Lecture 5 Data processing inequality and Fano inequality

October 24, 2024

Outline

1. Exercises Review
2. Continuity of Shannon's information measures for fixed finite alphabets
3. The range of the entropy function
4. Data processing inequality
5. Fano's inequality
6. Another inequality relating probability of error

Example

A $(7, 4)$ *Hamming code can correct any one error; might there be a* $(14, 8)$ *code that can correct any two errors?*

Proof.

When the decoder receives $\mathbf{r} = \mathbf{t} + \mathbf{n}$, his aim is to deduce both $\mathbf{t}$ and $\mathbf{n}$ from $\mathbf{r}$. If it is the case that the sender can select any transmission $\mathbf{t}$ from a code of size $S_{\mathbf{t}}$, and the channel can select any noise vector from a set of size $S_{\mathbf{n}}$, and those two selections can be recovered from the received bit string $\mathbf{r}$, which is one of at most 2^N possible strings, then it must be the case that

$$S_{\mathbf{t}} S_{\mathbf{n}} \leq 2^N.$$

So, for a (N, K) two-error-correcting code,

$$2^K [\binom{N}{2} + \binom{N}{1} + \binom{N}{0}] \leq 2^N.$$

however the inequality does not hold for $K = 8$ and $N = 14$, which rules out the possibility that there is a $(14, 8)$ code that is 2-error correcting. □

Outline

1. Exercises Review
2. Continuity of Shannon's information measures for fixed finite alphabets
3. The range of the entropy function
4. Data processing inequality
5. Fano's inequality
6. Another inequality relating probability of error

Definition

Let p and q be two probability distributions on a common alphabet $\mathcal{X}$, The variational distance between p and q is defined as

$$V(p,q) = \sum_{x \in \mathcal{X}} |p(x) - q(x)|.$$

For a fixed finite alphabet $\mathcal{X}$, let $\mathcal{P}_X$ be the set of all distributions on $\mathcal{X}$. Then the entropy of a distribution p on an alphabet $\mathcal{X}$ is defined as

$$H(p) = -\sum_{x \in \mathcal{S}_p} p(x) \log p(x),$$

where $\mathcal{S}_p$ denotes the support of p and $\mathcal{S}_p \subset \mathcal{X}$.

Theorem

$H(p)$ is continuous with respect to convergence in variational distance. More precisely, for $p \in \mathcal{P}_{\mathcal{X}}$ and for any $\varepsilon > 0$, there exists $\delta > 0$ such that

$$|H(p) - H(q)| < \varepsilon$$

for all $q \in \mathcal{P}_X$ satisfying

$$V(p, q) < \delta,$$

or equivalently,

$$\lim_{p' \to p} H(p') = H(\lim_{p' \to p} p') = H(p),$$

where the convergence $p' \to p$ is in variational distance.

Proof.

Since $a \log a \to 0$ as $a \to 0$, we define a function $l : [0, \infty) \to \mathbb{R}$ by

$$l(a) = \begin{cases} a \log a & \text{if } a > 0 \\ 0 & \text{if } a = 0 \end{cases}, \tag{2.1}$$

i.e., $l(a)$ is a continuous extension of $a \log a$. Then $H(p)$ can be rewrite as

$$H(p) = -\sum_{x \in \mathcal{X}} l(p(x)),$$

where the summation above is over all x in $\mathcal{X}$ instead of $\mathcal{S}_p$. Unon defining a function $l_x : \mathcal{P}_X \to \mathbb{R}$ for all $x \in \mathcal{X}$ by

$$l_x(p) = l(p(x)),$$

we have that

$$H(p) = -\sum_{x \in \mathcal{X}} l_x(p).$$

Evidently, $l_x(p)$ is continuous in p (with respect to convergence in variational distance). Since the summation above involves a finite number of terms, we conclude that $H(p)$ is a continuous functional of p. □

Outline

1. Exercises Review
2. Continuity of Shannon's information measures for fixed finite alphabets
3. The range of the entropy function
4. Data processing inequality
5. Fano's inequality
6. Another inequality relating probability of error

Corollary

The entropy of a random variable may take any nonnegative real value.

