3.1 Sufficient statistics and data-processing

Definition 3.1 (Sufficient Statistic)**.** Let

- P_X^θ be a collection of distributions of X parameterized by θ
- $P_{T|X}$ be some probability kernel. Let $P_T^\theta \triangleq P_{T|X} \circ P_X^\theta$ be the induced distribution on T for each θ.

We say that T is a *sufficient statistic* (s.s.) of X for θ if there exists a transition probability kernel $P_{X|T}$ so that $P_X^\theta P_{T|X} = P_T^\theta P_{X|T}$. (I.e.: $P_{X|T}$ can be chosen to not depend on θ).

Note:

- Know T, can forget X (T contains all the information that is sufficient to make inference about θ)
- Obviously any one-to-one transformation of X is sufficient. Therefore the interesting case is when T is a low-dimensional recap of X (dimensionality reduction)
- θ need not be a random variable (the definition does not involve any distribution on θ)

Theorem 3.1. *Let $\theta \to X \to T$. Then the following are equivalent*

1. *T is a s.s. of X for θ.*
2. *$\forall P_\theta$, $\theta \to T \to X$.*
3. *$\forall P_\theta$, $I(\theta; X|T) = 0$.*
4. *$\forall P_\theta$, $I(\theta; X) = I(\theta; T)$, i.e., data processing inequality for M.I. holds with equality.*

Theorem 3.2 (Fisher's factorization criterion)**.** *For all $\theta \in \Theta$, let P_X^θ have a density p_θ with respect to a measure μ (e.g., discrete – pmf, continuous – pdf). Let $T = T(X)$ be a deterministic function of X. Then T is a s.s. of X for θ iff*

$$p_\theta(x) = g_\theta(T(x))h(x)$$

for some measurable functions g_θ and h, $\forall \theta \in \Theta$.

Proof. We only give the proof in the discrete case (continuous case $\sum \to \int \mathrm{d}\mu$). Let $t = T(x)$.

"$\Rightarrow$": Suppose T is a s.s. of X for θ. Then $p_\theta(x) = P_\theta(X = x) = P_\theta(X = x, T = t) = P_\theta(X = x|T = t)P_\theta(T = t) = \underbrace{P(X = x|T = T(x))}_{h(x)} \underbrace{P_\theta(T = T(x))}_{g_\theta(T(x))}$

"⇐": Suppose the factorization holds. Then

$$P_\theta(X = x|T = t) = \frac{p_\theta(x)}{\sum_x \mathbf{1}_{\{T(x)=t\}} p_\theta(x)} = \frac{g_\theta(t)h(x)}{\sum_x \mathbf{1}_{\{T(x)=t\}} g_\theta(t)h(x)} = \frac{h(x)}{\sum_x \mathbf{1}_{\{T(x)=t\}} h(x)},$$

free of θ. □

Example:

1. *Normal mean model.* Let $\theta \in \mathbb{R}$ and observations $X_i \overset{\text{indep.}}{\sim} \mathcal{N}(\theta, 1), i \in [n]$. Then the sample mean $\bar{X} = \frac{1}{n}\sum_j X_j$ is a s.s. of X^n for θ.
   Verify: $P^\theta_{X^n}$ factorizes.

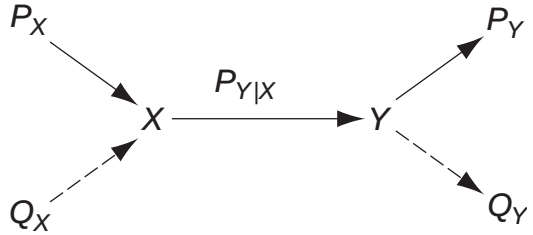
2. *Coin flips.* Let $B_i \overset{\text{i.i.d.}}{\sim} \text{Bern}(\theta)$. Then $\sum_{i=1}^n B_i$ is a s.s. of B^n for θ.

3. *Uniform distribution.* Let $U_i \overset{\text{i.i.d.}}{\sim} \text{uniform}[0, \theta]$. Then $\max_{i\in[n]} U_i$ is a s.s. of U^n for θ.

