4.1 Convexity of information measures

Theorem 4.1. $(P,Q) \mapsto D(P\|Q)$ *is convex.*

Proof. First proof: Let $X \in \{0,1\}$, $P_X = [\lambda, 1-\lambda]$. Select two conditional kernels:

$$P_{Y|X=0} = P_0, \quad P_{Y|X=1} = P_1 \tag{4.1}$$
$$Q_{Y|X=0} = Q_0, \quad Q_{Y|X=1} = Q_1 \tag{4.2}$$

Conditioning increases divergence, hence

$$D(P_{Y|X}\|Q_{Y|X}|P_X) \geq D(P_Y\|Q_Y)$$

Second proof: $(p,q) \to p\log\frac{p}{q}$ is convex on $\mathbb{R}^2_+$ [Verify by computing the Hessian matrix and showing that it is positive semidefinite][1]

Third proof: By the Donsker-Varadhan variational representation,

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

where for fixed f, $P \to \mathbb{E}_P[f(X)]$ is affine (hence convex), $Q \mapsto \log\mathbb{E}_Q[\exp\{f(X)\}]$ is concave. Therefore $(P,Q) \mapsto D(P\|Q)$ is pointwise supremum of convex functions, hence convex. □

Remark 4.1. The first proof shows that for an arbitrary measure of similarity $\mathcal{D}(P\|Q)$ convexity of $(P,Q) \mapsto \mathcal{D}(P\|Q)$ is *equivalent* to "conditioning increases divergence" property of $\mathcal{D}$. Convexity can also be understood as "mixing decreases divergence".

Remark 4.2 (f-divergences)**.** Any f-divergence, cf. (1.15), satisfies all the key properties of the usual divergence: positivity, monotonicity, data processing (DP), conditioning increases divergence (CID) and convexity in the pair. Indeed, by previous remark the last two are equivalent. Furthermore, proof of Theorem 2.2 showed that DP and CID are implied by monotonicity. Thus, consider P_{XY} and Q_{XY} and note

$$D_f(P_{XY}\|Q_{XY}) = \mathbb{E}_{Q_{XY}}\left[f\left(\frac{P_{XY}}{Q_{XY}}\right)\right] \tag{4.3}$$
$$= \mathbb{E}_{Q_Y}\,\mathbb{E}_{Q_{X|Y}}\left[f\left(\frac{P_Y}{Q_Y}\cdot\frac{P_{X|Y}}{Q_{X|Y}}\right)\right] \tag{4.4}$$
$$\geq \mathbb{E}_{Q_Y}\left[f\left(\frac{P_Y}{Q_Y}\right)\right], \tag{4.5}$$

where inequality follows by applying Jensen's inequality to convex function f. Finally, positivity follows from monotonicity by taking Y to be a constant and recalling that $f(1) = 0$.

[1] This is a general phenomenon: for a convex $f(\cdot)$ the *perspective* function $(p,q) \mapsto qf\left(\frac{p}{q}\right)$ is convex too.

Theorem 4.2 (Entropy)**.** *$P_X \mapsto H(P_X)$ is concave.*

Proof. If P_X is on a finite alphabet, then proof is complete by $H(X) = \log|\mathcal{X}| - D(P_X\|U_X)$. Otherwise, set

$$P_{X|Y} = \begin{cases} P_0 & Y = 0 \\ P_1 & Y = 1 \end{cases}, \; P_Y(Y=0) = \lambda$$

Then apply $H(X|Y) \le H(X)$. □

Recall that $I(X,Y)$ is a function of P_{XY}, or equivalently, $(P_X, P_{Y|X})$. Denote $I(P_X, P_{Y|X}) = I(X;Y)$.

Theorem 4.3 (Mutual Information)**.**

- *For fixed $P_{Y|X}$, $P_X \mapsto I(P_X, P_{Y|X})$ is concave.*
- *For fixed P_X, $P_{Y|X} \mapsto I(P_X, P_{Y|X})$ is convex.*

Proof.

- *First proof*: Introduce $\theta \in \mathrm{Bern}(\lambda)$. Define $P_{X|\theta=0} = P_X^0$ and $P_{X|\theta=1} = P_X^1$. Then $\theta \to X \to Y$. Then $P_X = \bar{\lambda}P_X^0 + \lambda P_X^1$. $I(X;Y) = I(X,\theta;Y) = I(\theta;Y) + I(X;Y|\theta) \ge I(X;Y|\theta)$, which is our desired $I(\bar{\lambda}P_X^0 + \lambda P_X^1, P_{Y|X}) \ge \bar{\lambda}I(P_X^0, P_{Y|X}) + \lambda I(P_X^0, P_{Y|X})$.

  Second proof: $I(X;Y) = \min_Q D(P_{Y|X}\|Q|P_X)$ – pointwise minimum of affine functions is concave.

  Third proof: Pick a Q and use the golden formula: $I(X;Y) = D(P_{Y|X}\|Q|P_X) - D(P_Y\|Q)$, where $P_X \mapsto D(P_Y\|Q)$ is convex, as the composition of the $P_X \mapsto P_Y$ (affine) and $P_Y \mapsto D(P_Y\|Q)$ (convex).
- $I(X;Y) = D(P_{Y|X}\|P_Y|P_X)$ □

4.2* Local behavior of divergence

Due to smoothness of the function $(p,q) \mapsto p\log\frac{p}{q}$ at $(1,1)$ it is natural to expect that the functional

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

should also be smooth as $P \to Q$. Due to non-negativity and convexity, it is then also natural to expect that this functional decays quadratically. In this section, we show that generally decay is sublinear and it is quadratic in the special case when $\chi^2(P\|Q) < \infty$ (see below).

