Unit 7: Properties of measures of information - 2

A Fano's inequality

We now prove Fano's inequality – a major and flexible tool that comes out of information theory. In fact, the proof we present uses nothing more than the data processing inequality.

Theorem 1 (Fano's inequality- version 1)**.** *Consider random variables X and Y where X is distributed uniformly over $\{1, ..., M\}$. Then, for every function $g : \mathcal{X} \to \mathcal{Y}$, we have*

$$\mathbb{P}\left(X \neq g(Y)\right) \geq 1 - \frac{I(X \wedge Y) + 1}{\log M}.$$

Proof. If X and Y were independent, any function g cannot have a probability of error less than $1 - 1/M$. Indeed, for independent X and Y, we have

$$\mathbb{P}\left(X = g(Y)\right) = \sum_y \mathrm{P}_Y\left(y\right) \mathrm{P}_X\left(g(y)\right) = \frac{1}{M} \sum_y \mathrm{P}_Y\left(y\right) = \frac{1}{M},$$

where we used the fact that X is uniform.

The main idea behind our proof is the following: For any distribution P_{XY}, the difference between the performance of g under P_{XY} and $\mathrm{P}_X \times \mathrm{P}_Y$, the independent distribution, is bounded by the "distance" between these distributions. We formalize this using the data processing inequality.

Formally, consider the channel $W : \mathcal{X} \times \mathcal{Y} \to \{0, 1\}$ given by $W(1|x, y) = \mathbb{1}_{\{x=g(y)\}}$.

Then, by the data processing inequality we get

$$D(W \circ \mathrm{P}_{XY} \| W \circ \mathrm{P}_X \mathrm{P}_Y) \leq D(\mathrm{P}_{XY} \| \mathrm{P}_X \mathrm{P}_Y) = I(X \wedge Y).$$

Further, denoting $p = \mathrm{P}_{XY}(X = g(Y))$ and $q = \mathrm{P}_X \mathrm{P}_Y(X = g(Y))$, we get

$$\begin{aligned} D(W \circ \mathrm{P}_{XY} \| W \circ \mathrm{P}_X \mathrm{P}_Y) &= p \log \frac{1}{q} + (1-p) \log \frac{1}{1-q} - h(p) \\ &= p \log \frac{M}{+} (1-p) \log \frac{M}{M-1} - h(p) \\ &\geq p \log M - h(p) \\ &\geq p \log M - 1. \end{aligned}$$

Upon combining the two bounds above, we obtain

$$\mathrm{P}_{XY}(X = g(Y)) = p \leq \frac{I(X \wedge Y) + 1}{\log M},$$

which completes the proof. □

Next, we present an alternative proof which is more standard. This proof works for any distribution on X, but yields a slightly different form. While we stated the previous version for functions g of Y, it can be extended to randomized functions. Note that if $\hat{X}$ is output of a randomized function of Y, we must have $X -\!\!\circ\!\!- Y -\!\!\circ\!\!- \hat{X}$.

Theorem 2 (Fano's inequality- version 2)**.** *Consider random variables X and Y. Then, for every $\hat{X} -\!\!\circ\!\!- X -\!\!\circ\!\!- Y$ where $\hat{X}$ takes values in $\mathcal{X}$, we have*

$$H(X|Y) \leq \mathbb{P}\left(X \neq \hat{X}\right) \log(|\mathcal{X}| - 1) + h(\mathbb{P}\left(X \neq \hat{X}\right)),$$

where $h(p) = p \log 1/p + (1-p) \log 1/(1-p)$ is the binary entropy function.

Proof. Denote by E the random variable $\mathbb{1}\{X = \hat{X}\}$. (This is the same as the random output of the channel W in the previous proof of Fano's inequality). Then,

$$H(X|\hat{X}) \le H(X, E|\hat{X}) = H(E|\hat{X}) + H(X|\hat{X}, E).$$

Further, since $H(X|\hat{X}, E = 1) = 0$ and X can take at most $|\mathcal{X}| - 1$ values conditioned on the event $E = 0$ and $\hat{X}$, we have

$$\begin{aligned} H(X|\hat{X}, E) &= \mathbb{P}\left(E = 0\right) H(X|\hat{X}, E = 0) + \mathbb{P}\left(E = 1\right) H(X|\hat{X}, E = 1) \\ &= \mathbb{P}\left(E = 0\right) H(X|\hat{X}, E = 0) \\ &\le \mathbb{P}\left(E = 0\right) \log(|\mathcal{X}| - 1). \end{aligned}$$

