad Y = X \oplus Z \\ I(X;Y) = 0 & \qquad \text{since } X \perp\!\!\!\perp Y \\ I(X;Y|Z) = I(X;X \oplus Z|Z) = H(X) = \log 2 & \end{aligned}$$

b) "<": $Z = Y$. Then $I(X;Y|Y) = 0$.

Note: (Chain rule for $I \Rightarrow$ Chain rule for H) Set $Y = X^n$. Then $H(X^n) = I(X^n;X^n) = \sum_{k=1}^n I(X_k;X^n|X^{k-1}) = \sum_{k=1}^n H(X_k|X^{k-1})$, since $H(X_k|X^n, X^{k-1}) = 0$.

Remark 2.3 (Data processing for mutual information via data processing of divergence)**.** We proved data processing for mutual information in Theorem 2.5 using Kolmogorov's identity. In fact, data processing for mutual information is *implied by* the data processing for divergence:

$$I(X;Z) = D(P_{Z|X} \| P_Z | P_X) \le D(P_{Y|X} \| P_Y | P_X) = I(X;Y),$$

where note that for each x, we have $P_{Y|X=x} \xrightarrow{P_{Z|Y}} P_{Z|X=x}$ and $P_Y \xrightarrow{P_{Z|Y}} P_Z$. Therefore if we have a bi-variate functional of distributions $\mathcal{D}(P\|Q)$ which satisfies data processing, then we can define an "M.I.-like" quantity via $I_{\mathcal{D}}(X;Y) \triangleq \mathcal{D}(P_{Y|X}\|P_Y|P_X) \triangleq \mathbb{E}_{x\sim P_X}\mathcal{D}(P_{Y|X=x}\|P_Y)$ which will satisfy data processing on Markov chains. A rich class of examples arises by taking $\mathcal{D} = D_f$ (an f-divergence, defined in (1.15)). That f-divergence satisfies data-processing is going to be shown in Remark 4.2.

2.5 Strong data-processing inequalities

For many random transformations $P_{Y|X}$, it is possible to improve the data-processing inequality (2.3): For any P_X, Q_X we have

$$D(P_Y \| Q_Y) \le \eta_{KL} D(P_X \| Q_X),$$

where $\eta_{KL} < 1$ and depends on the channel $P_{Y|X}$ only. Similarly, this gives an improvement in the data-processing inequality for mutual information: For any $P_{U,X}$ we have

$$U \to X \to Y \quad \Longrightarrow \quad I(U;Y) \le \eta_{KL} I(U;X).$$

For example, for $P_{Y|X} = BSC(\delta)$ we have $\eta_{KL} = (1-2\delta)^2$. Strong data-processing inequalities quantify the intuitive observation that noise inside the channel $P_{Y|X}$ must reduce the information that Y carries about the data U, regardless of how smart the hook up $U \to X$ is.

This is an active area of research, see [PW15] for a short summary.

2.6* How to avoid measurability problems?

As we mentioned in Remark 2.1 conditions imposed by Definition 2.1 on $P_{Y|X}$ are insufficient. Namely, we get the following two issues:

1. Radon-Nikodym derivatives such as $\frac{dP_{Y|X=x}}{dQ_Y}(y)$ may not be jointly measurable in (x, y)
2. Set $\{x : P_{Y|X=x} \ll Q_Y\}$ may not be measurable.

The easiest way to avoid all such problems is the following:

> **Agreement A1:** All conditional kernels $P_{Y|X} : \mathcal{X} \to \mathcal{Y}$ in these notes will be assumed to be defined by choosing a σ-finite measure μ_2 on $\mathcal{Y}$ and measurable function $\rho(y|x) \geq 0$ on $\mathcal{X} \times \mathcal{Y}$ such that
>
> $$P_{Y|X}(A|x) = \int_A \rho(y|x)\mu_2(dy)$$
>
> for all x and measurable sets A and $\int_{\mathcal{Y}} \rho(y|x)\mu_2(dy) = 1$ for all x.