Proof.

Consider a random variable X defined on a fixed finite alphabet $\mathcal{X}$. We see from the last theorem that $H(X) = \log |\mathcal{X}|$ is achieved when X is distributed uniformly on $\mathcal{X}$. On the other hand, $H(X) = 0$ is achieved when X is deterministic. For $0 \leq a \leq |\mathcal{X}|^{-1}$, let

$$\begin{aligned} g(a) &= H(\{1 - (|\mathcal{X}| - 1)a, a, \cdots, a\}) \\ &= -l(1 - (|\mathcal{X}| - 1)a) - (|\mathcal{X}| - 1)l(a), \end{aligned}$$

where $l(\cdot)$ is defined in (2.1). Note that $g(a)$ is continuous in a, with $g(0) = 0$ and $g(|\mathcal{X}|^{-1}) = \log |\mathcal{X}|$. For any value $0 < b < \log |\mathcal{X}|$, by the intermediate value theorem of continuous, there exists a distribution for X such that $H(X) = b$. Then we see that $H(X)$ can take any positive value by letting $|\mathcal{X}|$ be sufficiently large. This accomplishes the proof.

Outline

1. Exercises Review
2. Continuity of Shannon's information measures for fixed finite alphabets
3. The range of the entropy function
4. Data processing inequality
5. Fano's inequality
6. Another inequality relating probability of error

Data processing inequality

Lemma

If $X \to Y \to Z$, *then*

$$I(X;Y) \geq I(X;Z).$$

Proof.

By the chain rule, we can expand mutual information in two different ways:

$$\begin{aligned} I(X;Y,Z) &= I(X;Z) + I(X;Y|Z) \\ &= I(X;Y) + I(X;Z|Y). \end{aligned}$$

Since X and Z are conditionally independent given Y, we have $I(X;Z|Y) = 0$. Since $I(X;Y|Z) \geq 0$, we have

$$I(X;Y) \geq I(X;Z).$$

The equality holds if and only if $I(X;Y|Z) = 0$ (i.e., $X \to Z \to Y$ forms a Markov chain). Similarly, one can prove that $I(Y;Z) \geq I(X;Z)$. □

Corollary

If $Z = g(Y)$, *then* $I(X;Y) \geq I(X;g(Y))$.

Proof.

$X \to Y \to g(Y)$ forms a Markov chain. □

Corollary

If $X \to Y \to Z$, *then* $I(X;Y|Z) \leq I(X;Y)$.

Proof.

Note that

$$\begin{aligned} I(X;Y,Z) &= I(X;Z) + I(X;Y|Z) \\ &= I(X;Y) + I(X;Z|Y). \end{aligned}$$

Since X and Z are conditionally independent given Y, we have $I(X;Z|Y) = 0$. Since $I(X;Z) \geq 0$, we have that $I(X;Y|Z) \leq I(X;Y)$. □

Corollary

If $X_1 \to X_2 \to \cdots \to X_n$, *then for any* i, j, k, l *such that* $1 \leq i \leq j \leq k \leq l \leq n$, *we have that*

$$I(X_i; X_l) \leq I(X_j; X_k).$$

Note that it is also possible that $I(X;Y|Z) > I(X;Y)$ when X, Y and Z do not form a Markov chain.

For example, let X and Y be independent fair binary random variables, and let $Z = X + Y$. Then $I(X;Y) = 0$, but $I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = H(X|Z) = P(Z = 1)H(X|Z = 1) = \frac{1}{2}$ bit.