Upon combining the bounds above, we obtain

$$H(X|\hat{X}) \le H(E|\hat{X}) + \mathbb{P}\left(E = 0\right) \log(|\mathcal{X}| - 1) \le h(\mathbb{P}\left(E = 0\right)) + \mathbb{P}\left(E = 0\right) \log(|\mathcal{X}| - 1).$$

This is almost the bound we wanted, except that the left-side still depends on $\hat{X}$. Finally, we take recourse to the data processing inequality for conditional entropy to get $H(X|Y) \le H(X|\hat{X})$, which completes the proof. □

B Variational formulae

Loosely speaking, a variational formula expresses a quantity as a minimum/maximum. Such a formula has many applications. For us, it is simply a collection of tight upper/lower bounds.

B.1 Variational formula for KL divergence

Lemma 3. *For distributions P and Q on a finite set $\mathcal{X}$ such that $\texttt{supp}(P) \subset \texttt{supp}(Q)$, we have*

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

where the max *is over all R such that $\texttt{supp}(P) \subset \texttt{supp}(R)$. The* max *is attained by $R = P$.*

Proof. We have

$$\begin{aligned} D(P\|Q) &= \sum_x P(x) \log \frac{P(x)}{Q(x)} \\ &= \sum_x P(x) \log \frac{R(x)}{Q(x)} + D(P\|R) \\ &\geq \sum_x P(x) \log \frac{R(x)}{Q(x)}, \end{aligned}$$

with equality iff $P = R$. □

The following alternative form is perhaps more familiar.

Lemma 4. *For distributions P and Q on a finite set $\mathcal{X}$ such that $\texttt{supp}(P) \subset \texttt{supp}(Q)$, we have*

$$D(P\|Q) = \max_f \mathbb{E}_P\left[f(X)\right] - \log \mathbb{E}_Q\left[2^{f(X)}\right],$$

where the maximum is over all functions $f : \mathcal{X} \to \mathbb{R}$ and is attained by $f(x) = \log P(x)/Q(x)$.

Proof. We start by substituting $R(x) = Q(x)2^{f(x)}/\mathbb{E}_Q\left[2^{f(X)}\right]$ in the variational formula above to get

$$\begin{aligned} D(P\|Q) &\geq \sum_x P(x) \log \frac{Q(x)2^{f(x)}}{Q(x)\mathbb{E}_Q\left[2^{f(X)}\right]} \\ &= \mathbb{E}_P\left[f(X)\right] - \log \mathbb{E}_Q\left[2^{f(X)}\right]. \end{aligned}$$

The bound above holds for every f. Furthermore, for $f(x) = \log P(x)/Q(x)$, identity holds in the bound above. □

Note that if $\log P(x)/Q(x)$ belongs to a family $\mathcal{F}$, the maximum in the formula can be restricted to $f \in \mathcal{F}$. This idea has been used to model divergence for machine learning applications where $\mathcal{F}$ is chosen as a family where the maximization of the right-side can be computed efficiently by algorithms.

B.2 Variational formula for mutual information

Lemma 5. *For a discrete distribution P on $\mathcal{X}$ and a channel $W : \mathcal{X} \to \mathcal{Y}$,*

$$I(P; W) = \min_{Q} \sum_{x} P(x) D(W_x \| Q),$$

where the min *is over all distributions Q on $\mathcal{Y}$ and the minimum is attained for $Q(y) = (W \circ P)(y)$.*

Proof. We have

$$\begin{aligned}
I(P; W) &= \sum_{x} P(x) \sum_{y} W(y|x) \log \frac{W(y|x)}{(W \circ P)(y)} \\
&= \sum_{x} P(x) \sum_{y} W(y|x) \log \frac{W(y|x)}{Q(y)} + \sum_{x} P(x) \sum_{y} W(y|x) \log \frac{Q(y)}{(W \circ P)(y)} \\
&= \sum_{x} P(x) D(W_x \| Q) + \sum_{y} (W \circ P)(y) \log \frac{Q(y)}{(W \circ P)(y)} \\
&\leq \sum_{x} P(x) D(W_x \| Q),
\end{aligned}$$

where the inequality holds by nonnegativity of KL divergence and equality is attained for $Q = (W \circ P)$. □

This formula gives a pleasing, geometric interpretation of mutual information. We will use this in the next section to provide an alternative, *information radius* interpretation of the *channel capacity* (we will define both quantities in the next section).

