Review: Random variables

- Two methods to describe a random variable (R.V.) X:
  1. a function $X : \Omega \to \mathcal{X}$ from the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to a target space $\mathcal{X}$.
  2. a distribution P_X on some measurable space $(\mathcal{X}, \mathcal{F})$.
- Convention: capital letter – RV (e.g. X); small letter – realization (e.g. x_0).
- X — discrete if there exists a countable set $\mathcal{X} = \{x_j, j = 1, \ldots\}$ such that $\sum_{j=1}^{\infty} P_X(x_j) = 1$. $\mathcal{X}$ is called alphabet of X, $x \in \mathcal{X}$ – atoms and $P_X(x_j)$ – probability mass function (pmf).
- For discrete RV support $\mathrm{supp} P_X = \{x : P_X(x) > 0\}$.
- Vector RVs: $X_1^n \triangleq (X_1, \ldots, X_n)$. Also denoted just X^n.
- For a vector RV X^n and $S \subset \{1, \ldots, n\}$ we denote $X_S = \{X_i, i \in S\}$.

1.1 Entropy

Definition 1.1 (Entropy)**.** For a discrete R.V. X with distribution P_X:

$$\begin{aligned} H(X) &= \mathbb{E}\Big[\log \frac{1}{P_X(X)}\Big] \\ &= \sum_{x \in \mathcal{X}} P_X(x) \log \frac{1}{P_X(x)} \ . \end{aligned}$$

Definition 1.2 (Joint entropy)**.** $X^n = (X_1, X_2, \ldots, X_n)$ – a random vector with n components.

$$H(X^n) = H(X_1, \ldots, X_n) = \mathbb{E}\Big[\log \frac{1}{P_{X_1,\ldots,X_n}(X_1, \ldots, X_n)} \, .\Big]$$

Definition 1.3 (Conditional entropy)**.**

$$H(X|Y) = \mathbb{E}_{y \sim P_Y}[H(P_{X|Y=y})] = \mathbb{E}\Big[\log \frac{1}{P_{X|Y}(X|Y)}\Big],$$

i.e., the entropy of $H(P_{X|Y=y})$ averaged over P_Y.

Note:

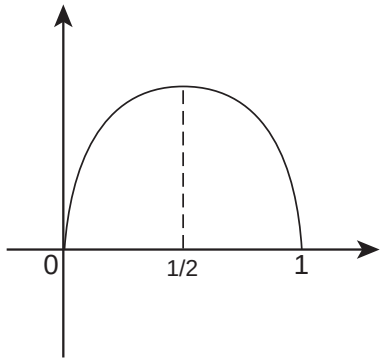
- Q: Why such definition, why log, why entropy?
  Name comes from thermodynamics. Definition is justified by theorems in this course (e.g. operationally by compression), but also by a number of experiments. For example, we can measure time it takes for ants-scouts to describe location of the food to ants-workers. It was found that when nest is placed at a root of a full binary tree of depth d and food at one of the leaves, the time was proportional to $\log 2^d = d$ – entropy of the random variable describing food location. It was estimated that ants communicate with about $0.7-1$ bit/min. Furthermore, communication time reduces if there are some regularities in path-description (e.g., paths like "left,right,left,right,left,right" were described faster). See [RZ86] for more.
- We agree that $0 \log \frac{1}{0} = 0$ (by continuity of $x \mapsto x \log \frac{1}{x}$)
- Also write $H(P_X)$ instead of $H(X)$ (abuse of notation, as customary in information theory).
- Basis of log — units

$$\begin{aligned} \log_2 &\leftrightarrow \text{bits} \\ \log_e &\leftrightarrow \text{nats} \\ \log_{256} &\leftrightarrow \text{bytes} \\ \log &\leftrightarrow \text{arbitrary units, base always matches exp} \end{aligned}$$

Example (Bernoulli): $X \in \{0,1\}$, $\mathbb{P}[X=1] = P_X(1) \triangleq p$

$$H(X) = h(p) \triangleq p \log \frac{1}{p} + \overline{p} \log \frac{1}{\overline{p}}$$

where $h(\cdot)$ is called the **binary entropy function**.



Proposition 1.1. *$h(\cdot)$ is continuous, concave on $[0,1]$ and*

$$h'(p) = \log \frac{\overline{p}}{p}$$

with infinite slope at 0 and 1.