Proposition 4.1. *When $D(P\|Q) < \infty$, the one-sided derivative in $\lambda = 0$ vanishes:*

$$\frac{d}{d\lambda}\Big|_{\lambda=0} D(\lambda P + \bar{\lambda}Q\|Q) = 0$$

Proof.

$$\frac{1}{\lambda}D(\lambda P + \bar{\lambda}Q\|Q) = \mathbb{E}_Q\left[\frac{1}{\lambda}(\lambda f + \bar{\lambda})\log(\lambda f + \bar{\lambda})\right]$$

where $f = \frac{dP}{dQ}$. As $\lambda \to 0$ the function under expectation decreases to $(f-1)\log e$ monotonically. Indeed, the function

$$\lambda \mapsto g(\lambda) \triangleq (\lambda f + \bar{\lambda})\log(\lambda f + \bar{\lambda})$$

is convex and equals zero at $\lambda = 0$. Thus $\frac{g(\lambda)}{\lambda}$ is increasing in λ. Moreover, by convexity of $x \mapsto x\log x$

$$\frac{1}{\lambda}(\lambda f + \overline{\lambda})(\log(\lambda f + \overline{\lambda})) \le \frac{1}{\lambda}(\lambda f \log f + \overline{\lambda}1\log 1) = f\log f$$

and by assumption $f\log f$ is Q-integrable. Thus the Monotone Convergence Theorem applies. □

Note: More generally, under suitable technical conditions,

$$\frac{d}{d\lambda}\Big|_{\lambda=0} D(\lambda P + \overline{\lambda}Q\|R) = \mathbb{E}_P\left[\log\frac{dQ}{dR}\right] - D(Q\|R)\,.$$

and

$$\frac{d}{d\lambda}\Big|_{\lambda=0} D(\bar{\lambda}P_1 + \lambda Q_1\|\bar{\lambda}P_0 + \lambda Q_0) = \mathbb{E}_{Q_1}\left[\log\frac{dP_1}{dP_0}\right] - D(P_1\|P_0) + \mathbb{E}_{P_1}\left[1 - \frac{dQ_0}{dP_0}\right]\log e$$

The message of Proposition 4.1 is that the function

$$\lambda \mapsto D(\lambda P + \bar{\lambda}Q\|Q)\,,$$

is $o(\lambda)$ as $\lambda \to 0$. In fact, in most cases it is quadratic in λ. To state a precise version, we need to define the concept of χ^2-divergence – a version of f-divergence (1.15):

$$\chi^2(P\|Q) \triangleq \int dQ\left(\frac{dP}{dQ} - 1\right)^2\,.$$

This is a very popular measure of distance between P and Q, frequently used in statistics. It has many important properties, but we will only mention that χ^2 dominates KL-divergence:

$$D(P\|Q) \le \log(1 + \chi^2(P\|Q))\,.$$

Our second result about local properties of KL-divergence is the following:

Proposition 4.2 (KL is locally χ^2-like)**.** *If $\chi^2(P\|Q) < \infty$ then*

$$D(\lambda P + \bar{\lambda}Q\|Q) = \frac{\lambda^2\log e}{2}\chi^2(P\|Q) + o(\lambda^2)\,, \qquad \lambda \to 0\,.$$

Proof. First, notice that

$$D(P\|Q) = \mathbb{E}_Q\left[g\left(\frac{dP}{dQ}\right)\right]\,,$$

where

$$g(x) \triangleq x\log x - (x-1)\log e\,.$$

Note that $x \mapsto \frac{g(x)}{(x-1)^2\log e} = \int_0^1 \frac{s\,ds}{x(1-s)+s}$ is decreasing in x on $(0,\infty)$. Therefore

$$0 \le g(x) \le (x-1)^2\log e\,,$$

and hence

$$0 \le \frac{1}{\lambda^2} g\left(\bar{\lambda} + \lambda \frac{dP}{dQ}\right) \le \left(\frac{dP}{dQ} - 1\right)^2 \log e.$$

By the dominated convergence theorem (which is applicable since $\chi^2(P\|Q) < \infty$) we have

$$\lim_{\lambda \to 0} \frac{1}{\lambda^2} \mathbb{E}_Q\left[g\left(\bar{\lambda} + \lambda \frac{dP}{dQ}\right)\right] = \frac{g''(1)}{2} \mathbb{E}_Q\left[\left(\frac{dP}{dQ} - 1\right)^2\right] = \frac{\log e}{2} \chi^2(P\|Q)\,.$$

□

4.3* Local behavior of divergence and Fisher information

Consider a parameterized set of distributions $\{P_\theta, \theta \in \Theta\}$ and assume Θ is an open subset of $\mathbb{R}^d$. Furthermore, suppose that distribution P_θ are all given in the form of

$$P_\theta(dx) = f(x|\theta)\mu(dx)\,,$$

where μ is some common dominating measure (e.g. Lebesgue or counting). If for a fixed x functions $\theta \to f(x|\theta)$ are smooth, one can define Fisher information matrix with respect to parameter θ as

$$J_F(\theta) \triangleq \mathbb{E}_{X \sim P_\theta}\left[VV^T\right], \quad V \triangleq \nabla_\theta \log f(X|\theta)\,. \tag{4.6}$$