C Applications of variational formulae

C.1 Pinsker's inequality

We obtain Pinsker's inequality as a consequence of the variational formula for KL divergence. In the HW, an alternative proof will be outlined that uses the data processing inequality. Recall the variational formula

$$D(P\|Q) = \max_f \mathbb{E}_P\left[f(X)\right] - \log \mathbb{E}_Q\left[2^{f(X)}\right].$$

Consider the set A such that $d(P,Q) = P(A) - Q(A)$, and let $f_\lambda(x) = \lambda(\mathbb{1}_{\{x\in A\}} - Q(A))$. Then, it is easy to see that $\mathbb{E}_P\left[f_\lambda(X)\right] = \lambda d(P,Q)$ and $\mathbb{E}_Q\left[f_\lambda(X)\right] = 0$. Using this specific choice of $f = f_\lambda$, we get

$$D(P\|Q) \geq \lambda d(P,Q) - \log \mathbb{E}_Q\left[2^{\lambda f_\lambda(X)}\right].$$

To proceed, we make use of a very useful bound called Hoeffding's lemma. It says that the log-moment generating function for bounded functions is quadratic, namely it behaves like that of Gaussian random variables[1]. We state this result below and will use it without a proof.

Lemma 6 (Hoeffding's lemma)**.** *For a random variable X such that $\mathbb{E}\left[X\right] = 0$ and taking values in the interval $[a,b]$, we have*

$$\ln \mathbb{E}\left[e^{\lambda X}\right] \leq \frac{(b-a)^2\lambda^2}{8}.$$

Note that our random variable $\mathbb{1}_{\{x\in A\}} - Q(A)$ is zero-mean under Q and takes values

[1] Recall that the log-moment generating function of a random variable X is given by $\phi(\lambda) = \ln \mathbb{E}\left[e^{\lambda X}\right]$, and for $X \sim \mathcal{N}(0,\sigma^2)$, $\phi(\lambda) = \sigma^2\lambda^2/2$.

values between $-Q(A)$ and $1-Q(A)$. Thus, by Hoeffding's lemma,

$$\log \mathbb{E}\left[2^{f_\lambda(X)}\right] = \frac{1}{\ln 2}\ln \mathbb{E}\left[e^{\ln 2 f_\lambda(X)}\right] \leq \frac{\ln 2\,(1-2Q(A))^2\lambda^2}{8} \leq \frac{\ln 2\,\lambda^2}{8},$$

Upon combining the two bounds above, we obtain

$$D(P\|Q) \geq \lambda d(P,Q) - \frac{\ln 2\,\lambda^2}{8},$$

which on maximizing the right-side over λ gives

$$D(P\|Q) \geq \frac{2}{\ln 2}\, d(P,Q)^2,$$

which is Pinsker's inequality.

C.2 Channel capacity and information radius

The *capacity of a channel* W will be defined operationally later in the course as the maximum number of bits that can be transmitted error-free per channel use. In this section, we will be interested in the formula that characterizes channel capacity, which, with an abuse of terminology, also will be termed channel capacity. Specifically, the capacity $C(W)$ of a channel W is given by

$$C(W) = \max_P I(P;W),$$

that is, the maximum mutual information between the input and the output of the channel. Using the variational formula for mutual information we saw earlier, we obtain:

$$\begin{aligned} C(W) &= \max_P I(P;W) \\ &= \max_P \min_Q \sum_x P(x) D(W_x\|Q). \end{aligned}$$

We now take recourse to a theorem from the theory of convex functions: Sion's minmax theorem. The expression on the right-side above is of the form $\max_u \min_v f(u,v)$. In general, $\max_u \min_v f(u,v) \le \min_v \max_u f(u,v)$ (show this). But in special cases, equality holds. Sion's minmax theorem states, roughly, that equality holds if $f(u,v)$ is concave and continuous in u and convex in v. Our function $\sum_x P(x)D(W_x\|Q)$ is, in fact, continuous and linear in P, whereby it is concave in P, and convex in Q. Thus, by Sion's minmax theorem[2],

$$\max_P \min_Q \sum_x P(x)D(W_x\|Q) = \min_Q \max_P \sum_x P(x)D(W_x\|Q).$$

Note that $\max_P \sum_x P(x)D(W_x\|Q) = \max_x D(W_x\|Q)$, whereby

$$C(W) = \min_Q \max_x D(W_x\|Q).$$

The expression on the right-side is called the *information radius* of W, and rightly so, since it is the minimum of the maximum "distance" of W_xs from a fixed point Q in the set of distributions on $\mathcal{Y}$. Note that in Fano's inequality version 1, we can upper bound channel capacity to get

$$P_e^* \ge 1 - \frac{\max_x D(W_x\|Q) + 1}{\log M},$$

for every distribution Q on $\mathcal{Y}$. This is a very useful bound in practice where we can freely choose a distribution Q of our choice (and not only the ones that can appear on the output of the channel for some input distributions). In particular, it gives

$$P_e^* \ge 1 - \frac{\max_{x,x'} D(W_x\|W_{x'}) + 1}{\log M},$$

which, too, is used often.