Example (Geometric): $X \in \{0,1,2,\ldots\}$ $\mathbb{P}[X=i] = P_x(i) = p \cdot (\overline{p})^i$

$$\begin{aligned} H(X) &= \sum_{i=0}^{\infty} p \cdot \overline{p}^i \log \frac{1}{p \cdot \overline{p}^i} = \sum_{i=0}^{\infty} p\overline{p}^i \Big(i \log \frac{1}{\overline{p}} + \log \frac{1}{p} \Big) \\ &= \log \frac{1}{p} + p \cdot \log \frac{1}{\overline{p}} \cdot \frac{1-p}{p^2} = \frac{h(p)}{p} \end{aligned}$$

Example (Infinite entropy): Can $H(X) = +\infty$? Yes, $\mathbb{P}[X=k] = \frac{c}{k \ln^2 k}, k = 2,3,\cdots$

Review: Convexity

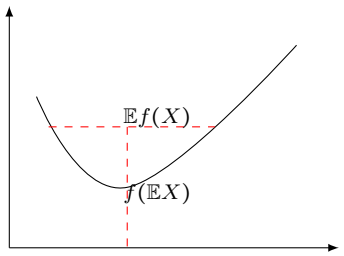
- *Convex set*: A subset S of some vector space is convex if $x, y \in S \Rightarrow \alpha x + \bar{\alpha} y \in S$ for all $\alpha \in [0,1]$. (Notation: $\bar{\alpha} \triangleq 1 - \alpha$.)

  e.g., unit interval $[0,1]$; $S = \{\text{probability distributions on } \mathcal{X}\}$, $S = \{P_X : \mathbb{E}[X] = 0\}$.

- *Convex function*: $f : S \to \mathbb{R}$ is
  - convex if $f(\alpha x + \bar{\alpha} y) \le \alpha f(x) + \bar{\alpha} f(y)$ for all $x, y \in S, \alpha \in [0,1]$.
  - strictly convex if $f(\alpha x + \bar{\alpha} y) < \alpha f(x) + \bar{\alpha} f(y)$ for all $x \ne y \in S, \alpha \in (0,1)$.
  - (strictly) concave if $-f$ is (strictly) convex.

  e.g., $x \mapsto x \log x$ is strictly convex; the mean $P \mapsto \int x \mathrm{d}P$ is convex but not strictly convex, variance is concave (Q: is it strictly concave? Think of zero-mean distributions.).

- *Jensen's inequality*: For any S-valued random variable X
  - f is convex $\Rightarrow f(\mathbb{E}X) \le \mathbb{E}f(X)$
  - f is strictly convex $\Rightarrow f(\mathbb{E}X) < \mathbb{E}f(X)$ unless X is a constant ($X = \mathbb{E}X$ a.s.)



Famous puzzle: A man says, "I am the average height and average weight of the population. Thus, I am an average man." However, he is still considered to be a little overweight. Why?
Answer: The weight is roughly proportional to the volume, which is roughly proportional to the third power of the height. Let P_X denote the distribution of the height among the population. So by Jensen's inequality, since $x \mapsto x^3$ is strictle convex on $x > 0$, we have $(\mathbb{E}X)^3 < \mathbb{E}X^3$, regardless of the distribution of X.
Source: [Yos03, Puzzle 94] or online [Har].

Theorem 1.1. *Properties of H:*

1. *(Positivity) $H(X) \ge 0$ with equality iff $X = x_0$ a.s. for some $x_0 \in \mathcal{X}$.*
2. *(Uniform maximizes entropy) $H(X) \le \log |\mathcal{X}|$, with equality iff X is uniform on $\mathcal{X}$.*
3. *(Invariance under relabeling) $H(X) = H(f(X))$ for any bijective f.*
4. *(Conditioning reduces entropy)*

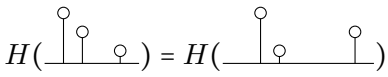
$$H(X|Y) \le H(X), \quad \text{with equality iff } X \text{ and } Y \text{ are independent.}$$

5. *(Small chain rule)*
$$H(X,Y) = H(X) + H(Y|X) \le H(X) + H(Y)$$

6. *(Entropy under functions)* $H(X) = H(X, f(X)) \ge H(f(X))$ *with equality iff* f *is one-to-one on the support of* P_X*,*

7. *(Full chain rule)*
$$H(X_1,\dots,X_n) = \sum_{i=1}^{n} H(X_i|X^{i-1}) \le \sum_{i=1}^{n} H(X_i), \tag{1.1}$$
$$\uparrow \text{ equality iff } X_1,\dots,X_n \text{ mutually independent} \tag{1.2}$$

Proof. 1. Expectation of non-negative function
2. Jensen's inequality
3. H only depends on the values of P_X, not locations:



4. Later (Lecture 2)
5. $\mathbb{E}\log\frac{1}{P_{XY}(X,Y)} = \mathbb{E}\Big[\log\frac{1}{P_X(X)\cdot P_{Y|X}(Y|X)}\Big]$
6. *Intuition:* $(X, f(X))$ contains the same amount of information as X. Indeed, $x \mapsto (x, f(x))$ is 1-1. Thus by 3 and 5:
$$H(X) = H(X, f(X)) = H(f(X)) + H(X|f(X)) \ge H(f(X))$$
The bound is attained iff $H(X|f(X)) = 0$ which in turn happens iff X is a *constant* given $f(X)$.
7. Telescoping:
$$P_{X_1X_2\cdots X_n} = P_{X_1}P_{X_2|X_1}\cdots P_{X_n|X^{n-1}}$$

□

Note: To give a preview of the *operational meaning* of entropy, let us play the following game. We are allowed to make queries about some unknown discrete R.V. X by asking yes-no questions. The objective of the game is to guess the realized value of the R.V. X. For example, $X \in \{a,b,c,d\}$ with $\mathbb{P}[X=a] = 1/2$, $\mathbb{P}[X=b] = 1/4$, and $\mathbb{P}[X=c] = \mathbb{P}[X=c] = 1/8$. In this case, we can ask "$X = a$?". If not, proceed by asking "$X = b$?". If not, ask "$X = c$?", after which we will know for sure the realization of X. The resulting average number of questions is $1/2 + 1/4 \times 2 + 1/8 \times 3 + 1/8 \times 3 = 1.75$, which equals $H(X)$ in bits. It turns out (chapter 2) that the minimal average number of yes-no questions to pin down the value of X is always between $H(X)$ `bits` and $H(X) + 1$ `bits`. In this special case the above scheme is optimal because (intuitively) it always splits the probability in half.

1.1.1 Entropy: axiomatic characterization

One might wonder why entropy is defined as $H(P) = \sum p_i \log\frac{1}{p_i}$ and if there are other definitions. Indeed, the information-theoretic definition of entropy is related to entropy in statistical physics. Also, it arises as answers to specific operational problems, e.g., the minimum average number of bits to describe a random variable as discussed above. Therefore it is fair to say that it is not pulled out of thin air.

Shannon has also showed that entropy can be defined *axiomatically*, as a function satisfying several natural conditions. Denote a probability distribution on m letters by $P = (p_1,\dots,p_m)$ and consider a functional $H_m(p_1,\dots,p_m)$. If H_m obeys the following axioms:

a) Permutation invariance

b) Expansible: $H_m(p_1, \ldots, p_{m-1}, 0) = H_{m-1}(p_1, \ldots, p_{m-1})$.

c) Normalization: $H_2(\frac{1}{2}, \frac{1}{2}) = \log 2$.

d) Continuity: $H_2(p, 1-p) \to 0$ as $p \to 0$.

e) Subadditivity: $H(X,Y) \leq H(X) + H(Y)$. Equivalently, $H_{mn}(r_{11}, \ldots, r_{mn}) \leq H_m(p_1, \ldots, p_m) + H_n(q_1, \ldots, q_n)$ whenever $\sum_{j=1}^{n} r_{ij} = p_i$ and $\sum_{i=1}^{m} r_{ij} = q_j$.

f) Additivity: $H(X,Y) = H(X)+H(Y)$ if $X \perp\!\!\!\perp Y$. Equivalently, $H_{mn}(p_1q_1, \ldots, p_mq_n) \leq H_m(p_1, \ldots, p_m)+ H_n(q_1, \ldots, q_n)$.

then $H_m(p_1, \ldots, p_m) = \sum_{i=1}^{m} p_i \log \frac{1}{p_i}$ is the only possibility. The interested reader is referred to [CT06, p. 53] and the reference therein.

1.1.2 History of entropy

In the early days of industrial age, engineers wondered if it is possible to construct a perpetual motion machine. After many failed attempts, a law of conservation of energy was postulated: a machine cannot produce more work than the amount of energy it consumed from the ambient world (this is also called the *first law* of thermodynamics). The next round of attempts was then to construct a machine that would draw energy in the form of heat from a warm body and convert it to equal (or approximately equal) amount of work. An example would be a steam engine. However, again it was observed that all such machines were highly inefficiencient, that is the amount of work produced by absorbing heat Q was $\ll Q$. The remainder of energy was dissipated to the ambient world in the form of heat. Again after many rounds of attempting various designs Clausius and Kelvin proposed another law:

> *Second law* of thermodynamics: There does not exist a machine that operates in a cycle (i.e. returns to its original state periodically), produces useful work and whose only other effect on the outside world is drawing heat from a warm body. (That is, every such machine, should expend some amount of heat to some cold body too!)[1]

Equivalent formulation is: There does not exist a cyclic process that transfers heat from a cold body to a warm body (that is, every such process needs to be helped by expending some amount of external work).