Under suitable regularity conditions, Fisher information matrix has several equivalent expressions:

$$\begin{aligned} J_F(\theta) &= \operatorname{cov}_{X \sim P_\theta}\left[\nabla_\theta \log f(X|\theta)\right] && (4.7)\\ &= (4 \log e) \int \mu(dx)(\nabla_\theta \sqrt{f(x|\theta)})(\nabla_\theta \sqrt{f(x|\theta)})^T && (4.8)\\ &= -(\log e)\, \mathbb{E}_\theta[\operatorname{Hess}_\theta(\log f(X|\theta))]\,, && (4.9) \end{aligned}$$

where the latter is obtained by differentiating

$$0 = \int \mu(dx) f(x|\theta) \frac{\partial}{\partial \theta_i} \log f(x|\theta)$$

in θ_j.

Trace of this matrix is called Fisher information and similarly can be expressed in a variety of forms:

$$\begin{aligned} \operatorname{tr} J_F(\theta) &= \int \mu(dx) \frac{\|\nabla_\theta f(x|\theta)\|^2}{f(x|\theta)} && (4.10)\\ &= 4 \int \mu(dx) \|\nabla_\theta \sqrt{f(x|\theta)}\|^2 && (4.11)\\ &= -(\log e) \cdot \mathbb{E}_{X \sim P_\theta}\left[\sum_{i=1}^d \frac{\partial^2}{\partial \theta_i \partial \theta_i} \log f(X|\theta)\right], && (4.12) \end{aligned}$$

Significance of Fisher information matrix arises from the fact that it gauges the local behaviour of divergence for smooth parametric families. Namely, we have (again under suitable technical conditions):

$$D(P_{\theta_0}\|P_{\theta_0+\xi}) = \frac{1}{2 \log e} \xi^T J_F(\theta_0) \xi + o(\|\xi\|^2)\,, \tag{4.13}$$

which is obtained by integrating the Taylor expansion:

$$\log f(x|\theta_0+\xi) = \log f(x|\theta_0) + \xi^T \nabla_\theta \log f(x|\theta_0) + \frac{1}{2}\xi^T \mathrm{Hess}_\theta(\log f(x|\theta_0))\xi + o(\|\xi\|^2)\,.$$

Property (4.13) is of paramount importance in statistics. We should remember it as: *Divergence is locally quadratic on the parameter space, with Hessian given by the Fisher information matrix.*

Remark 4.3. It can be seen that if one introduces another parametrization $\tilde{\theta} \in \tilde{\Theta}$ by means of a smooth invertible map $\tilde{\Theta} \to \Theta$, then Fisher information matrix changes as

$$J_F(\tilde{\theta}) = A^T J_F(\theta) A\,, \tag{4.14}$$

where $A = \frac{d\theta}{d\tilde{\theta}}$ is the Jacobian of the map. So we can see that J_F transforms similarly to the metric tensor in Riemannian geometry. This idea can be used to define a Riemannian metric on the space of parameters Θ, called Fisher-Rao metric. This is explored in a field known as information geometry [AN07].

Example: Consider Θ to be the interior of a simplex of all distributions on a finite alphabet $\{0,\dots,d\}$. We will take $\theta_1,\dots,\theta_d$ as free parameters and set $\theta_0 = 1 - \sum_{i=1}^d \theta_i$. So all derivatives are with respect to $\theta_1,\dots,\theta_d$ only. Then we have

$$P_\theta(x) = f(x|\theta) = \begin{cases} \theta_x, & x = 1,\dots,d \\ 1 - \sum_{x\neq 0} \theta_x, & x = 0 \end{cases}$$

and for Fisher information matrix we get

$$J_F(\theta) = (\log^2 e)\left\{\mathrm{diag}(\frac{1}{\theta_1},\dots,\frac{1}{\theta_d}) + \frac{1}{1-\sum_{i=1}^d \theta_i} 1\cdot 1^T\right\}, \tag{4.15}$$

where $1\cdot 1^T$ is the $d\times d$ matrix of all ones. For future reference, we also compute determinant of $J_F(\theta)$. To that end notice that $\det(A + xy^T) = \det A \cdot \det(I + A^{-1}xy^T) = \det A \cdot (1 + y^T A^{-1} x)$, where we used the identity $\det(I+AB) = \det(I+BA)$. Thus, we have

$$\det J_F(\theta) = (\log e)^{2d} \prod_{x=0}^d \frac{1}{\theta_x} = (\log e)^{2d} \frac{1}{1-\sum_{x=1}^d \theta_x} \prod_{x=1}^d \frac{1}{\theta_x}\,. \tag{4.16}$$

4.4 Extremization of mutual information

Two problems of interest

- Fix $P_{Y|X} \to \max_{P_X} I(X;Y)$ — channel coding
  Note: This maximum is called "capacity" of a set of distributions $\{P_{Y|X=x}, x\in\mathcal{X}\}$.
- Fix $P_X \to \min_{P_{Y|X}} I(X;Y)$ — lossy compression

Theorem 4.4 (Saddle point)**.** *Let $\mathcal{P}$ be a convex set of distributions on $\mathcal{X}$. Suppose there exists $P_X^* \in \mathcal{P}$ such that*

$$\sup_{P_X\in\mathcal{P}} I(P_X, P_{Y|X}) = I(P_X^*, P_{Y|X}) \triangleq C$$

and let $P_X^* \xrightarrow{P_{Y|X}} P_Y^*$. Then for all $P_X \in \mathcal{P}$ and for all Q_Y, we have

$$D(P_{Y|X}\|P_Y^*|P_X) \le D(P_{Y|X}\|P_Y^*|P_X^*) \le D(P_{Y|X}\|Q_Y|P_X^*). \tag{4.17}$$

Note: P_X^* (resp., P_Y^*) is called a capacity-achieving input (resp., output) distribution, or a *caid* (resp., the *caod*).