Notes:

1. Given another kernel $Q_{Y|X}$ specified via $\rho'(y|x)$ and μ_2' we may first replace μ_2 and μ_2' via $\mu_2'' = \mu_2 + \mu_2'$ and thus assume that both $P_{Y|X}$ and $Q_{Y|X}$ are specified in terms of the same dominating measure μ_2''. (This modifies $\rho(y|x)$ to $\rho(y|x)\frac{d\mu_2}{d\mu_2''}(y)$.)

2. Given two kernels $P_{Y|X}$ and $Q_{Y|X}$ specified in terms of the same dominating measure μ_2 and functions $\rho_P(y|x)$ and $\rho_Q(y|x)$, respectively, we may set

   $$\frac{dP_{Y|X}}{dQ_{Y|X}} \triangleq \frac{\rho_P(y|x)}{\rho_Q(y|x)}$$

   outside of $\rho_Q = 0$. When $P_{Y|X=x} \ll Q_{Y|X=x}$ the above gives a version of the Radon-Nikodym derivative, which is automatically measurable in (x, y).

3. Given Q_Y specified as

   $$dQ_Y = q(y)d\mu_2$$

   we may set

   $$A_0 = \{x : \int_{\{q=0\}} \rho(y|x)d\mu_2 = 0\}$$

   This set plays a role of $\{x : P_{Y|X=x} \ll Q_Y\}$. Unlike the latter A_0 is guaranteed to be measurable by Fubini [Ç11, Prop. 6.9]. By "plays a role" we mean that it allows to prove statements like: For any P_X

   $$P_{X,Y} \ll P_X Q_Y \iff P_X[A_0] = 1\,.$$

So, while our agreement resolves the two measurability problems above, it introduces a new one. Indeed, given a joint distribution $P_{X,Y}$ on standard Borel spaces, it is always true that one can extract a conditional distribution $P_{Y|X}$ satisfying Definition 2.1 (this is called disintegration). However, it is not guaranteed that $P_{Y|X}$ will satisfy Agreement A1. To work around this issue as well, we add another agreement:

Agreement A2: All joint distributions $P_{X,Y}$ are specified by means of data: μ_1,μ_2 – σ-finite measures on $\mathcal{X}$ and $\mathcal{Y}$, respectively, and measurable function $\lambda(x,y)$ such that

$$P_{X,Y}(E) \triangleq \int_E \lambda(x,y)\mu_1(dx)\mu_2(dy)\,.$$

Notes:

1. Again, given a finite or countable collection of joint distributions $P_{X,Y}, Q_{X,Y}, \ldots$ satisfying A2 we may without loss of generality assume they are defined in terms of a common μ_1, μ_2.

2. Given $P_{X,Y}$ satisfying A2 we can disintegrate it into conditional (satisfying A1) and marginal:

$$P_{Y|X}(A|x) = \int_A \rho(y|x)\mu_2(dy) \qquad \rho(y|x) \triangleq \frac{\lambda(x,y)}{p(x)} \tag{2.9}$$

$$P_X(A) = \int_A p(x)\mu_1(dx) \qquad p(x) \triangleq \int_{\mathcal{Y}} \lambda(x,\eta)\mu_2(d\eta) \tag{2.10}$$

with $\rho(y|x)$ defined arbitrarily for those x, for which $p(x) = 0$.

Remark 2.4. The first problem can also be resolved with the help of Doob's version of Radon-Nikodym theorem [Ç11, Chapter V.4, Theorem 4.44]: If the σ-algebra on $\mathcal{Y}$ is separable (satisfied whenever $\mathcal{Y}$ is a Polish space, for example) and $P_{Y|X=x} \ll Q_{Y|X=x}$ then there exists a jointly measurable version of Radon-Nikodym derivative

$$(x,y) \mapsto \frac{dP_{Y|X=x}}{dQ_{Y|X=x}}(y)$$