Notice that there is something annoying about the second law as compared to the first law. In the first law there is a quantity that is conserved, and this is somehow logically easy to accept. The second law seems a bit harder to believe in (and some engineers did not, and only their recurrent failures to circumvent it finally convinced them). So Clausius, building on an ingenious work of S. Carnot, figured out that there is an "explanation" to why any cyclic machine should expend heat. He proposed that there must be some hidden quantity associated to the machine, entropy of it (translated as transformative content), whose value must return to its original state. Furthermore, under any reversible (i.e. quasi-stationary, or "very slow") process operated on this machine the change of entropy is proportional to the ratio of absorbed heat and the temperature of the machine:

$$\Delta S = \frac{\Delta Q}{T}. \tag{1.3}$$

[1] Note that the reverse effect (that is converting work into heat) is rather easy: friction is an example.

So that if heat Q is absorbed at temperature T_{hot} then to return to the original state, one must return some Q' amount of heat. Q' can be significantly smaller than Q if Q' is returned at temperature $T_{cold} < T_{hot}$. Further logical arguments can convince one that for irreversible cyclic process the change of entropy at the end of the cycle can only be positive, and hence *entropy cannot reduce*.

There were a great many experimentally verified consequences that second law produced. However, what is surprising is that the mysterious entropy did not have any formula for it (unlike say energy), and thus had to be computed indirectly on the basis of relation (1.3). This was changed with the revolutionary work of Boltzmann and Gibbs, who showed that for a system of n particles the entropy of a given macro-state can be computed as

$$S = kn \sum_{j=1}^{\ell} p_j \log \frac{1}{p_j},$$

where k is the Boltzmann constant, we assume that each particle can only be in one of ℓ molecular states (e.g. spin up/down, or if we quantize the phase volume into ℓ subcubes) and p_j is the fraction of particles in j-th molecular state.

1.1.3* Entropy: submodularity

Recall that $[n]$ denotes a set $\{1, \ldots, n\}$, $\binom{S}{k}$ denotes subsets of S of size k and 2^S denotes all subsets of S. A set function $f : 2^S \to \mathbb{R}$ is called submodular if for any $T_1, T_2 \subset S$

$$f(T_1 \cup T_2) + f(T_1 \cap T_2) \le f(T_1) + f(T_2)$$

Submodularity is similar to concavity, in the sense that "adding elements gives diminishing returns". Indeed consider $T' \subset T$ and $b \notin T$. Then

$$f(T \cup b) - f(T) \le f(T' \cup b) - f(T').$$

Theorem 1.2. *Let X^n be discrete RV. Then $T \mapsto H(X_T)$ is submodular.*

Proof. Let $A = X_{T_1 \setminus T_2}, B = X_{T_1 \cap T_2}, C = X_{T_2 \setminus T_1}$. Then we need to show

$$H(A, B, C) + H(B) \le H(A, B) + H(B, C).$$

This follows from a simple chain

$$\begin{aligned} H(A, B, C) + H(B) &= H(A, C|B) + 2H(B) && (1.4)\\ &\le H(A|B) + H(C|B) + 2H(B) && (1.5)\\ &= H(A, B) + H(B, C) && (1.6) \end{aligned}$$

□

Note that entropy is not only submodular, but also monotone:

$$T_1 \subset T_2 \implies H(X_{T_1}) \le H(X_{T_2}).$$

So fixing n, let us denote by Γ_n the set of all non-negative, monotone, submodular set-functions on $[n]$. Note that via an obvious enumeration of all non-empty subsets of $[n]$, Γ_n is a closed convex cone in $\mathbb{R}_+^{2^n-1}$. Similarly, let us denote by Γ_n^* the set of all set-functions corresponding to

distributions on X^n. Let us also denote $\bar{\Gamma}_n^*$ the closure of Γ_n^*. It is not hard to show, cf. [ZY97], that $\bar{\Gamma}_n^*$ is also a closed convex cone and that

$$\Gamma_n^* \subset \bar{\Gamma}_n^* \subset \Gamma_n \,.$$

The astonishing result of [ZY98] is that

$$\Gamma_2^* = \bar{\Gamma}_2^* = \Gamma_2 \tag{1.7}$$
$$\Gamma_3^* \subsetneq \bar{\Gamma}_3^* = \Gamma_3 \tag{1.8}$$
$$\Gamma_n^* \subsetneq \bar{\Gamma}_n^* \subsetneq \Gamma_n \qquad n \geq 4 \,. \tag{1.9}$$

This follows from the fundamental new information inequality not implied by the submodularity of entropy (and thus called *non-Shannon inequality*). Namely, [ZY98] shows that for any 4 discrete random variables:

$$I(X_3; X_4) - I(X_3; X_4|X_1) - I(X_3; X_4|X_2) \leq \frac{1}{2} I(X_1; X_2) + \frac{1}{4} I(X_1; X_3, X_4) + \frac{1}{4} I(X_2; X_3, X_4) \,.$$

(see Definition 2.3).