Proof. Right inequality: obvious from $C = I(P_X^*, P_{Y|X}) = \min_{Q_Y} D(P_{Y|X}\|Q_Y|P_X^*)$.

Left inequality: If $C = \infty$, then trivial. In the sequel assume that $C < \infty$, hence $I(P_X, P_{Y|X}) < \infty$ for all $P_X \in \mathcal{P}$. Let $P_{X_\lambda} = \lambda P_X + \bar{\lambda} P_X^* \in \mathcal{P}$ by convexity of $\mathcal{P}$, and introduce $\theta \sim \mathrm{Bern}(\lambda)$, so that $P_{X_\lambda|\theta=0} = P_X^*$, $P_{X_\lambda|\theta=1} = P_X$, and $\theta \to X_\lambda \to Y_\lambda$. Then

$$\begin{aligned} C \ge I(X_\lambda; Y_\lambda) &= I(\theta, X_\lambda; Y_\lambda) = I(\theta; Y_\lambda) + I(X_\lambda; Y_\lambda|\theta) \\ &= D(P_{Y_\lambda|\theta}\|P_{Y_\lambda}|P_\theta) + \lambda I(P_X, P_{Y|X}) + \bar{\lambda} C \\ &= \lambda D(P_Y\|P_{Y_\lambda}) + \bar{\lambda} D(P_Y^*\|P_{Y_\lambda}) + \lambda I(P_X, P_{Y|X}) + \bar{\lambda} C \\ &\ge \lambda D(P_Y\|P_{Y_\lambda}) + \lambda I(P_X, P_{Y|X}) + \bar{\lambda} C. \end{aligned}$$

Since $I(P_X, P_{Y|X}) < \infty$, we can subtract it to obtain

$$\lambda(C - I(P_X, P_{Y|X})) \ge \lambda D(P_Y\|P_{Y_\lambda}).$$

Dividing both sides by λ, taking the lim inf and using lower semicontinuity of D, we have

$$\begin{aligned} & C - I(P_X, P_{Y|X}) \ge \liminf_{\lambda\to 0} D(P_Y\|P_{Y_\lambda}) \ge D(P_Y\|P_Y^*) \\ \Longrightarrow\ & C \ge I(P_X, P_{Y|X}) + D(P_Y\|P_Y^*) = D(P_{Y|X}\|P_Y|P_X) + D(P_Y\|P_Y^*) = D(P_{Y|X}\|P_Y^*|P_X). \end{aligned}$$

Here is an even shorter proof:

$$\begin{aligned} C \ge I(X_\lambda; Y_\lambda) &= D(P_{Y|X}\|P_{Y_\lambda}|P_{X_\lambda}) && (4.18)\\ &= \lambda D(P_{Y|X}\|P_{Y_\lambda}|P_X) + \bar{\lambda} D(P_{Y|X}\|P_{Y_\lambda}|P_X^*) && (4.19)\\ &\ge \lambda D(P_{Y|X}\|P_{Y_\lambda}|P_X) + \bar{\lambda} C && (4.20)\\ &= \lambda D(P_{X,Y}\|P_X P_{Y_\lambda}) + \bar{\lambda} C\,, && (4.21) \end{aligned}$$

where inequality is by the right part of (4.17) (already shown). Thus, subtracting $\bar{\lambda} C$ and dividing by λ we get

$$D(P_{X,Y}\|P_X P_{Y_\lambda}) \le C$$

and the proof is completed by taking $\liminf_{\lambda\to 0}$ and applying lower semincontinuity of divergence. □

Corollary 4.1. *In addition to the assumptions of Theorem 4.4, suppose $C < \infty$. Then caod P_Y^* is unique. It satisfies the property that for any P_Y induced by some $P_X \in \mathcal{P}$ (i.e. $P_Y = P_{Y|X} \circ P_X$) we have*

$$D(P_Y\|P_Y^*) \le C < \infty \tag{4.22}$$

and in particular $P_Y \ll P_Y^*$.

