Functionals of $p(x, y)$ invariant under marginal bijections and which satisfy the data-processing inequality

Charles Zheng and Yuval Benjamini

August 24, 2017

These are preliminary notes.

1 Motivation

The mutual information

$$\mathrm{I}(X;Y) = \mathrm{I}[p(x,y)] = \int_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \log\left(\frac{p(x,y)}{p_x(x)p_y(y)}\right) p(x)p(y)dxdy$$

and the k-class average Bayes accuracy

$$\mathrm{ABA}_k(X;Y) = \mathrm{ABA}_k[p(x,y)] = \int_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \max_{i=1}^{k} p_{x|y}(x|y_i)p_x(x)dx \prod_{i=1}^{k} p_y(y_i)dy_i$$

are two examples of functionals used to measure the strength of dependence between continuous random variables X and Y based on their joint density $p(x,y)$. Both have been used in neuroscience applications as tools for model selection. Their usefulness in this regard is due to the fact that each satisfies the data-processing inequality. A functional $\mathrm{F}[p(x,y)]$ satisfies the data-processing inequality if, given a *marginal* transition kernel Q_x or Q_y, we have

$$\mathrm{F}[p] \geq \mathrm{F}[Q_x p]$$

and

$$\mathrm{F}[p] \geq \mathrm{F}[Q_y p]$$

for all joint densities $p(x,y)$ and marginal transition kernels Q_x and Q_y. A marginal transition kernel is a probability transition kernel which operates only on x or y but not both. Marginal kernels Q_x which operate on x are defined as collections of probability distributions Q_x on $\mathcal{X}$ for each $z \in \mathcal{X}$. The kernel Q_x acts on $p(x,y)$ to transform it into the joint density $\tilde{p}(x,y)$ on $\mathcal{X} \times \mathcal{Y}$

$$\tilde{p}(x,y) = Q_x p(x,y) = \int_{\mathcal{X}} p(z,y) dQ_x(z)$$

Kernels Q_y which operate on y are defined analagously, such that Q_y is a distribution on $\mathcal{Y}$ for each $y \in \mathcal{Y}$ such that

$$\tilde{p}(x,y) = Q_y p(x,y) = \int_{\mathcal{Y}} p(x,z) dQ_y(z).$$

The question we wish to address in these notes is: what other functionals of continuous joint densities $p(x,y)$ satisfy the data-processing inequality? We call such functionals *information coefficients*.

2 Information coefficients from divergences

Generalizing the mutual information readily gives more examples. The mutual information is the KL-divergence between the join density $p(x,y)$ and the product distribution of the marginals, $p_x(x)p_y(y)$:

$$\mathrm{I}[p(x,y)] = D_{KL}(p(x,y)||p_x(x)p_y(y))$$

where

$$D_{KL}(p(x)||q(x)) = \int \log\left(\frac{p(x)}{q(x)}\right) p(x)dx.$$

However, the KL-divergence is closely related to the family of Renyi α-divergences

$$D_\alpha(p(x)||q(x)) = \frac{1}{1-\alpha} \log \int \left(\frac{p(x)}{q(x)}\right)^\alpha q(x)dx$$

for $\alpha \geq 0$ with $\alpha \neq 1$. In fact, $D_{KL} = \lim_{\alpha \to 1} D_\alpha$. Therefore, define the symmetric α-information as

$$\mathrm{I}_\alpha[p(x,y)] = D_\alpha(p(x,y)||p_x(x)p_y(y)) = \frac{1}{1-\alpha} \log \int \left(\frac{p(x,y)}{p_x(x)p_y(y)}\right)^\alpha p_x(x)p_y(y)dx.$$

As we will see, I_α also satisfies the data-processing inequality. In fact, we can generalize even beyond α-divergences. Let $D(p||q)$ be any divergence between probability distributions which satisfies the *contraction inequality*

$$D(p||q) \geq D(Qp||Qq)$$

for all densities p, q and transition kernels Q. Then define the functional I_D as the divergence between the joint density and the product density,

$$\mathrm{I}_D[p(x,y)] = D(p(x,y)||p_x(x)p_y(y)).$$

It can be easily seen that the data-processing inequality for I_D follows as a special case of the contraction inequality for D. But which divergences satisfy the contraction inequality?