Example: *Binary hypothesis testing.* $\theta = \{0, 1\}$. Given $\theta = 0$ or 1, $X \sim P_X$ or Q_X. Then Y – the output of $P_{Y|X}$ – is a s.s. of X for θ iff $D(P_{X|Y}\|Q_{X|Y}|P_Y) = 0$, i.e., $P_{X|Y} = Q_{X|Y}$ holds P_Y-a.s. Indeed, the latter means that for kernel $Q_{X|Y}$ we have

$$P_X P_{Y|X} = P_Y Q_{X|Y} \quad \text{and} \quad Q_X P_{Y|X} = Q_Y Q_{X|Y},$$

which is precisely the definition of s.s. when $\theta \in \{0, 1\}$. This example explains condition for equality in the data-processing for divergence:



Then assuming $D(P_Y\|Q_Y) < \infty$ we have:

$$D(P_X\|Q_X) = D(P_Y\|Q_Y) \quad \Longleftrightarrow \quad Y \text{ – s.s. for testing } P_X \text{ vs. } Q_X$$

Proof: Let $Q_{XY} = Q_X P_{Y|X}$, $P_{XY} = P_X P_{Y|X}$, then

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

with equality iff $D(P_{X|Y}\|Q_{X|Y}|P_Y) = 0$, which is equivalent to Y being a s.s. for testing P_X vs Q_X as desired.

3.2 Geometric interpretation of mutual information

Mutual information as "weighted distance":

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

Theorem 3.3 (Golden formula)**.** $\forall Q_Y$ *such that* $D(P_Y\|Q_Y) < \infty$

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

Proof. For discrete case: $I(X;Y) = \mathbb{E}\log\frac{P_{Y|X}Q_Y}{P_Y Q_Y}$, group $P_{Y|X}$ and Q_Y. □

Corollary 3.1 (mutual information as center of gravity)**.**

$$I(X;Y) = \min_Q D(P_{Y|X}\|Q|P_X),$$

achieved at $Q = P_Y$.

Note: This representation is useful to bound mutual information from above.

Theorem 3.4 (mutual information as distance to product distributions)**.**

$$I(X;Y) = \min_{Q_X,Q_Y} D(P_{XY}\|Q_X Q_Y)$$

Proof. $I(X;Y) = \mathbb{E}\log\frac{P_{XY}Q_X Q_Y}{P_X P_Y Q_X Q_Y}$, group P_{XY} and $Q_X Q_Y$ and bound marginal divergences $D(P_X\|Q_X)$ and $D(P_Y\|Q_Y)$ by zero. □

Note: Generalization to conditional mutual information.

$$I(X;Z|Y) = \min_{Q_{XYZ}:X\to Y\to Z} D(P_{XYZ}\|Q_{XYZ})$$

Proof. By chain rule,

$$\begin{aligned}
&D(P_{XYZ}\|Q_X Q_{Y|X} Q_{Z|Y}) \\
=&D(P_{XYZ}\|P_X P_{Y|X} P_{Z|Y}) + D(P_X\|Q_X) + D(P_{Y|X}\|Q_{Y|X}|P_X) + D(P_{Z|Y}\|Q_{Z|Y}|P_Y) \\
=&D(P_{XYZ}\|P_Y P_{X|Y} P_{Z|Y}) + \dots \\
=&\underbrace{D(P_{XZ|Y}\|P_{X|Y}P_{Z|Y}|P_Y)}_{I(X;Z|Y)} + \dots
\end{aligned}$$

□

Interpretation: The most general graphical model for the triplet (X,Y,Z) is a 3-clique. What is the information flow on the edge $X \to Z$? To answer, notice that removing this edge restricts possible joint distributions to a Markov chain $X \to Y \to Z$. Thus, it is natural to ask what is the minimum distance between a given $P_{X,Y,Z}$ and the set of all distributions $Q_{X,Y,Z}$ satisfying the Markov chain constraint. By the above calculation, optimal $Q_{X,Y,Z} = P_Y P_{X|Y} P_{Z|Y}$ and hence the distance is $I(X;Z|Y)$. It is natural to take this number as the information flowing on the edge $X \to Z$.