1.1.4 Entropy: Han's inequality

Theorem 1.3 (Han's inequality)**.** *Let X^n be discrete n-dimensional RV and denote $\bar{H}_k(X^n) = \frac{1}{\binom{n}{k}} \sum_{T \subset \binom{[n]}{k}} H(X_T)$ – the average entropy of a k-subset of coordinates. Then $\frac{\bar{H}_k}{k}$ is decreasing in k:*

$$\frac{1}{n} \bar{H}_n \leq \cdots \leq \frac{1}{k} \bar{H}_k \cdots \leq \bar{H}_1 \,. \tag{1.10}$$

Furthermore, the sequence $\bar{H}_k$ is increasing and concave in the sense of decreasing slope:

$$\bar{H}_{k+1} - \bar{H}_k \leq \bar{H}_k - \bar{H}_{k-1} \,. \tag{1.11}$$

Proof. Denote for convenience $\bar{H}_0 = 0$. Note that $\frac{\bar{H}_m}{m}$ is an average of differences:

$$\frac{1}{m} \bar{H}_m = \frac{1}{m} \sum_{k=1}^{m} (\bar{H}_k - \bar{H}_{k-1})$$

Thus, it is clear that (1.11) implies (1.10) since increasing m by one adds a smaller element to the average. To prove (1.11) observe that from submodularity

$$H(X_1, \ldots, X_{k+1}) + H(X_1, \ldots, X_{k-1}) \leq H(X_1, \ldots, X_k) + H(X_1, \ldots, X_{k-1}, X_{k+1}) \,.$$

Now average this inequality over all $n!$ permutations of indices $\{1, \ldots, n\}$ to get

$$\bar{H}_{k+1} + \bar{H}_{k-1} \leq 2\bar{H}_k$$

as claimed by (1.11).

Alternative proof: Notice that by "conditioning decreases entropy" we have

$$H(X_{k+1}|X_1, \ldots, X_k) \leq H(X_{k+1}|X_2, \ldots, X_k) \,.$$

Averaging this inequality over all permutations of indices yields (1.11). □

Note: Han's inequality holds for any submodular set-function.
Example: Another submodular set-function is

$$S \mapsto I(X_S; X_{S^c}) .$$

Han's inequality for this one reads

$$0 = \frac{1}{n} I_n \le \cdots \le \frac{1}{k} I_k \cdots \le I_1 ,$$

where $I_k = \frac{1}{\binom{n}{k}} \sum_{S:|S|=k} I(X_S; X_{S^c})$ – gauges the amount of k-subset coupling in the random vector X^n.

1.2 Divergence

Review: Measurability

In this course we will assume that all alphabets are standard Borel spaces. Some of the nice properties of standard Borel spaces:

- all complete separable metric spaces, endowed with Borel σ-algebras are standard Borel. In particular, countable alphabets and $\mathbb{R}^n$ and $\mathbb{R}^\infty$ (space of sequences) are standard Borel.
- if $\mathcal{X}_i, i = 1, \ldots$ are s.B.s. then so is $\prod_{i=1}^\infty \mathcal{X}_i$
- singletons $\{x\}$ are measurable sets
- diagonal $\Delta = \{(x, x) : x \in \mathcal{X}\}$ is measurable in $\mathcal{X} \times \mathcal{X}$
- (Most importantly) for any probability distribution $P_{X,Y}$ on $\mathcal{X} \times \mathcal{Y}$ there exists a transition probability kernel (also called a regular branch of a conditional distribution) $P_{Y|X}$ s.t.

$$P_{X,Y}[E] = \int_{\mathcal{X}} P_X(dx) \int_{\mathcal{Y}} P_{Y|X=x}(dy) 1\{(x, y) \in E\} .$$

Intuition: $D(P\|Q)$ gauges the **dissimilarity** between P and Q.