Proof. The statement is: $I(P_X, P_{Y|X}) = C \Rightarrow P_Y = P_Y^*$. Indeed:

$$\begin{aligned} C = D(P_{Y|X}\|P_Y|P_X) &= D(P_{Y|X}\|P_Y^*|P_X) - D(P_Y\|P_Y^*) \\ &\le D(P_{Y|X}\|P_Y^*|P_X^*) - D(P_Y\|P_Y^*) \\ &= C - D(P_Y\|P_Y^*) \Rightarrow P_Y = P_Y^* \end{aligned}$$

Statement (4.22) follows from the left inequality in (4.17) and "conditioning increases divergence". □

Notes:

- Finiteness of C is necessary. Counterexample: The identity channel $Y = X$, where X takes values on integers. Then any distribution with infinite entropy is caid or caod.
- *Non-uniqueness of caid.* Unlike the caod, caid does not need to be unique. Let $Z_1 \sim \text{Bern}(\frac{1}{2})$. Consider $Y_1 = X_1 \oplus Z_1$ and $Y_2 = X_2$. Then $\max_{P_{X_1X_2}} I(X_1, X_2; Y_1, Y_2) = \log 4$, achieved by $P_{X_1X_2} = \text{Bern}(p) \times \text{Bern}(\frac{1}{2})$ for any p. Note that the *caod* is unique: $P^*_{Y_1Y_2} = \text{Bern}(\frac{1}{2}) \times \text{Bern}(\frac{1}{2})$.

Review: Minimax and saddlepoint

Suppose we have a bivariate function f. Then we always have the *minimax inequality*:

$$\inf_y \sup_x f(x,y) \geq \sup_x \inf_y f(x,y).$$

When does it hold with equality?

1. It turns out minimax equality is implied by the existence of a saddle point (x^*, y^*), i.e.,
$$f(x, y^*) \leq f(x^*, y^*) \leq f(x^*, y) \qquad \forall x, y$$
Furthermore, minimax equality also implies existence of saddle point if inf and sup are achieved c.f. [BNO03, Section 2.6]) for all x, y [Straightforward to check. See proof of corollary below].
2. There are a number of known criteria establishing
$$\inf_y \sup_x f(x,y) = \sup_x \inf_y f(x,y)$$
They usually require some continuity of f, compactness of domains and convexity in x and concavity in y. One of the most general version is due to M. Sion [Sio58].
3. The mother result of all this minimax theory is a theorem of von Neumann on bilinear functions: Let A and B have finite alphabets, and $g(a, b)$ be arbitrary, then
$$\min_{P_A} \max_{P_B} \mathbb{E}[g(A,B)] = \max_{P_B} \min_{P_A} \mathbb{E}[g(A,B)]$$
Here $(x, y) \leftrightarrow (P_A, P_B)$ and $f(x, y) \leftrightarrow \sum_{a,b} P_A(a)P_B(b)g(a, b)$.
4. A more general version is: if $\mathcal{X}$ and $\mathcal{Y}$ are compact convex domains in $\mathbb{R}^n$, $f(x, y)$ continuous in (x, y), concave in x and convex in y then
$$\max_{x \in \mathcal{X}} \min_{y \in \mathcal{Y}} f(x,y) = \min_{y \in \mathcal{Y}} \max_{x \in \mathcal{X}} f(x,y)$$

Applying Theorem 4.4 to conditional divergence gives the following result.

Corollary 4.2 (Minimax)**.** *Under assumptions of Theorem 4.4, we have*

$$\begin{aligned} \max_{P_X \in \mathcal{P}} I(X;Y) &= \max_{P_X \in \mathcal{P}} \min_{Q_Y} D(P_{Y|X} \| Q_Y | P_X) \\ &= \min_{Q_Y} \max_{P_X \in \mathcal{P}} D(P_{Y|X} \| Q_Y | P_X) \end{aligned}$$

Proof. This follows from saddle-point trivially: Maximizing/minimizing the leftmost/rightmost sides of (4.17) gives

$$\min_{Q_Y}\max_{P_X\in\mathcal{P}} D(P_{Y|X}\|Q_Y|P_X) \le \max_{P_X\in\mathcal{P}} D(P_{Y|X}\|P_Y^*|P_X) \le D(P_{Y|X}\|P_Y^*|P_X^*)$$
$$\le \min_{Q_Y} D(P_{Y|X}\|Q_Y|P_X^*) \le \max_{P_X\in\mathcal{P}}\min_{Q_Y} D(P_{Y|X}\|Q_Y|P_X).$$

but by definition $\min\max \ge \max\min$. □

4.5 Capacity = information radius

Review: Radius and diameter

Let (X,d) be a metric space. Let A be a bounded subset.

1. *Radius* (aka Chebyshev radius) of A: the radius of the smallest ball that covers A, i.e., $\operatorname{rad}(A) = \inf_{y\in X}\sup_{x\in A} d(x,y)$.
2. *Diameter* of A: $\operatorname{diam}(A) = \sup_{x,y\in A} d(x,y)$.
3. Note that the radius and the diameter both measure how big/rich a set is.
4. From definition and triangle inequality we have
$$\frac{1}{2}\operatorname{diam}(A) \le \operatorname{rad}(A) \le \operatorname{diam}(A)$$
5. In fact, the rightmost upper bound can frequently be improved. A result of Bohnenblust [Boh38] shows that in $\mathbb{R}^n$ equipped with any norm we always have $\operatorname{rad}(A) \le \frac{n}{n+1}\operatorname{diam}(A)$. For $\mathbb{R}^n$ with ℓ_∞-norm the situation is even simpler: $\operatorname{rad}(A) = \frac{1}{2}\operatorname{diam}(A)$ (such spaces are called centrable).

The next simple corollary shows that capacity is just the radius of the set of distributions $\{P_{Y|X=x}, x\in\mathcal{X}\}$ when distances are measured by divergence (although, we remind, divergence is not a metric).