[2] We gave a rather informal argument. The actual conditions and their verification will require some more terminology from analysis. Please read the wikipedia entry for Sion's minmax theorem and the original paper cited there.

D Continuity of entropy

In this final section of this technical unit, we present a bound that related $H(P)-H(Q)$ to $d(P,Q)$. To show that entropy is a continuous function of P, we need to show that $H(P)$ approaches $H(Q)$ when P approaches Q. But before we show this, we need to agree in what sense is P approaching Q. We consider these limits in the total variation distance, namely $d(P,Q)$ goes to 0. In fact, we will now only show that $|H(P)-H(Q)|$ approaches 0 as $d(P,Q)$ approaches 0, but show that $|H(P)-H(Q)|$ is almost bounded by a constant (depending on $\mathcal{X}$) times $d(P,Q)$.

To prove this result, we will make use of another beautiful result from probability theory: *the maximum coupling lemma.* There are other elementary proofs available, but we would like to use this excuse to introduce the maximum coupling lemma.

Consider a joint distribution P_{XY} for X and Y taking values in the same set $\mathcal{X}$. Suppose that $\mathrm{P}_X = P$ and $\mathrm{P}_Y = Q$. Then, for any x,

$$P(x) = \mathbb{P}\left(X=x, Y\neq X\right)+\mathbb{P}\left(X=x, Y=X\right) \leq \mathbb{P}\left(X=x, Y\neq X\right)+\mathbb{P}\left(Y=x\right) = \mathbb{P}\left(X=x, Y\neq X\right)+Q(x).$$

Thus, for every x we have

$$P(x) - Q(x) \leq \mathbb{P}\left(X=x, Y\neq X\right).$$

Summing over x such that $P(x) > Q(x)$, we get

$$d(P,Q) \leq \sum_{x:P(x)>Q(x)} \mathbb{P}\left(X=x, Y\neq X\right) \leq \mathbb{P}\left(Y\neq X\right).$$

This bound holds for any joint distribution P_{XY} with marginals of X and Y fixed to P and Q, respectively. In fact, this bound is tight: there exists a joint distribution with marginals P and Q for which equality holds in this bound. A joint distribution with marginals P and Q is called a *coupling* of P and Q. We denote by $\pi(P,Q)$ the set of all couplings of P

and Q

Lemma 7 (Maximum coupling lemma)**.** *For distributions P and Q on $\mathcal{X}$, we have*

$$d(P,Q) = \max_{\mathrm{P}_{XY} \in \pi(P,Q)} \mathbb{P}\left(X \neq Y\right).$$

We have shown one side of the proof; we skip the other side. Instead, we show how this maximum coupling lemma can be used to establish a bound for $|H(P) - H(Q)|$.

Consider a coupling P_{XY} of P and Q ($\mathrm{P}_X = P$ and $\mathrm{P}_Y = Q$). Then, $H(X) = H(P)$ and $H(Y) = H(Q)$, whereby

$$|H(P) - H(Q)| = |H(X) - H(Y)| = |H(X|Y) - H(Y|X)| \leq \max\{H(X|Y), H(Y|X)\}.$$

By Fano's inequality, we have

$$\max\{H(X|Y), H(Y|X)\} \leq \mathbb{P}\left(X \neq Y\right) \log(|\mathcal{X}| - 1) + h(\mathbb{P}\left(X \neq Y\right)).$$

The bound above holds for every coupling. Therefore, choosing the coupling that attains the lower bound of $d(P,Q)$ in maximal coupling lemma (a maximal coupling), we get

$$\max\{H(X|Y), H(Y|X)\} \leq d(P,Q) \log(|\mathcal{X}| - 1) + h(d(P,Q)).$$

Upon combining all the bounds above, we obtain the following result.

Lemma 8. *For P and Q on $\mathcal{X}$, we have*

$$|H(P) - H(Q)| \leq d(P,Q) \log(|\mathcal{X}| - 1) + h(\min\{d(P,Q), 1/2\}),$$

where $h(\cdot)$ denotes the binary entropy function.