§1. Entropy, Divergence, and Mutual Information

$H(X) \coloneqq -\sum_x p_X(x)\log p_X(x) = -\mathbb{E}(\log p_X(x))$ $\quad D(p||q) \coloneqq \sum_x p(x)\frac{p(x)}{q(x)} = \mathbb{E}(\log\frac{1}{q(X)}) - H(X)$

$I(X;Y) \coloneqq \sum_{x,y} p_{X,Y}(x,y)\log\frac{p_{X,Y}(x,y)}{p_X(x)p_y(y)} = D(p_{X,Y}||p_Xp_Y) = H(X)+H(Y)-H(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$

$H(Y|X) = -\sum_{x,y} p_{X,Y}(x,y)\log p_{Y|X=x}(y) = -\sum_x p_X(x)\mathbb{E}(\log p_{Y|X=x}(Y)) = -\mathbb{E}(\log p_{Y|X}(Y)) = \sum_x H(Y|X=x)\mathbb{P}(X=x) = H(X,Y) - H(X)$

$I(X;Y|Z) \coloneqq H(X|Z) - H(X|Y,Z)$

Gibb's Inequality $H(X) = -\sum_x p(x)\log p(x) \leq -\sum_x p(x)\log q(x)$ $\quad$ ($=$ iff $p=q$)

$D(p_{X,Y}||p_{\hat{X},\hat{Y}}) = D(p_{Y|X}||p_{\hat{Y}|\hat{X}}|p_X) + D(p_X||p_{\hat{X}})$

$D(p_{Y|X}||p_{\hat{X},\hat{Y}}|p_X) = D(p_Xp_{Y|X}||p_Xp_{\hat{Y}|\hat{X}})$

$D(p_{Y|X}||q_{Y|X}|p_X) = \sum_x p_X(x)D\left(p_{Y_1|X=x}||q_{Y_2|X=x}\right)$

Logsum: $a_i, b_i \geq 0 \Longrightarrow \sum_{i=1}^n a_i\log\left(\frac{a_i}{b_i}\right) \geq \left(\sum_{i=1}^n \log\left(\frac{\sum_{i=1}^n a_i}{\sum_{i=1}^n b_i}\right)\right)$ Pf) $q_i = a_i/\sum_{i=1}^n a_i$ and Gibbs

$D(\lambda p_1 + (1-\lambda)p_2||\lambda q_1 + (1-\lambda)q_2) \leq \lambda D\left(p_1||q_1\right) + (1-\lambda)\,D\left(p_2||q_2\right)$ for $\lambda\in[0,1]$ (Pf: logsum)

$I(X;Y)\geq 0$ $\quad$ ($=$ iff $X\perp Y$) $\qquad I(X;Y) = H(X) - H(X|Y)$

$H(X,Y) = H(X) + H(Y|X)$ $\qquad I(X_1,\cdots,X_n;Y) = \sum_{i=1}^n I(X_i;Y|X_{i-1},\cdots,X_1)$ (Chain)

Data Processing $(X\perp Z)|Y \Longrightarrow I(X;Y)\geq I(X;Z)$ Pf) Chain $I(Y,Z;X)$ and $I(Z,Y;X)$

$f:\mathcal{X}\to\mathcal{Y} \Longrightarrow I(X;Y)\geq I(X;f(Y))$

$0\leq H(X)\leq \log\left(|\mathcal{X}|\right)$ $\quad$ ($=$ iff X const, X uniform respectively)

$0\leq H(X|Y)\leq H(X)$ $\quad$ (2^{nd} $=$ iff $X\perp Y$ iff $X=f(Y)$)

$H(f(X))\leq H(X)$ $\quad$ ($=$ iff f bijective)

$H(X_1,\cdots,X_n) = \sum_{i=1}^n H(X_i|X_{i-1},\cdots,X_1) \leq \sum_{i=1}^n H(X_i)$ $\quad$ ($=$ iff X_i indep)