A first step is to note that the contraction inequality implies invariance under bijections. This is due to the fact that any deterministic bijection is equivalent to some transition kernel: if ϕ is a bijection which acts on densities $p(x)$ by

$$\tilde{p}(x) = \phi p = \frac{p(\phi^{-1}(x))}{|\det J\phi|}$$

where $J\phi$ is the determinant of the Jacobian of ϕ, then the transition kernel Q, defined by

$$Q_x = \delta_{\phi(x)}$$

satisfies

$$\phi p = Q_x p$$

for all densities p. From bijection invariance, it follows that any divergence which satisfies the contraction inequality must be an f-divergence, that is,

$$D_f(p||q) = \int f\left(\frac{p(x)}{q(x)}\right) q(x)dx. \tag{1}$$

However, it remains to determine which functions f result in divergences which satisfy the contraction inequality.

3 The first-order contraction inequality

It will be useful to consider very *small* transition kernels, where by "small" we mean a transition kernel very close to the identity operator. For any given transition kernel Q, we can define a family of δ-shrunken kernels defined by

$$Q^\delta = (1-\delta)I + \delta Q,$$

where I is the identity operator. That is,

$$Q^\delta p = (1-\delta)p + \delta Q p. \tag{2}$$

The *first-order* contraction inequality is obtained by requiring $D(Q^\delta p||Q^\delta q) \leq D(p||q)$ for small δ. In the limit of small δ, the contraction inequality therefore becomes:

Definition (First-order contraction inequality): We say that the divergence D satisfies the first-order contraction inequality if and only if

$$\frac{d}{d\delta} D(Q^\delta p||Q^\delta q)|_{\delta=0} \leq 0$$

for all densities p and all transition kernels Q.

It is easy to see that the contraction inequality implies the first-order contraction inequality, since the contraction inequality holds for all transition kernels, including ones arbitrarily close to identity. However, we will also show that the first-order contraction inequality implies the contraction inequality given certain regularity conditions on D: for this class of divergences D, they are equivalent.

Define an *infinitely*-divisible transition kernel Q as a kernel which satisfies

$$Q = \lim_{n\to\infty} \prod_{i=1}^{n} Q^{\delta(n)}$$

for some sequence $\delta(n)$ with $\lim_{n\to\infty} \delta(n) = 0$. It is clear that the first-order contraction inequality implies that the contraction inequality holds for all infinitely-divisible kernels Q. To extend to all transition kernels in general, we have to show that any transition kernel Q can be arbitrarily well-approximated by finite products of infinitely-divisible kernels–and that this approximation carries through to the divergence $D(Qp||Qq)$ (under regularity conditions on f).

To elaborate, we propose the following strategy to establish the equivalence of first-order contraction inequality and the original contraction inequality:

1. Begin by the case where $\mathcal{X}$ is compact, and extend to general Euclidean $\mathcal{X}$ by a limiting argument.

2. Establish regularity conditions on f so that for any sequence of transition kernels $Q^{[i]}$ with
$$\lim_{i\to\infty} Q^{[i]} = Q,$$
we have
$$\lim_{i\to\infty} D_f(Q^{[i]}p||Q^{[i]}q) = D_f(Qp||Qq)$$
for all densities p and q.

3. Extend any divergence D defined for densities p, q on $\mathcal{X}$ to a divergence D defined for densities p, q on the space $\mathcal{X} \times \{0, 1\} = \mathcal{X}^0 \cup \mathcal{X}^1$. This is a technical step.

4. Show that any transition kernel Q can be arbitrarily well-approximated by *discretized* transition kernels. A discretized kernel Q has an associated finite partition of $\mathcal{X}$, $\mathcal{X} = \sqcup_{i=1}^m A_i$. The discretized kernel Q which can be written as a density $q(x, z) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ of the form
$$q(x, z) = \sum_{i=1}^{m}\sum_{j=1}^{m} q_{ij} I_{A_i}(x) I_{A_j}(z)$$
where q_{ij} are non-negative real numbers.