3.3 Variational characterizations of divergence: Donsker-Varadhan

Why variational characterization (sup- or inf-representation): $F(x) = \sup_{\lambda\in\Lambda} f_\lambda(x)$

1. Regularity, e.g., recall
   a) Pointwise supremum of convex functions is convex
   b) Pointwise supremum of lower semicontinuous (lsc) functions is lsc
2. Give bounds by choosing a (suboptimal) λ

Theorem 3.5 (Donsker-Varadhan)**.** *Let P, Q be probability measures on $\mathcal{X}$ and let $\mathcal{C}$ denote the set of functions $f : \mathcal{X} \to \mathbb{R}$ such that $\mathbb{E}_Q[\exp\{f(X)\}] < \infty$. If $D(P\|Q) < \infty$ then for every $f \in \mathcal{C}$ expectation $\mathbb{E}_P[f(X)]$ exists and furthermore*

$$D(P\|Q) = \sup_{f\in\mathcal{C}} \mathbb{E}_P[f(X)] - \log \mathbb{E}_Q[\exp\{f(X)\}]\,. \tag{3.1}$$

Proof. "$\le$": take $f = \log\frac{dP}{dQ}$.

"$\ge$": Fix $f \in \mathcal{C}$ and define a probability measure Q^f (*tilted version* of Q) via $Q^f(dx) \triangleq \frac{\exp\{f(x)\}Q(dx)}{\int_{\mathcal{X}}\exp\{f(x)\}Q(dx)}$, or equivalently,

$$Q^f(dx) = \exp\{f(x) - Z_f\}Q(dx)\,, \qquad Z_f \triangleq \log \mathbb{E}_Q[\exp\{f(X)\}]\,.$$

Then, obviously $Q^f \ll Q$ and we have

$$\mathbb{E}_P[f(X)] - Z_f = \mathbb{E}_P\left[\log\frac{dQ^f}{dQ}\right] = \mathbb{E}_P\left[\log\frac{dPdQ^f}{dQdP}\right] = D(P\|Q) - D(P\|Q^f) \le D(P\|Q)\,. \quad \square$$

Notes:

1. What is Donsker-Varadhan good for? By setting $f(x) = \epsilon\cdot g(x)$ with $\epsilon \ll 1$ and linearizing exp and log we can see that when $D(P\|Q)$ is small, expectations under P can be approximated by expectations over Q (change of measure): $\mathbb{E}_P[g(X)] \approx \mathbb{E}_Q[g(X)]$. This holds for all functions g with finite exponential moment under Q. Total variation distance provides a similar bound, but for a narrower class of bounded functions:
$$|\mathbb{E}_P[g(X)] - \mathbb{E}_Q[g(X)]| \le \|g\|_\infty \mathrm{TV}(P, Q)\,.$$
2. More formally, inequality $\mathbb{E}_P[f(X)] \le \log \mathbb{E}_Q[\exp f(X)] + D(P\|Q)$ is useful in estimating $\mathbb{E}_P[f(X)]$ for complicated distribution P (e.g. over large-dimensional vector X^n with lots of weak inter-coordinate dependencies) by making a smart choice of Q (e.g. with iid components).
3. In the next lecture we will show that $P \mapsto D(P\|Q)$ is convex. A general method of obtaining variational formulas like (3.1) is by Young-Fenchel inequality. Indeed, (3.1) is exactly this inequality since the Fenchel-Legendre conjugate of $D(\cdot\|Q)$ is given by a convex map $f \mapsto Z_f$.