Corollary 4.3. *For fixed kernel $P_{Y|X}$, let $\mathcal{P} = \{$all dist. on $\mathcal{X}\}$ and $\mathcal{X}$ is finite, then*

$$\max_{P_X} I(X;Y) = \max_x D(P_{Y|X=x}\|P_Y^*)$$
$$= D(P_{Y|X=x}\|P_Y^*) \qquad \forall x: P_X^*(x) > 0\ .$$

The last corollary gives a geometric interpretation to capacity: it equals the radius of the smallest divergence-"ball" that encompasses all distributions $\{P_{Y|X=x} : x\in\mathcal{X}\}$. Moreover, P_Y^* is a convex combination of some $P_{Y|X=x}$ and it is **equidistant** to those.

4.6 Existence of caod (general case)

We have shown above that the solution to

$$C = \sup_{P_X\in\mathcal{P}} I(X;Y)$$

can be a) interpreted as a saddle point; b) written in the minimax form and c) that caod P_Y^* is unique. This was all done under the extra assumption that supremum over P_X is attainable. It turns out, properties b) and c) can be shown without that extra assumption.

Theorem 4.5 (Kemperman)**.** *For any $P_{Y|X}$ and a convex set of distributions $\mathcal{P}$ such that*

$$C = \sup_{P_X \in \mathcal{P}} I(P_X, P_{Y|X}) < \infty \tag{4.23}$$

there exists a unique P_Y^* with the property that

$$C = \sup_{P_X \in \mathcal{P}} D(P_{Y|X} \| P_Y^* | P_X) \,. \tag{4.24}$$

Furthermore,

$$\begin{aligned} C &= \sup_{P_X \in \mathcal{P}} \min_{Q_Y} D(P_{Y|X} \| Q_Y | P_X) && (4.25)\\ &= \min_{Q_Y} \sup_{P_X \in \mathcal{P}} D(P_{Y|X} \| Q_Y | P_X) && (4.26)\\ &= \min_{Q_Y} \sup_{x \in \mathcal{X}} D(P_{Y|X=x} \| Q_Y) \,, \qquad \textit{(if } \mathcal{P} = \{\textit{all } P_X\}\textit{.)} && (4.27) \end{aligned}$$

Note: Condition (4.23) is automatically satisfied if there is any Q_Y such that

$$\sup_{P_X \in \mathcal{P}} D(P_{Y|X} \| Q_Y | P_X) < \infty \,. \tag{4.28}$$

Example: *Non-existence of caid.* Let $Z \sim \mathcal{N}(0,1)$ and consider the problem

$$C = \sup_{P_X : \substack{\mathbb{E}[X]=0, \mathbb{E}[X^2]=P \\ \mathbb{E}[X^4]=s}} I(X; X+Z) \,. \tag{4.29}$$

If we remove the constraint $\mathbb{E}[X^4] = s$ the unique caid is $P_X = \mathcal{N}(0,P)$, see Theorem 4.6. When $s \neq 3P^2$ then such P_X is no longer inside the constraint set $\mathcal{P}$. However, for $s > 3P^2$ the maximum

$$C = \frac{1}{2}\log(1+P)$$

is still attainable. Indeed, we can add a small "bump" to the gaussian distribution as follows:

$$P_X = (1-p)\mathcal{N}(0,P) + p\delta_x \,,$$

where $p \to 0$, $px^2 \to 0$ but $px^4 \to s - 3P^2 > 0$. This shows that for the problem (4.29) with $s > 3P^2$ the caid does not exist, the caod $P_Y^* = \mathcal{N}(0, 1+P)$ exists and unique as Theorem 4.5 postulates.

Proof of Theorem 4.5. Let P'_{X_n} be a sequence of input distributions achieving C, i.e., $I(P'_{X_n}, P_{Y|X}) \to C$. Let $\mathcal{P}_n$ be the convex hull of $\{P'_{X_1}, \ldots, P'_{X_n}\}$. Since $\mathcal{P}_n$ is a finite-dimensional simplex, the concave function $P_X \mapsto I(P_X, P_{Y|X})$ attains its maximum at some point $P_{X_n} \in \mathcal{P}_n$, i.e.,

$$I_n \triangleq I(P_{X_n}, P_{Y|X}) = \max_{P_X \in \mathcal{P}_n} I(P_X, P_{Y|X}) \,.$$

Denote by P_{Y_n} be the sequence of output distributions corresponding to P_{X_n}. We have then:

$$\begin{aligned} D(P_{Y_n} \| P_{Y_{n+k}}) &= D(P_{Y|X} \| P_{Y_{n+k}} | P_{X_n}) - D(P_{Y|X} \| P_{Y_n} | P_{X_n}) && (4.30)\\ &\le I(P_{X_{n+k}}, P_{Y|X}) - I(P_{X_n}, P_{Y|X}) && (4.31)\\ &\le C - I_n \,, && (4.32) \end{aligned}$$

where in (4.31) we applied Theorem 4.4 to $(\mathcal{P}_{n+k}, P_{Y_{n+k}})$. By the Pinsker-Csiszár inequality (1.14) and since $I_n \nearrow C$, we conclude that the sequence P_{Y_n} is Cauchy in total variation:

$$\sup_{k\ge 1} \mathrm{TV}(P_{Y_n}, P_{Y_{n+k}}) \to 0\,, \qquad n\to\infty\,.$$