Outline

1. Exercises Review
2. Continuity of Shannon's information measures for fixed finite alphabets
3. The range of the entropy function
4. Data processing inequality
5. Fano's inequality
6. Another inequality relating probability of error

Fano's inequality

Theorem

Let X and Y be two random variables, correlated in general. with alphabet $\mathcal{X}$ and $\mathcal{Y}$, respectively, where $\mathcal{X}$ is finite but $\mathcal{Y}$ can be countably infinite. Let $\hat{X} := g(Y)$ be an estimate of X from observing Y, where $g : \mathcal{Y} \to \mathcal{X}$ is a given estimation function. Define the probability of error as

$$P_e := Pr[\hat{X} \neq X].$$

Then the following inequality holds

$$H(X|Y) \leq H(X|\hat{X}) \leq h_b(P_e) + P_e \cdot \log_2(|\mathcal{X}| - 1),$$

where $h_b(x) := -x \log_2 x - (1-x) \log_2 (1-x)$ for $0 \leq x \leq 1$ is the binary entropy function.

Proof.

Define a new random variable,

$$E := \begin{cases} 1, & \text{if } \hat{X} \neq X \\ 0, & \text{if } \hat{X} = X. \end{cases}$$

Then using the chain rule for conditional entropy, we obtain

$$\begin{aligned} H(E, X|\hat{X}) &= H(X|\hat{X}) + H(E|X, \hat{X}) \\ &= H(E|\hat{X}) + H(X|E, \hat{X}). \end{aligned}$$

Observe that E is a function of X and $\hat{X}$; hence, $H(E|X, \hat{X}) = 0$. Since conditioning never increases entropy, $H(E|\hat{X}) \leq H(E) = h_b(P_e)$.

Proof.

The remaining term, $H(X|E,\hat{X})$, can be bounded as follows:

$$\begin{aligned} H(X|E,\hat{X}) &= Pr[E=0]H(X|\hat{X},E=0) + Pr[E=1]H(X|\hat{X},E=1) \\ &\leq (1-P_e)\cdot 0 + P_e \cdot \log_2(|\mathcal{X}|-1), \end{aligned}$$