Theorem 3.6 (Weak lower-semicontinuity of divergence)**.** *Let $\mathcal{X}$ be a metric space with Borel σ-algebra $\mathcal{H}$. If P_n and Q_n converge weakly (in distribution) to P, Q, then*

$$D(P\|Q) \le \liminf_{n\to\infty} D(P_n\|Q_n)\,. \tag{3.2}$$

Proof. First method: On a metric space $\mathcal{X}$ bounded continuous functions ($\mathcal{C}_b$) are dense in the set of all integrable functions. Then in Donsker-Varadhan (3.1) we can replace $\mathcal{C}$ by $\mathcal{C}_b$ to get

$$D(P_n\|Q_n) = \sup_{f\in\mathcal{C}_b} \mathbb{E}_{P_n}[f(X)] - \log \mathbb{E}_{Q_n}[\exp\{f(X)\}]\,.$$

Recall $P_n \to P$ weakly if and only if $\mathbb{E}_{P_n} f(X) \to \mathbb{E}_P f(X)$ for all $f \in \mathcal{C}_b$. Taking the limit concludes the proof.

Second method (less mysterious): Let $\mathcal{A}$ be the algebra of Borel sets E whose boundary has zero $(P+Q)$ measure, i.e.

$$\mathcal{A} = \{E \in \mathcal{H} : (P+Q)(\partial E) = 0\}\,.$$

By the property of weak convergence P_n and Q_n converge pointwise on $\mathcal{A}$. Thus by (3.8) we have

$$D(P_{\mathcal{A}}\|Q_{\mathcal{A}}) \le \lim_{n\to\infty} D(P_{n,\mathcal{A}}\|Q_{n,\mathcal{A}})$$

If we show $\mathcal{A}$ is $(P+Q)$-dense in $\mathcal{H}$, we are done by (3.7). To get an idea, consider $\mathcal{X} = \mathbb{R}$. Then open sets are $(P+Q)$-dense in $\mathcal{H}$ (since finite measures are regular), while the algebra $\mathcal{F}$ generated by open intervals is $(P+Q)$-dense in the open sets. Since there are at most countably many points $a \in \mathcal{X}$ with $P(a)+Q(a) > 0$, we may further approximate each interval (a,b) whose boundary has non-zero $(P+Q)$ measure by a slightly larger interval from $\mathcal{A}$. □

Note: In general, $D(P\|Q)$ is *not* continuous in either P or Q. Example: Let $B_1,\dots,B_n \overset{\text{i.i.d.}}{\sim} \{\pm 1\}$ equiprobably. Then $S_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n B_i \xrightarrow{\text{D}} \mathcal{N}(0,1)$. But $D(\underbrace{P_{S_n}}_{\text{discrete}}\|\underbrace{\mathcal{N}(0,1)}_{\text{cont's}}) = \infty$ for all n. Note that this is an example for strict inequality in (3.2).

Note: Why do we care about continuity of information measures? Let's take divergence as an example.

1. *Computation.* For complicated P and Q direct computation of $D(P\|Q)$ might be hard. Instead, one may want to discretize them then let the computer compute. **Question**: Is this procedure stable, i.e., as the quantization becomes finer, does this procedure guarantee to converge to the true value? Yes! Continuity w.r.t. discretization is guaranteed by the next theorem.

2. *Estimating information measures.* In many statistical setups, oftentimes we do not know P or Q, if we estimate the distribution from data (e.g., estimate P by empirical distribution $\hat{P}_n$ from n samples) and then plug in, does $D(\hat{P}_n\|Q)$ provide a good estimator for $D(P\|Q)$? Well, note from the first example that this is a bad idea if Q is continuous, since $D(\hat{P}_n\|Q) = \infty$ for n. In fact, if one convolves the empirical distribution with a tiny bit of, say, Gaussian distribution, then it will always have a density. If we allow the variance of the Gaussian to vanish with n appropriately, we will have convergence. This leads to the idea of *kernel density estimators*. All these need regularity properties of divergence.

3.4 Variational characterizations of divergence: Gelfand-Yaglom-Perez

The point of the following theorem is that divergence on general alphabets can be defined via divergence on finite alphabets and discretization. Moreover, as the quantization becomes finer, we approach the value of divergence.