Since the space of probability distributions is complete in total variation, the sequence must have a limit point $P_{Y_n} \to P_Y^*$. By taking a limit as $k\to\infty$ in (4.32) and applying the lower semi-continuity of divergence (Theorem 3.6) we get

$$D(P_{Y_n}\|P_Y^*) \le \lim_{k\to\infty} D(P_{Y_n}\|P_{Y_{n+k}}) \le C - I_n\,,$$

and therefore, $P_{Y_n}\to P_Y^*$ in the (stronger) sense of $D(P_{Y_n}\|P_Y^*)\to 0$. Therefore,

$$D(P_{Y|X}\|P_Y^*|P_{X_n}) = I_n + D(P_{Y_n}\|P_Y^*) \to C\,. \tag{4.33}$$

Take any $P_X \in \bigcup_{k\ge 1}\mathcal{P}_k$. Then $P_X\in\mathcal{P}_n$ for all sufficiently large n and thus by Theorem 4.4

$$D(P_{Y|X}\|P_{Y_n}|P_X) \le I_n \le C\,, \tag{4.34}$$

which by lower semi-continuity of divergence implies

$$D(P_{Y|X}\|P_Y^*|P_X) \le C\,. \tag{4.35}$$

Finally, to prove that (4.35) holds for arbitrary $P_X\in\mathcal{P}$, we may repeat the argument above with $\mathcal{P}_n$ replaced by $\tilde{\mathcal{P}}_n = \mathrm{conv}(P_X\cup\mathcal{P}_n)$, denoting the resulting sequences by $\tilde{P}_{X_n}, \tilde{P}_{Y_n}$ and the limit point by $\tilde{P}_Y^*$ we have:

$$\begin{aligned} D(P_{Y_n}\|\tilde{P}_{Y_n}) &= D(P_{Y|X}\|\tilde{P}_{Y_n}|P_{X_n}) - D(P_{Y|X}\|P_{Y_n}|P_{X_n}) &(4.36)\\ &\le C - I_n\,, &(4.37)\end{aligned}$$

where (4.37) follows from (4.35) since $P_{X_n}\in\tilde{\mathcal{P}}_n$. Hence taking limit as $n\to\infty$ we have $\tilde{P}_Y^* = P_Y^*$ and therefore (4.35) holds.

Finally, to see (4.26), note that by definition capacity as a max-min is at most the min-max, i.e.,

$$C = \sup_{P_X\in\mathcal{P}}\min_{Q_Y} D(P_{Y|X}\|Q_Y|P_X) \le \min_{Q_Y}\sup_{P_X\in\mathcal{P}} D(P_{Y|X}\|Q_Y|P_X) \le \sup_{P_X\in\mathcal{P}} D(P_{Y|X}\|P_Y^*|P_X) = C$$

in view of (4.34). □

Corollary 4.4. *Let $\mathcal{X}$ be countable and $\mathcal{P}$ a convex set of distributions on $\mathcal{X}$. If $\sup_{P_X\in\mathcal{P}} H(X) < \infty$ then*

$$\sup_{P_X\in\mathcal{P}} H(X) = \min_{Q_X}\sup_{P_X\in\mathcal{P}} \sum_x P_X(x)\log\frac{1}{Q_X(x)} < \infty$$

and the optimizer Q_X^* exist and is unique. If $Q_X^*\in\mathcal{P}$ then it is also a unique maximizer of $H(X)$.

Proof. Just apply Kemperman's result to channel $Y = X$. □

Example: Assume that $f:\mathbb{Z}\to\mathbb{R}$ is such that $\sum_{n\in\mathbb{Z}}\exp\{-\lambda f(n)\} < \infty$ for all $\lambda > 0$. Then

$$\max_{X:\mathbb{E}[f(X)]\le a} H(X) \le \inf_{\lambda>0} \lambda a + \log\sum_n \exp\{-\lambda f(n)\}\,.$$

This follows from taking $Q(n) = c\exp\{-\lambda f(n)\}$. This bound is often tight and achieved by $P_X(n) = c\exp\{-\lambda f(n)\}$, known as the Gibbs distribution for energy function f.

4.7 Gaussian saddle point

For additive noise, there is also a different kind of saddle point between P_X and the distribution of noise:

Theorem 4.6. *Let $X_g \sim \mathcal{N}(0, \sigma_X^2)$, $N_g \sim \mathcal{N}(0, \sigma_N^2)$, $X_g \perp\!\!\!\perp N_g$. Then:*

1. *"Gaussian capacity":*
$$C = I(X_g; X_g + N_g) = \frac{1}{2} \log \left(1 + \frac{\sigma_X^2}{\sigma_N^2}\right)$$

2. *"Gaussian input is the best": For all $X \perp\!\!\!\perp N_g$ and $\mathsf{var} X \le \sigma_X^2$,*
$$I(X; X + N_g) \le I(X_g; X_g + N_g),$$
with equality iff $X \overset{\mathrm{D}}{=} X_g$.

3. *"Gaussian noise is the worst": For for all N s.t. $\mathbb{E}[X_g N] = 0$ and $\mathbb{E}N^2 \le \sigma_N^2$,*
$$I(X_g; X_g + N) \ge I(X_g; X_g + N_g),$$
with equality iff $N \overset{\mathrm{D}}{=} N_g$ and independent of X_g.

