Information Theory Condensed Notes

Kexing Ying

January 6, 2023

Basic Definitions and Properties

We always work in logarithmic base 2 unless explicitly stated otherwise (e.g. $\log_e$ denotes the natural logarithm).

Definition (Entropy)**.** Given a random variable X on some finite space $\mathscr{A}$, denoting P the probability mass function of X, X has entropy

$$H(X) = H(P) := -\sum_{x \in \mathscr{A}} P(x) \log P(x) = \mathbb{E}[-\log P(X)].$$

By convention we take $0 \log 0 := 0$

Proposition (Bernoulli entropy)**.** If $X \sim \text{Bern}(p)$ then $H(X) = -p \log p - (1-p) \log(1-p)$.

Definition (Fixed-rate code)**.** A fixed-rate lossless compression code for a source (X_n) (always iid.) on $\mathscr{A}$ is a sequence of codebooks $B_n \subseteq \mathscr{A}^n$.

The idea of a compression using a fixed-rate code is to index the codebooks using $\lceil \log |B_n| \rceil$ bits. Then to transmit $x_1^n \in \mathscr{A}^n$, if $x_1^n \in B_n$, we transmit 1 postfixed with the index corresponding to x_1^n in B_n. This costs $1 + \lceil \log |B_n| \rceil$ bits. On the other hand if $x_1^n \notin B_n$, we transmit 0 and the entire string x_1^n. This costs $1 + \lceil \log |\mathscr{A}^n| \rceil 1 + \lceil n \log |\mathscr{A}| \rceil$ bits.

Definition (Rate and error probability)**.** Given a fixed-rate code B_n for the source (X_n), the rate of the code is defined as

$$R_n = \frac{1}{n}(1 + \lceil \log |B_n| \rceil),$$

and its probability of error is

$$P_e^{(n)} = \mathbb{P}(X_1^n \notin B_n).$$

Definition (Relative entropy)**.** The relative entropy of the pmfs P, Q on $\mathscr{A}$ is

$$D(P \| Q) = \sum_{x \in \mathscr{A}} P(x) \log \frac{P(x)}{Q(x)} = \mathbb{E}\left[\log \frac{P(X)}{Q(X)}\right],$$

for some random variable $X \sim P$. Again we introduce the convention $0 \log 0 = 0, 0 \log \frac{0}{0} = 0$.

Theorem 1 (Log-sum inequality)**.** For non-negative constants $a_1, \cdots, a_n$ and $b_1, \cdots, b_n$,

$$\sum_{i=1}^n a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^n a_i\right) \log \frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i}.$$

Equality is achieved if and only if a_i/b_i is some fixed constant.

Proposition. Let P, Q be two pmfs on $\mathscr{A}$, then

- $0 \leq H(P) \leq \log|\mathscr{A}|$ and $H(P) = 0$ if and only if P is a Dirac measure and $H(P) = \log|\mathscr{A}|$ if and only if P is uniform.
- $D(P\|Q) \geq 0$ with equality if and only if $P = Q$.

Definition (Conditional entropy). Given X, Y random variables on $\mathscr{A}$ with joint pmf P_{XY}, the conditional entropy of X given Y is

$$H(Y \mid X) = -\sum_{x,y\in\mathscr{A}} P_{XY}(x,y)\log P_{Y|X(y|x)} = \mathbb{E}[-\log P_{Y|X}(Y \mid X)]$$

where $P_{Y|X}(y \mid x) = \frac{P_{XY}(x,y)}{P_X(x)}$.

Proposition. Given X, Y, Z random variables and $(X_n), (Y_n)$ sequences of random variables (not necessary independent) on $\mathscr{A}$,

- $H(X,Y) = H(X) + H(Y \mid X)$;
- $H(Y \mid X) \leq H(Y)$ with equality if and only if X and Y are independent;
- $H(f(X)) \leq H(X)$ with equality if and only if f is bijective;
- $H(f(X) \mid X) = 0$;
- $H(X,Z \mid Y) = H(X \mid Y) + H(Z \mid X,Y)$;
- $H(X,Z \mid Y) \leq H(X \mid Y) + H(Z \mid Y)$ with equality if and only if X and Z are conditionally independent given Y;
- $H(X \mid Y,Z) \leq H(X \mid Y)$ with equality if and only if X and Z are conditionally independent given Y;
- $H(X_1^n) = \sum_{i=1}^n H(X_i \mid X_1^{i-1}) = H(X_1) + H(X_2 \mid X_1) + \cdots + H(X_n \mid X_1^{n-1})$;
- $H(X_1^n) \leq \sum_{i=1}^n H(X_i)$ with equality if and only if (X_n) is independent;
- if $f : \mathscr{A} \to \mathscr{B}$ be some function, $D(P_{f(X)}\|P_{f(Y)}) \leq D(P_X\|P_Y)$;