Theorem 3.7 (Gelfand-Yaglom-Perez)**.** *Let P, Q be two probability measures on $\mathcal{X}$ with σ-algebra $\mathcal{F}$. Then*

$$D(P\|Q) = \sup_{\{E_1,\ldots,E_n\}} \sum_{i=1}^{n} P[E_i] \log \frac{P[E_i]}{Q[E_i]}, \tag{3.3}$$

where the supremum is over all finite $\mathcal{F}$-measurable partitions: $\bigcup_{j=1}^{n} E_j = \mathcal{X}, E_j \cap E_i = \varnothing$, and $0 \log \frac{1}{0} = 0$ and $\log \frac{1}{0} = \infty$ per our usual convention.

Remark 3.1. This theorem, in particular, allows us to prove all general identities and inequalities for the cases of discrete random variables.

Proof. "$\geq$": Fix a finite partition $E_1, \ldots E_n$. Define a function (quantizer/discretizer) $f : \mathcal{X} \to \{1, \ldots, n\}$ as follows: For any x, let $f(x)$ denote the index j of the set E_j to which X belongs. Let X be distributed according to either P or Q and set $Y = f(X)$. Applying data processing inequality for divergence yields

$$\begin{aligned} D(P\|Q) &= D(P_X\|Q_X) \\ &\geq D(P_Y\|Q_Y) \\ &= \sum_i P[E_i] \log \frac{P[E_i]}{Q[E_i]}. \end{aligned} \tag{3.4}$$

"$\leq$": To show $D(P\|Q)$ is indeed achievable, first note that if $P \not\ll Q$, then by definition, there exists B such that $Q(B) = 0 < P(B)$. Choosing the partition $E_1 = B$ and $E_2 = B^c$, we have $D(P\|Q) = \infty = \sum_{i=1}^{2} P[E_i] \log \frac{P[E_i]}{Q[E_i]}$. In the sequel we assume that $P \ll Q$, hence the likelihood ratio $\frac{dP}{dQ}$ is well-defined. Let us define a partition of $\mathcal{X}$ by partitioning the range of $\log \frac{dP}{dQ}$: $E_j = \{x : \log \frac{dP}{dQ} \in \epsilon \cdot [j - n/2, j + 1 - n/2)\}$, $j = 1, \ldots, n-1$ and $E_n = \{x : \log \frac{dP}{dQ} < 1 - n/2 \text{ or } \log \frac{dP}{dQ} \geq n/2)\}$.[1]Note that on E_j, $\log \frac{dP}{dQ} \leq \epsilon(j + 1 - n/2) \leq \log \frac{P(E_j)}{Q(E_j)} + \epsilon$. Hence $\sum_{j=1}^{n-1} \int_{E_j} dP \log \frac{dP}{dQ} \leq \sum_{j=1}^{n-1} \epsilon P(E_j) + P(E_j) \log \frac{P(E_j)}{Q(E_j)} \leq \epsilon + \sum_{j=1}^{n} \epsilon P(E_j) + P(E_j) \log \frac{P(E_j)}{Q(E_j)} + P(E_n) \log \frac{1}{P(E_n)}$. In other words, $\sum_{j=1}^{n} P(E_j) \log \frac{P(E_j)}{Q(E_j)} \geq \int_{E_n^c} dP \log \frac{dP}{dQ} - \epsilon - P(E_n) \log \frac{1}{P(E_n)}$. Let $n \to \infty$ and $\epsilon \to 0$ be such that $n\epsilon \to \infty$ (e.g., $\epsilon = 1/\sqrt{n}$). The proof is complete by noting that $P(E_n) \to 0$ and $\int \mathbf{1}_{\left\{|\log \frac{dP}{dQ}| \leq \epsilon n\right\}} dP \log \frac{dP}{dQ} \xrightarrow{\epsilon n \uparrow \infty} \int dP \log \frac{dP}{dQ} = D(P\|Q)$. □

3.5 Continuity of divergence. Dependence on σ-algebra.

For finite alphabet $\mathcal{X}$ it is easy to establish continuity of entropy and divergence:

Proposition 3.1. *Let $\mathcal{X}$ be finite, fix distribution Q on $\mathcal{X}$ with $Q(x) > 0$ for all $x \in \mathcal{X}$. Then map*

$$P \mapsto D(P\|Q)$$

is continuous. In particular,

$$P \mapsto H(P) \tag{3.5}$$

is continuous.