Note: Intuitive remarks

1. For AWGN channel, Gaussian input is the most favorable. Indeed, immediately from the second statement we have
$$\max_{X: \mathsf{var} X \le \sigma_X^2} I(X; X + N_g) = \frac{1}{2} \log \left(1 + \frac{\sigma_X^2}{\sigma_N^2}\right)$$
which is the capacity formula for the AWGN channel.

2. For Gaussian source, additive Gaussian noise is the worst in the sense that it minimizes the mutual information provided by the noisy version.

Proof. WLOG, assume all random variables have zero mean. Let $Y_g = X_g + N_g$. Define
$$g(x) = D(P_{Y_g|X_g=x} \| P_{Y_g}) = D(\mathcal{N}(x, \sigma_N^2) \| \mathcal{N}(0, \sigma_X^2 + \sigma_N^2)) = \underbrace{\frac{1}{2} \log \left(1 + \frac{\sigma_X^2}{\sigma_N^2}\right)}_{=C} + \frac{\log e}{2} \frac{x^2 - \sigma_X^2}{\sigma_X^2 + \sigma_N^2}$$

1. Compute $I(X_g; X_g + N_g) = \mathbb{E}[g(X_g)] = C$

2. Recall the inf-representation $I(X;Y) = \min_Q D(P_{Y|X} \| Q | P_X)$. Then
$$I(X; X + N_g) \le D(P_{Y_g|X_g} \| P_{Y_g} | P_X) = \mathbb{E}[g(X)] \le C < \infty\,.$$
Furthermore, if $I(X; X + N_g)$ then uniqueness of caod, cf. Corollary 4.1, implies $P_Y = P_{Y_g}$. But $P_Y = P_X * \mathcal{N}(0, \sigma_N^2)$. Then it must be that $X \sim \mathcal{N}(0, \sigma_X^2)$ simply by considering characteristic functions:
$$\Psi_X(t) \cdot \mathrm{e}^{-\frac{1}{2}\sigma_N^2 t^2} = \mathrm{e}^{-\frac{1}{2}(\sigma_X^2 + \sigma_N^2) t^2} \Rightarrow \Psi_X(t) = \mathrm{e}^{-\frac{1}{2}\sigma_X^2 t^2} \Longrightarrow X \sim \mathcal{N}(0, \sigma_X^2)$$

3. Let $Y = X_g + N$ and let $P_{Y|X_g}$ be the respective kernel. [Note that here we only assume that N is *uncorrelated* with X_g, i.e., $\mathbb{E}[NX_g] = 0$, not necessarily independent.] Then

$$\begin{aligned}
I(X_g; X_g + N) &= D(P_{X_g|Y} \| P_{X_g} | P_Y) \\
&= D(P_{X_g|Y} \| P_{X_g|Y_g} | P_Y) + \mathbb{E} \log \frac{P_{X_g|Y_g}(X_g|Y)}{P_{X_g}(X_g)} \\
&\geq \mathbb{E} \log \frac{P_{X_g|Y_g}(X_g|Y)}{P_{X_g}(X_g)} && (4.38) \\
&= \mathbb{E} \log \frac{P_{Y_g|X_g}(Y|X_g)}{P_{Y_g}(Y)} && (4.39) \\
&= C + \frac{\log e}{2} \mathbb{E}\Big[\frac{Y^2}{\sigma_X^2 + \sigma_N^2} - \frac{N^2}{\sigma_N^2}\Big] && (4.40) \\
&= C + \frac{\log e}{2} \frac{\sigma_X^2}{\sigma_X^2 + \sigma_N^2}\Big(1 - \frac{\mathbb{E}N^2}{\sigma_N^2}\Big) && (4.41) \\
&\geq C, && (4.42)
\end{aligned}$$

where

- (4.39): $\frac{P_{X_g|Y_g}}{P_{X_g}} = \frac{P_{Y_g|X_g}}{P_{Y_g}}$
- (4.41): $\mathbb{E}[X_g N] = 0$ and $\mathbb{E}[Y^2] = \mathbb{E}[N^2] + \mathbb{E}[X_g^2]$.
- (4.42): $\mathbb{E}N^2 \leq \sigma_N^2$.

Finally, the conditions for equality in (4.38) say

$$D(P_{X_g|Y} \| P_{X_g|Y_g} | P_Y) = 0$$

Thus, $P_{X_g|Y} = P_{X_g|Y_g}$, i.e., X_g is conditionally Gaussian: $P_{X_g|Y=y} = \mathcal{N}(by, c^2)$ for some constant b, c. In other words, under $P_{X_g Y}$, we have

$$X_g = bY + cZ \quad , \quad Z \sim \text{Gaussian} \perp\!\!\!\perp Y.$$

But then Y must be Gaussian itself by Cramer's Theorem or simply by considering characteristic functions:

$$\Psi_Y(t) \cdot e^{ct^2} = e^{c't^2} \Rightarrow \Psi_Y(t) = e^{c''t^2} \Longrightarrow \text{Y– Gaussian}$$

Therefore, (X_g, Y) must be jointly Gaussian and hence $N = Y - X_g$ is Gaussian. Thus we conclude that it is only possible to attain $I(X_g; X_g + N) = C$ if N is Gaussian of variance σ_N^2 and independent of X_g. □

6.441 Information Theory

Spring 2016