Definition (Total variation). The total variation of two pmfs P, Q on $\mathscr{A}$ is

$$\|P - Q\|_{\mathrm{TV}} := \sum_{x\in\mathscr{A}} |P(x) - Q(x)|.$$

Proposition. $D(P\|Q)$ is jointly convex in P, Q, i.e. for pmfs $P_i, Q_i, i = 1,2$ and $\lambda \in (0,1)$,

$$D(\lambda P_1 + (1-\lambda)P_2\|\lambda Q_1 + (1-\lambda)Q_2) \leq \lambda D(P_1\|Q_1) + (1-\lambda)D(P_2\|Q_2).$$

Proposition. $H(P)$ is concave in P, i.e. for pmfs $P_i, i = 1,2$ and $\lambda \in (0,1)$,

$$H(\lambda P_1 + (1-\lambda)P_2) \geq \lambda H(P_1) + (1-\lambda)H(P_2).$$

Named theorems

Proposition (Asymptotic equipartition property (AEP)). Given (X_n) a iid. sequence of random variables on $\mathscr{A}$ finite with pmf P and entropy $H = H(X_i)$, then

- for all $\epsilon > 0$, defining the set of typical strings

$$B_n^* := \{2^{-n(H+\epsilon)} \leq P^n(x_1^n) \leq 2^{-n(H-\epsilon)}\} \subseteq \mathscr{A}^n,$$

  B_n^* satisfies $|B_n^*| \leq 2^{n(H+\epsilon)}$ and $\mathbb{P}(X_1^n \in B_n^*) = P^n(B_n^*) \to 1$.

- for all sequences of sets $B_n \subseteq \mathscr{A}^n$ satisfying $\mathbb{P}(X_1^n \in B_n) \to 1$, given $\epsilon > 0$, we have $|B_n| \geq (1-\epsilon)2^{n(H-\epsilon)}$ eventually.

Proposition (Fixed-rate coding). Given the source (X_n) on $\mathscr{A}$ with pmf P and entropy H,

- for all $\epsilon > 0$, there exists a fixed-rate code (B_n^*) with $P_e^{(n)} \to 0$ and

$$R_n \leq H + \epsilon + \frac{2}{n}.$$

- for all fixed-rate code (B_n) with $P_e^{(n)} \to 0$, for any ϵ, B_n has rate eventually satisfying

$$R_n > H - \epsilon.$$

Proposition (Stein's lemma). Suppose (X_n) is a sequence of iid. random variables on $\mathscr{A}$ and P, Q are two pmfs on $\mathscr{A}$. Then, given a sequence of sets $B_n \subseteq \mathscr{A}^n$ of decision regions, we denote the probability of errors

$$e_1^{(n)} = \mathbb{P}(X_1^n \in B_n \mid X_i \sim Q), \text{ and } e_2^{(n)} = \mathbb{P}(X_1^n \notin B_n \mid X_i \sim P).$$

Then,

- for all $\epsilon > 0$, there exists decision regions (B_n^*) such that for all n

$$e_1^{(n)} \leq 2^{-n(D(P\|Q)-\epsilon)} \text{ and } \lim_{n\to\infty} e_2^{(n)} = 0.$$

- if (B_n) are decision regions such that $e_2^{(n)} \to 0$ as $n \to \infty$, then for all $\epsilon > 0$,

$$e_1^{(n)} \geq 2^{-n(D+\epsilon+n^{-1})}.$$

Proposition (Neyman-Pearson lemma). For P, Q pmfs on $\mathscr{A}$ and $x_1^n \in \mathscr{A}^n$ we define the Neyman-Pearson decision region

$$B_{\mathrm{NP}} := \left\{ \frac{P^n(x_1^n)}{Q^n(x_1^n)} \geq T \right\}$$

for some threshold $T > 0$ with probability of error $e_{1,\mathrm{NP}}^{(n)} = Q^n(B_{\mathrm{NP}})$ and $e_{2,\mathrm{NP}}^{(n)} = P^n(B_{\mathrm{NP}^c})$. Then, for any other decision region $B_n \subseteq \mathscr{A}^n$, such that $e_2^{(n)} \leq e_{2,\mathrm{NP}}^{(n)}$, we have $e_1^{(n)} \geq e_{1,\mathrm{NP}}^{(n)}$.

Proposition. The Neyman-Pearson decision region B_{NP} can also be expressed in terms of relative entropy as

$$B_{\mathrm{NP}} = \{D(\hat{P}_n \| Q) \geq D(\hat{P}_n \| P) + T'\}$$

where $T' = \frac{1}{n} \log T$.

Proposition (Fano's inequality)**.** Given X, Y random variables taking value in $\mathscr{A}$ and $\mathscr{B}$ respectively, and $f : \mathscr{B} \to \mathscr{A}$ is som function, we have

$$H(X \mid Y) \leq h(P_e) + P_e \log(|\mathscr{A}| - 1),$$

where h is the Bernoulli entropy and $P_e := \mathbb{P}(f(Y) \neq X)$.

Proposition (Pinsker's inequality)**.** Given two pmfs P, Q on $\mathscr{A}$,

$$\|P - Q\|_{\mathrm{TV}}^2 \leq (2 \log_e 2) D(P \| Q).$$