[1] *Intuition*: The main idea is to note that the loss in the inequality (3.4) is in fact $D(P_X\|Q_X) = D(P_Y\|Q_Y) + D(P_{X|Y}\|Q_{X|Y}|P_Y)$, and we want to show that the conditional divergence is small. Note that $P_{X|Y=j} = P_{X|X\in E_j}$ and $Q_{X|Y=j} = Q_{X|X\in E_j}$. Hence $\frac{dP_{X|Y=j}}{dQ_{X|Y=j}} = \frac{dP}{dQ} \frac{Q(E_j)}{P(E_j)} \mathbf{1}_{E_j}$. Once we partitioned the likelihood ratio sufficiently finely, these two conditional distribution are very close to each other.

Warning: Divergence is never continuous in the pair, even for finite alphabets: $d(\frac{1}{n}\|2^{-n}) \not\to 0$.

Proof. Notice that

$$D(P\|Q) = \sum_x P(x)\log\frac{P(x)}{Q(x)}$$

and each term is a continuous function of $P(x)$. □

Our next goal is to study continuity properties of divergence for general alphabets. First, however, we need to understand dependence on the σ-algebra of the space. Indeed, divergence $D(P\|Q)$ implicitly depends on the σ-algebra $\mathcal{F}$ defining the measurable space $(\mathcal{X},\mathcal{F})$. To emphasize the dependence on $\mathcal{F}$ we will write

$$D(P_{\mathcal{F}}\|Q_{\mathcal{F}})\,.$$

We want to understand how does $D(P_{\mathcal{F}}\|Q_{\mathcal{F}})$ depend upon refining $\mathcal{F}$. Notice that we can even define $D(P_{\mathcal{F}}\|Q_{\mathcal{F}})$ for any *algebra* of sets $\mathcal{F}$ and two positive additive set-functions P, Q on $\mathcal{F}$. For this we take (3.3) as the definition. Note that when $\mathcal{F}$ is not a σ-algebra or P, Q are not σ-additive, we do not have Radon-Nikodym theorem and thus our original definition is not applicable.

Corollary 3.2 (Measure-theoretic properties of divergence)**.** *Let P, Q be probability measures on the measurable space $(\mathcal{X},\mathcal{H})$. Assume all algebras below are sub-algebras of $\mathcal{H}$. Then:*

- *(Monotonicity) If $\mathcal{F}\subseteq\mathcal{G}$ then*

$$D(P_{\mathcal{F}}\|Q_{\mathcal{F}}) \le D(P_{\mathcal{G}}\|Q_{\mathcal{G}})\,. \tag{3.6}$$

- *Let $\mathcal{F}_1\subseteq\mathcal{F}_2\ldots$ be an increasing sequence of algebras and let $\mathcal{F}=\bigcup_n\mathcal{F}_n$ be their limit, then*

$$D(P_{\mathcal{F}_n}\|Q_{\mathcal{F}_n}) \nearrow D(P_{\mathcal{F}}\|Q_{\mathcal{F}})\,.$$

- *If $\mathcal{F}$ is $(P+Q)$-dense in $\mathcal{G}$ then*[2]

$$D(P_{\mathcal{F}}\|Q_{\mathcal{F}}) = D(P_{\mathcal{G}}\|Q_{\mathcal{G}})\,. \tag{3.7}$$

- *(Monotone convergence theorem) Let $\mathcal{F}_1\subseteq\mathcal{F}_2\ldots$ be an increasing sequence of algebras and let $\mathcal{F}=\bigvee_n\mathcal{F}_n$ be the σ-algebra generated by them, then*

$$D(P_{\mathcal{F}_n}\|Q_{\mathcal{F}_n}) \nearrow D(P_{\mathcal{F}}\|Q_{\mathcal{F}})\,.$$