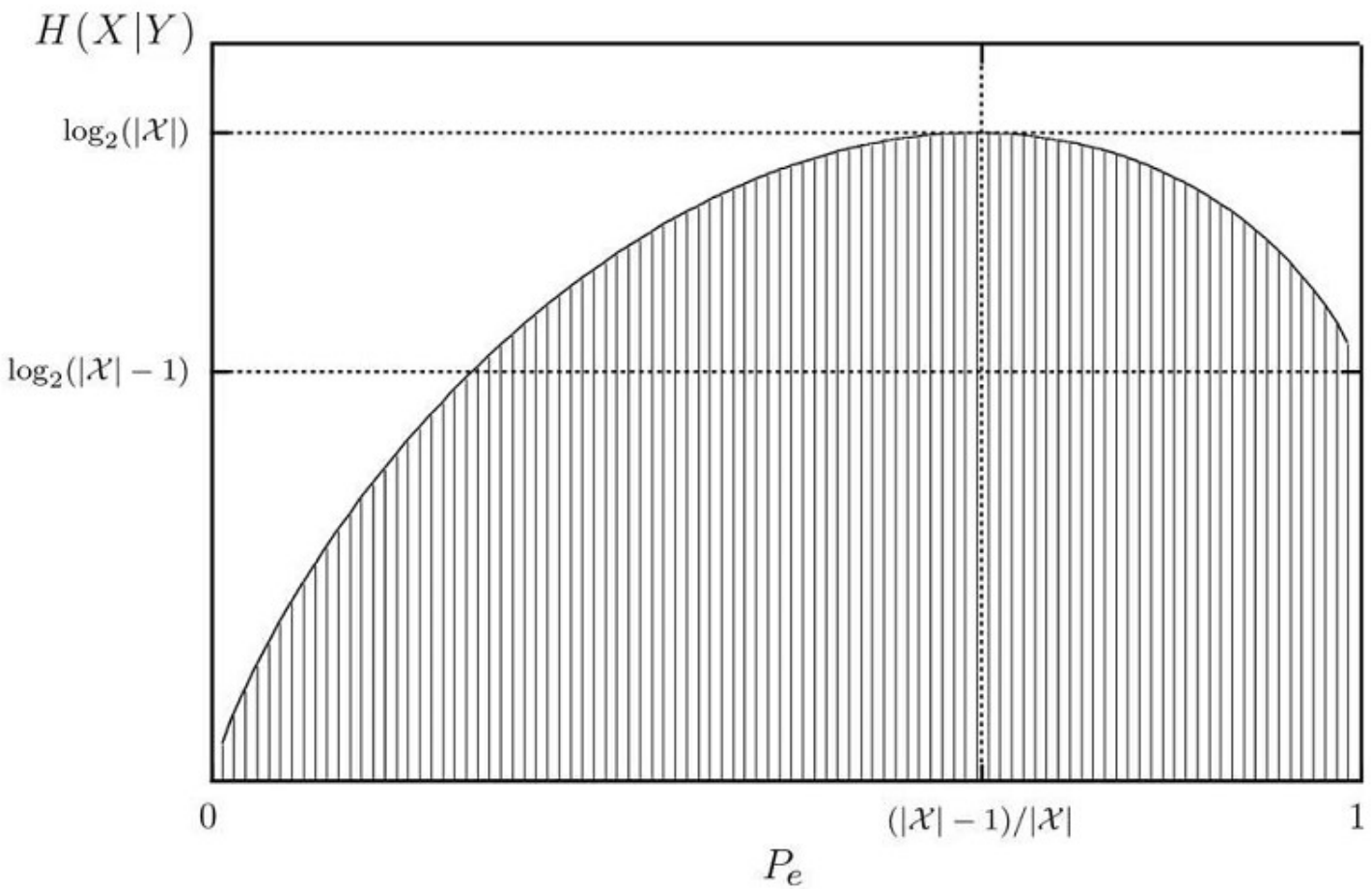
since $X=\hat{X}$ for $E=0$, and given $E=1$, we can upper bound the conditional entropy by the logarithm of the number of remaining outcomes, i.e., $(|\mathcal{X}|-1)$. Combining these results we have that

$$H(X|\hat{X}) \leq h_b(P_e) + P_e \cdot \log_2(|\mathcal{X}|-1).$$

Since $X \to Y \to \hat{X}$ is a Markov chain, by the data-processing inequality, $I(X;\hat{X}) \leq I(X;Y)$, and therefor $H(X|\hat{X}) \geq H(X|Y)$. Thus,

$$H(X|Y) \leq H(X|\hat{X}) \leq h_b(P_e) + P_e \cdot \log_2(|\mathcal{X}|-1). \qquad \square$$

Permissible $(P_e, H(X|Y))$ region due to Fano's inequality



Fano's inequality yields upper and lower bounds on P_e in terms of $H(X|Y)$. This is illustrated in last page, where we plot the region for the pairs $(P_e, H(X|Y))$ that are permissible under Fano's inequality.

In the figure, the boundary of the permissible (dashed) region is given by the function

$$f(P_e) := h_b(P_e) + P_e \cdot \log_2(|\mathcal{X}| - 1).$$

We obtain that when

$$\log_2(|\mathcal{X}| - 1) \leq H(X|Y) \leq \log_2(|\mathcal{X}|),$$

P_e can be upper and lower bounded as follows:

$$0 < \inf\{a : f(a) \geq H(X|Y)\} \leq P_e \leq \sup\{a : f(a) \geq H(X|Y)\} < 1.$$

Furthermore, when

$$0 < H(X|Y) \leq \log_2(|\mathcal{X}| - 1),$$

only the lower bound holds:

$$P_e \geq \inf\{a : f(a) \geq H(X|Y)\} > 0.$$

Thus for all nonzero values of $H(X|Y)$, we obtain a lower bound (of the same form above) on P_e; the bound implies that if $H(X|Y)$ is bounded away from zero, P_e is also bounded away from zero.