Definition 1.4 (Divergence)**.** Let P, Q be distributions on

- $\mathcal{A}$ = discrete alphabet (finite or countably infinite)

$$D(P\|Q) \triangleq \sum_{a \in \mathcal{A}} P(a) \log \frac{P(a)}{Q(a)} ,$$

where we agree:

(1) $0 \cdot \log \frac{0}{0} = 0$
(2) $\exists a : Q(a) = 0, P(a) > 0 \Rightarrow D(P\|Q) = \infty$

- $\mathcal{A} = \mathbb{R}^k$, P and Q have densities f_P and f_Q

$$D(P\|Q) = \begin{cases} \int_{\mathbb{R}^k} \log \frac{f_P(x^k)}{f_Q(x^k)} f_P(x^k) dx^k & , \quad \mathrm{Leb}\{f_P > 0, f_Q = 0\} = 0 \\ +\infty & , \quad \text{otherwise} \end{cases}$$

- $\mathcal{A}$ — measurable space:

$$D(P\|Q) = \begin{cases} \mathbb{E}_Q \frac{dP}{dQ} \log \frac{dP}{dQ} = \mathbb{E}_P \log \frac{dP}{dQ} & , \quad P \ll Q \\ +\infty & , \quad \text{otherwise} \end{cases}$$

(Also known as information divergence, Kullback–Leibler divergence, relative entropy.)

Notes:

- (Radon-Nikodym theorem) Recall that for two measures P and Q, we say P is absolutely continuous w.r.t. Q (denoted by $P \ll Q$) if $Q(E) = 0$ implies $P(E) = 0$ for all measurable E. If $P \ll Q$, then there exists a function $f : \mathcal{X} \to \mathbb{R}_+$ such that for any measurable set E,

$$P(E) = \int_E f \mathrm{d}Q. \qquad \text{[change of measure]}$$

  Such f is called a density (or a Radon-Nikodym derivative) of P w.r.t. Q, denoted by $\frac{\mathrm{d}P}{\mathrm{d}Q}$. For finite alphabets, we can just take $\frac{dP}{dQ}(x)$ to be the ratio of the pmfs. For P and Q on $\mathbb{R}^n$ possessing pdfs we can take $\frac{dP}{dQ}(x)$ to be the ratio of pdfs.

- (Infinite values) $D(P\|Q)$ can be ∞ also when $P \ll Q$, but the two cases of $D(P\|Q) = +\infty$ are consistent since $D(P\|Q) = \sup_\Pi D(P_\Pi \| Q_\Pi)$, where Π is a finite partition of the underlying space $\mathcal{A}$ (proof: later)

- (Asymmetry) $D(P\|Q) \neq D(Q\|P)$. Asymmetry can be very useful. Example: $P(H) = P(T) = 1/2$, $Q(H) = 1$. Upon observing HHHHHHH, one tends to believe it is Q but can never be absolutely sure; Upon observing HHT, know for sure it is P. Indeed, $D(P\|Q) = \infty$, $D(Q\|P) = 1\,\mathtt{bit}$.

- (Pinsker's inequality) There are many other measures for dissimilarity, e.g., total variation (L_1-distance)

$$\mathrm{TV}(P,Q) \triangleq \sup_E P[E] - Q[E] \tag{1.12}$$

$$= \frac{1}{2} \int |\mathrm{d}P - \mathrm{d}Q| \qquad = (\text{discrete case}) \frac{1}{2} \sum_x |P(x) - Q(x)| \,. \tag{1.13}$$

  This one is symmetric. There is a famous Pinsker's (or Pinsker-Csiszár) inequality relating D and TV:

$$\mathrm{TV}(P,Q) \le \sqrt{\frac{1}{2 \log e} D(P\|Q)} \,. \tag{1.14}$$

- (Other divergences) A general class of divergence-like measures was proposed by Csiszár. Fixing a convex function $f : \mathbb{R}_+ \to \mathbb{R}$ with $f(1) = 0$ we define *f-divergence* D_f as

$$D_f(P\|Q) \triangleq \mathbb{E}_Q \left[f\left(\frac{dP}{dQ} \right) \right] . \tag{1.15}$$

This encompasses total variation, χ^2-distance, Hellinger, Tsallis etc. Inequalities between various f-divergences such as (1.14) was once an active field of research. It was made largely irrelevant by a work of Harremoës and Vajda [HV11] giving a simple method for obtaining best possible inequalities between any two f-divergences.