  In particular,

$$D(P_{X^\infty}\|Q_{X^\infty}) = \lim_{n\to\infty} D(P_{X^n}\|Q_{X^n})\,.$$

- *(Lower-semicontinuity of divergence) If $P_n\to P$ and $Q_n\to Q$ pointwise on the algebra $\mathcal{F}$, then*[3]

$$D(P_{\mathcal{F}}\|Q_{\mathcal{F}}) \le \liminf_{n\to\infty} D(P_{n,\mathcal{F}}\|Q_{n,\mathcal{F}})\,. \tag{3.8}$$

Proof. Straightforward applications of (3.3) and the observation that any algebra $\mathcal{F}$ is μ-dense in the σ-algebra $\sigma\{\mathcal{F}\}$ it generates, for any μ on $(\mathcal{X},\mathcal{H})$.[4] □

Note: Pointwise convergence on $\mathcal{H}$ is weaker than convergence in total variation and stronger than convergence in distribution (aka "weak convergence"). However, (3.8) can be extended to this mode of convergence (see Theorem 3.6).

[2] Note: $\mathcal{F}$ is μ-dense in $\mathcal{G}$ if $\forall E\in\mathcal{G}, \epsilon>0 \exists E'\in\mathcal{F}$ s.t. $\mu[E\Delta E']\le\epsilon$.

[3] $P_n\to P$ pointwise on some algebra $\mathcal{F}$ if $\forall E\in\mathcal{F}: P_n[E]\to P[E]$.

[4] This may be shown by transfinite induction: to each ordinal ω associate an algebra $\mathcal{F}_\omega$ generated by monotone limits of sets from $\mathcal{F}_{\omega'}$ with $\omega'<\omega$. Then $\sigma\{\mathcal{F}\}=\mathcal{F}_{\omega_0}$, where ω_0 is the first ordinal for which $\mathcal{F}_\omega$ is a monotone class. But $\mathcal{F}$ is μ-dense in each $\mathcal{F}_\omega$ by transfinite induction.

3.6 Variational characterizations and continuity of mutual information

Again, similarly to Proposition 3.1, it is easy to show that in the case of finite alphabets mutual information is continuous in the distribution:

Proposition 3.2. *Let $\mathcal{X}$ and $\mathcal{Y}$ be finite alphabets. Then*

$$P_{X,Y} \mapsto I(X;Y)$$

is continuous.

Proof. Apply representation

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

and (3.5). □

Further properties of mutual information follow from $I(X;Y) = D(P_{XY}\|P_X P_Y)$ and corresponding properties of divergence, e.g.

1. $$I(X;Y) = \sup_f \mathbb{E}[f(X,Y)] - \log \mathbb{E}[\exp\{f(X,\bar{Y})\}]\,,$$

   where $\bar{Y}$ is a copy of Y, independent of X and supremum is over bounded, or even bounded continuous functions.

2. If $(X_n, Y_n) \xrightarrow{d} (X,Y)$ converge in distribution, then

   $$I(X;Y) \le \liminf_{n\to\infty} I(X_n;Y_n)\,. \tag{3.9}$$

   Good example of strict inequality: $X_n = Y_n = \frac{1}{n}Z$. In this case $(X_n,Y_n) \xrightarrow{d} (0,0)$ but $I(X_n;Y_n) = H(Z) > 0 = I(0;0)$.

3. $$I(X;Y) = \sup_{\{E_i\}\times\{F_j\}} \sum_{i,j} P_{XY}[E_i \times F_j] \log \frac{P_{XY}[E_i \times F_j]}{P_X[E_i]P_Y[F_j]}\,,$$

   where supremum is over finite partitions of spaces $\mathcal{X}$ and $\mathcal{Y}$.[5]

4. (Monotone convergence):

   $$I(X^\infty;Y) = \lim_{n\to\infty} I(X^n;Y) \tag{3.10}$$
   $$I(X^\infty;Y^\infty) = \lim_{n\to\infty} I(X^n;Y^n) \tag{3.11}$$

   This implies that all mutual information between two-processes X^∞ and Y^∞ is contained in their finite-dimensional projections, leaving nothing for the tail σ-algebra.

[5] To prove this from (3.3) one needs to notice that algebra of measurable rectangles is dense in the product σ-algebra.

6.441 Information Theory

Spring 2016