A weaker but simpler version of Fano's inequality can be directly obtained from Fano's inequality by noting that $h_b(P_e) \leq 1$:

$$H(X|Y) \leq 1 + P_e \log_2(|\mathcal{X}| - 1),$$

which in turn yields that

$$P_e \geq \frac{H(X|Y) - 1}{\log_2(|\mathcal{X}| - 1)} \text{ (for } |\mathcal{X}| > 2)$$

which is weaker then the above lower bound on P_e.

Fano's inequality cannot be improved in the sense that the lower bound, $H(X|Y)$, can be achieved for some specific cases. Any bound that can be achieved in some cases is often referred to as sharp.

Fano's inequality cannot be improved in the sense that the lower bound, $H(X|Y)$, can be achieved for some specific cases. Any bound that can be achieved in some cases is often referred to as sharp.

From the proof of the above lemma, we can observe that equality holds in Fano's inequality, if $H(E|Y) = H(E)$ and $H(X|Y, E = 1) = \log_2(|\mathcal{X}| - 1)$. The former is equivalent to E being independent of Y, and the latter holds iff $P_{X|Y}(\cdot|y)$ is uniformly distributed over the set $\mathcal{X}\backslash\{g(y)\}$. We can therefore create an example in which equality holds in Fano's inequality.

Corollary

Let X and $\hat{X}$ be random variables taking values in the same alphabet $\mathcal{X}$. Then

$$H(X|\hat{X}) \leq h_b(P_e) + P_e \log(|\mathcal{X}| - 1),$$

where h_b is the binary entropy function.

Example

Suppose that X and Y are two independent random variables which are both uniformly distributed on the alphabet $\{0, 1, 2\}$. Let the estimating function be given by $g(y) = y$. Then

$$P_e = Pr[g(Y) \neq X] = Pr[Y \neq X] = 1 - \sum_{x=0}^{2} P_X(x) P_Y(x) = \frac{2}{3}.$$

In this case, equality is achieved in Fano's inequality, i.e.,

$$h_b(\frac{2}{3}) + \frac{2}{3} \cdot \log_2(3-1) = H(X|Y) = H(X) = \log_2 3.$$

Outline

1. Exercises Review
2. Continuity of Shannon's information measures for fixed finite alphabets
3. The range of the entropy function
4. Data processing inequality
5. Fano's inequality
6. Another inequality relating probability of error

Let X and X' be two independent identically distributed random variables with entropy $H(X)$. The probability at $X = X'$ is given by

$$Pr(X = X') = \sum_x p^2(x).$$

Lemma

If X and X' are i.i.d. with entropy $H(X)$.

$$Pr(X = X') \geq 2^{-H(X)},$$

with equality if and only if X has a uniform distribution.

Proof.

Suppose that $X \sim p(x)$. By Jensen's inequality, we have

$$2^{E \log p(x)} \leq E 2^{\log p(x)},$$

which implies that

$$2^{-H(X)} = 2^{\sum p(x) \log p(x)} \leq \sum p(x) 2^{\log p(x)} = \sum p^2(x).$$

Corollary

Let X, X' be independent with $X \sim p(x)$, $X' \sim r(x)$, $x, x' \in \mathcal{X}$. Then

$$\begin{aligned} P(X = X') &\geq 2^{-H(p)-D(p\|r)}, \\ P(X = X') &\geq 2^{-H(r)-D(r\|p)}. \end{aligned}$$

Proof.

We have

$$\begin{aligned} 2^{-H(p)-D(p\|r)} &= 2^{\sum p(x)\log p(x)+\sum p(x)\log \frac{r(x)}{p(x)}} \\ &= 2^{\sum p(x)\log r(x)} \\ &\le \sum p(x)2^{\log r(x)} \\ &= \sum p(x)r(x) \\ &= P(X=X'), \end{aligned}$$

where the inequality follows from Jensen's inequality and the convexity of the function $f(y)=2^y$. □