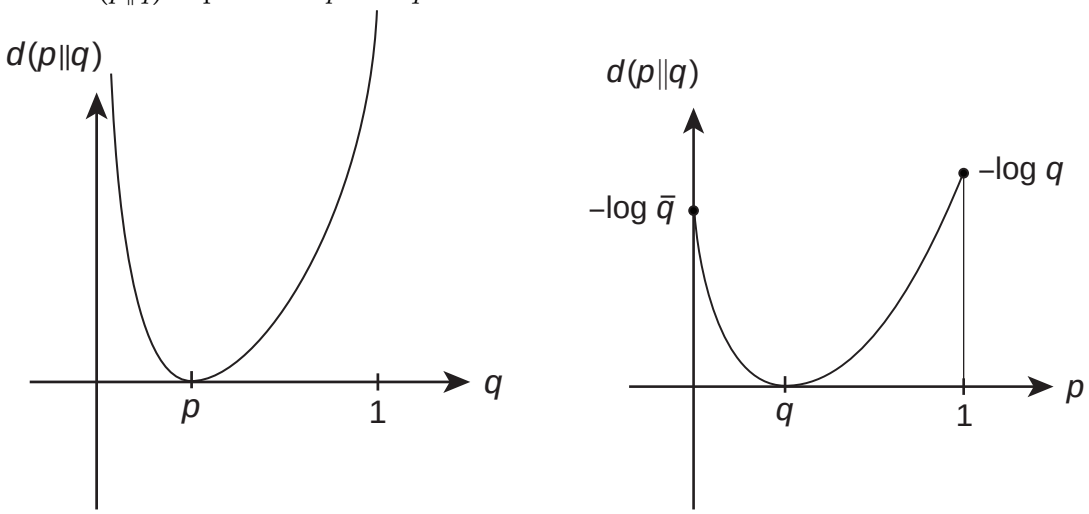
Theorem 1.4 (H v.s. D)**.** *If distribution P is supported on $\mathcal{A}$ with $|\mathcal{A}| < \infty$, then*

$$H(P) = \log|\mathcal{A}| - D(P\|\underbrace{U_{\mathcal{A}}}_{\text{uniform distribution on } \mathcal{A}}).$$

Example (Binary divergence): $\mathcal{A} = \{0,1\}$; $P = [p, \overline{p}]$; $Q = [q, \overline{q}]$

$$D(P\|Q) = d(p\|q) \triangleq p\log\frac{p}{q} + \overline{p}\log\frac{\overline{p}}{\overline{q}}$$

Here is how $d(p\|q)$ depends on p and q:



Quadratic lower bound (homework):

$$d(p\|q) \geq 2(p-q)^2\log e$$

Example (Real Gaussian): $\mathcal{A} = \mathbb{R}$

$$D(\mathcal{N}(m_1,\sigma_1^2)\|\mathcal{N}(m_0,\sigma_0^2)) = \frac{1}{2}\log\frac{\sigma_0^2}{\sigma_1^2} + \frac{1}{2}\Big[\frac{(m_1-m_0)^2}{\sigma_0^2} + \frac{\sigma_1^2}{\sigma_0^2} - 1\Big]\log e \tag{1.16}$$

Example (Complex Gaussian): $\mathcal{A} = \mathbb{C}$. The pdf of $\mathcal{N}_c(m,\sigma^2)$ is $\frac{1}{\pi\sigma^2}e^{-|x-m|^2/\sigma^2}$, or equivalently:

$$\mathcal{N}_c(m,\sigma^2) = \mathcal{N}\left(\begin{bmatrix}\mathrm{Re}(m) & \mathrm{Im}(m)\end{bmatrix}, \begin{bmatrix}\sigma^2/2 & 0\\ 0 & \sigma^2/2\end{bmatrix}\right) \tag{1.17}$$

$$D(\mathcal{N}_c(m_1,\sigma_1^2)\|\mathcal{N}_c(m_0,\sigma_0^2)) = \log\frac{\sigma_0^2}{\sigma_1^2} + \Big[\frac{|m_1-m_0|^2}{\sigma_0^2} + \frac{\sigma_1^2}{\sigma_0^2} - 1\Big]\log e \tag{1.18}$$

Example (Vector Gaussian): $\mathcal{A} = \mathbb{C}^k$

$$\begin{aligned} D(\mathcal{N}_c(m_1,\Sigma_1)\|\mathcal{N}_c(m_0,\Sigma_0)) &= \log\det\Sigma_0 - \log\det\Sigma_1 + (m_1-m_0)^H\Sigma_0^{-1}(m_1-m_0)\log e \\ &\quad + \mathrm{tr}(\Sigma_0^{-1}\Sigma_1 - I)\log e \end{aligned}$$

(assume $\det \Sigma_0 \neq 0$).

Note: The definition of $D(P\|Q)$ extends verbatim to measures P and Q (not necessarily probability measures), in which case $D(P\|Q)$ can be negative. A sufficient condition for $D(P\|Q) \geq 0$ is that P is a probability measure and Q is a sub-probability measure, i.e., $\int \mathrm{d}Q \leq 1 = \int \mathrm{d}P$.