5. Any discretized transition kernel Q with associated partition $\{A_i\}_{i=1}^m$ and coefficients q_{ij} can be extended to a transition kernel $\tilde{Q}$ on $\mathcal{X} \times \{0, 1\}$ such that the restriction of $\tilde{Q}$ on $\mathcal{X} \times \{0\}$ is isomorphic to Q. $\tilde{Q}$ has associated partitions $\{A_i^0\}_{i=1}^m \cup \{A_i^1\}_{i=1}^m$ where $A_i^j = A_i \times \{j\}$ for $j = \{0, 1\}$. The extended kernel $\tilde{Q}$ on $\mathcal{X} \times \{0, 1\}$ can be written as the product of $m + 1$ infinitely-divible kernels,
$$\tilde{Q} = \pi V^{[m]} \cdots V^{[1]}.$$
where the π kernel is a projection from $\mathcal{X} \times \{1\}$ into $\mathcal{X} \times \{0\}$,
$$\pi_{(x,j)} = \delta_{(x,0)},$$
for all $x \in \mathcal{X}$ and $j \in \{0, 1\}$, and $V^{[i]}$ is the transition kernel with
$$V^{[i]}_{(x,j)} = \begin{cases} \delta_{(x,j)} & \text{for } x \notin A_i \text{ or } j = 1 \\ \text{given by density } \sum_{k=1}^m q_{ij} I_{A_i^1} & \text{for } (x, j) \in A_i^0 \end{cases}$$

That is, $V^{[i]}$ acts the same way on $(x, 0) \in A_m^0$ as Q does on $x \in A_m$, except that it sends those points from $\mathcal{X}^0$ (the original space) to $\mathcal{X}^1$. This means that the product $\prod_{i=1}^m V^{[i]}$ maps a density p restricted to $\mathcal{X}^0$ to the density Qp on $\mathcal{X}^1$. We can see that the splitting of $\mathcal{X}$ into the two clones $\mathcal{X}^0$ and $\mathcal{X}^1$ is done so that the individual transition kernels $V^{[i]}$, $V^{[k]}$ do not interfere with each other. Finally, π merely moves Qp back into the correct space.

6. Argue that by chaining the above approximations, any transition kernel Q can be arbitrarily well-approximated by finite products of infinitely divisible transition kernels. Since the first-order DPI implies DPI for the infinitely divisible kernels, the approximation implies DPI for Q as well.

Let us expand the first-order contraction inequality plugging in the definitions (1) and (2). The left-hand side of the first-order DPI then simplifies to

$$
\begin{aligned}
\frac{d}{d\delta} D_f(Q^\delta p || Q^\delta q) =& \frac{d}{d\delta} \int f\left(\frac{Q^\delta p(x)}{Q^\delta q(x)}\right) Q^\delta q(x) dx \\
=& \frac{d}{d\delta} \int f\left(\frac{p(x) + \delta(Qp(x) - p(x))}{q(x) + \delta(Qq(x) - q(x))}\right) (q(x) + \delta(Qq(x) - q(x)) dx \\
=& \frac{d}{d\delta} \int \left[f\left(\frac{p(x)}{q(x)}\right) + \delta f'\left(\frac{p(x)}{q(x)}\right) \frac{q(x)(Qp(x) - p(x)) - p(x)(Qq(x) - q(x))}{q(x)^2} \right] \\
& \times (q(x) + \delta(Qq(x) - q(x)) dx \\
=& \int f\left(\frac{p(x)}{q(x)}\right) (Qq(x) - q(x)) dx \\
& + \int f'\left(\frac{p(x)}{q(x)}\right) \left(Qp(x) - p(x) - \frac{p(x)(Qq(x) - q(x))}{q(x)}\right) dx \\
=& \int \left[f\left(\frac{p(x)}{q(x)}\right) - \frac{p(x)}{q(x)} f'\left(\frac{p(x)}{q(x)}\right) \right] (Qq(x) - q(x)) dx \\
& + \int f'\left(\frac{p(x)}{q(x)}\right) (Qp(x) - p(x)) dx \\
=& \int f\left(\frac{p(x)}{q(x)}\right) (Qq(x) - q(x)) + f'\left(\frac{p(x)}{q(x)}\right) \left(Qp(x) - \frac{p(x)}{q(x)} Qq(x)\right)
\end{aligned}
$$

For the special case of KL divergence, $f(x) = x\log(x)$, we have

$$\begin{aligned}
\frac{d}{d\delta}D_f(Q^\delta p||Q^\delta q) &= \int (Qp(x) - p(x))\log\left(\frac{p(x)}{q(x)}\right) + Qp(x) - \left(\frac{p(x)}{q(x)}\right)Qq(x)dx \\
&= \int \left(p(x)\log\left(\frac{p(x)}{q(x)}\right)\right)(Qq(x) - q(x)) \\
&\quad + \left(\log\left(\frac{p(x)}{q(x)}\right) + 1\right)\left(Qp(x) - \left(\frac{p(x)}{q(x)}\right)Qq(x)\right)dx
\end{aligned}$$