1.3 Differential entropy

The notion of *differential entropy* is simply the divergence with respect to the Lebesgue measure:

Definition 1.5. The differential entropy of a random vector X^k is

$$h(X^k) = h(P_{X^k}) \triangleq -D(P_{X^k}\|\mathrm{Leb}). \tag{1.19}$$

In particular, if X^k has probability density function (pdf) p, then $h(X^k) = \mathbb{E}\log\frac{1}{p(X^k)}$; otherwise $h(X^k) = -\infty$. Conditional differential entropy $h(X^k|Y) \triangleq \mathbb{E}\log\frac{1}{p_{X^k|Y}(X^k|Y)}$ where $p_{X^k|Y}$ is a conditional pdf.

Warning: Even for X with pdf $h(X)$ can be positive, negative, take values of $\pm\infty$ or even be undefined[2].

Nevertheless, differential entropy shares many properties with the usual entropy:

Theorem 1.5 (Properties of differential entropy)**.** *Assume that all differential entropies appearing below exists and are finite (in particular all RVs have pdfs and conditional pdfs). Then the following hold :*

1. *(Uniform maximizes diff. entropy) If $\mathbb{P}[X^n \in S] = 1$ then $h(X^n) \leq \mathrm{Leb}\{S\}$ with equality iff X^n is uniform on S.*
2. *(Conditioning reduces diff. entropy) $h(X|Y) \leq h(X)$ (here Y could be arbitrary, e.g. discrete)*
3. *(Chain rule)*
$$h(X^n) = \sum_{k=1}^{n} h(X_k|X^{k-1})\,.$$
4. *(Submodularity) The set-function $T \mapsto h(X_T)$ is submodular.*
5. *(Han's inequality) The function $k \mapsto \frac{1}{k\binom{n}{k}}\sum_{T\in\binom{[n]}{k}} h(X_T)$ is decreasing in k.*

1.3.1 Application of differential entropy: Loomis-Whitney and Bollobás-Thomason

The following famous result shows that n-dimensional rectangle simultaneously minimizes volumes of all projections:[3]

Theorem 1.6 (Bollobás-Thomason Box Theorem)**.** *Let $K \subset \mathbb{R}^n$ be a compact set. For $S \subset [n]$ denote by K_S – projection of K on the subset S of coordinate axes. Then there exists a rectangle A s.t. $\mathrm{Leb}\{A\} = \mathrm{Leb}\{K\}$ and for all $S \subset [n]$:*

$$\mathrm{Leb}\{A_S\} \leq \mathrm{Leb}\{K_S\}$$

[2] For an example, consider piecewise-constant pdf taking value $e^{(-1)^n n}$ on the n-th interval of width $\Delta_n = \frac{c}{n^2}e^{-(-1)^n n}$.

[3] Note that since K is compact, its projection and slices are all compact and hence measurable.

Proof. Let X^n be uniformly distributed on K. Then $h(X^n) = \log \mathrm{Leb}\{K\}$. Let A be rectangle $a_1 \times \cdots \times a_n$ where

$$\log a_i = h(X_i|X^{i-1}) \,.$$

Then, we have by 1. in Theorem 1.5

$$h(X_S) \le \log \mathrm{Leb}\{K_S\}$$

On the other hand, by the chain rule

$$\begin{aligned} h(X_S) &= \sum_{i=1}^{n} 1\{i \in S\} h(X_i|X_{[i-1]\cap S}) && (1.20)\\ &\ge \sum_{i\in S} h(X_i|X^{i-1}) && (1.21)\\ &= \log \prod_{i\in S} a_i && (1.22)\\ &= \log \mathrm{Leb}\{A_S\} && (1.23) \end{aligned}$$

□

Corollary 1.1 (Loomis-Whitney)**.** *Let K be a compact subset of $\mathbb{R}^n$ and let K_{j^c} denote projection of K on coordinate axes $[n] \setminus j$. Then*

$$\mathrm{Leb}\{K\} \le \prod_{j=1}^{n} \mathrm{Leb}\{K_{j^c}\}^{\frac{1}{n-1}} \,. \tag{1.24}$$

Proof. Apply previous theorem to construct rectangle A and note that

$$\mathrm{Leb}\{K\} = \mathrm{Leb}\{A\} = \prod_{j=1}^{n} \mathrm{Leb}\{A_{j^c}\}^{\frac{1}{n-1}}$$

By previous theorem $\mathrm{Leb}\{A_{j^c}\} \le \mathrm{Leb}\{K_{j^c}\}$. □

The meaning of Loomis-Whitney inequality is best understood by introducing the average width of K in direction j: $w_j \triangleq \frac{\mathrm{Leb}\{K\}}{\mathrm{Leb}\{K_{j^c}\}}$. Then (1.24) is equivalent to

$$\mathrm{Leb}\{K\} \ge \prod_{j=1}^{n} w_j \,,$$

i.e. that volume of K is greater than volume of the rectangle of average widths.

6.441 Information Theory

Spring 2016