X, Y iid $\Longrightarrow \mathbb{P}(X=Y)\geq 2^{-H(X)}$ $\quad$ ($=$ iff uniform, proof by Jensen)

Fano's Inequality $H(X|Y)\leq H(\mathbf{1}_{X\neq Y}) + \mathbb{P}(X\neq Y)\log\left(|\mathcal{X}|-1\right)$ $\quad$ (Proof by defining $Z=\mathbf{1}_{X\neq Y}$ and $H(Z|X,Y)=0$, so that $H(X|Y) = H(Z|Y) + H(X|Y,Z)$)

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

Max Entropy Tip: Use Gibb's to bound $H(q)$, and equality iff $p=q$ where p specified as:

For $\mathbb{E}(X) = const.$ on $\{1,\cdots\}$, $X\sim\text{geom}(p)\in\{1,\cdots\}$, $\mathbb{E}(X) = \frac{1}{p}$, $H(X) = \frac{-(1-p)\log 1-p-p\log p}{p}$

For $\mathbb{E}(X) = const.$ on $\{0,\cdots\}$, $X\sim\text{geom}(p)\in\{0,\cdots\}$, $\mathbb{E}(X) = \frac{1-p}{p}$, $H(X) = \frac{-(1-p)\log 1-p-p\log p}{p}$

Max Entropy for $\mathbb{E}(f(X)) = C$, $X\sim p_X(x) = \frac{e^{-\lambda f(x)}}{\sum_x e^{-\lambda f(x)}}$ where λ chosen s.t. $\mathbb{E}(f(X)) = C$

---

§2. AEP WLLN: $\bar{X}_n \xrightarrow{P} \mu$ (ie: $\lim_{n\to\infty}\mathbb{P}\left(|\bar{X}_n - \mu| < \epsilon\right) = 1$ for any $\epsilon > 0$)

Weak AEP 1: $-\frac{1}{n}\log p_{X_1,\cdots,X_n}(X_1,\cdots,X_n) \xrightarrow{P} H(X)$

$\mathcal{T}_n^\epsilon \coloneqq \left\{(x_1,\cdots,x_n)\in\mathcal{X}^n : |-\frac{1}{n}\log p_{X_1,\cdots,X_n}(x_1,\cdots,x_n) - H(X)| \leq \epsilon\right\}$

$\forall\epsilon>0$, $\exists N$ s.t. $\forall n\geq N$,

(i) $p_{X_1,\cdots,X_n}(x_1,\cdots,x_n) \in [2^{-n(H(X)+\epsilon)}, 2^{-n(H(X)-\epsilon)}]$ for any $(x_1,\cdots,x_n)\in\mathcal{T}_n^\epsilon$

(ii) $\mathbb{P}((X_1,\cdots,X_n)\in\mathcal{T}_n^\epsilon) \geq 1-\epsilon$ (From weak AEP 1)

(iii) $|\mathcal{T}_n^\epsilon| \in \left[(1-\epsilon)2^{n(H(X)-\epsilon}, 2^{n(H(X))+\epsilon}\right]$ $\quad$ Pf) $1 = \sum_x p_{\mathbf{X}}(\mathbf{x}) \geq \sum_{\mathcal{T}_n^\epsilon} p_{\mathbf{X}}\mathbf{x} \geq \sum_{\mathcal{T}_n^\epsilon} 2^{-n(H(X)+\epsilon)}$ and $1-\epsilon \leq \mathbb{P}\left((X_1,\cdots,X_n)\in\mathcal{T}_n^\epsilon\right) \leq \sum_{\mathcal{T}_n^\epsilon} 2^{-n(H(X)-\epsilon)} = 2^{-n(H(X)-\epsilon)}|\mathcal{T}_n^\epsilon|$

Shannon's 1st Thrm: $\forall\epsilon>0$, $\exists n\in\mathbb{Z}$ and $c:\mathcal{X}^*\to\{0,1\}^*$ s.t. $\cup_{k\geq 0}\mathcal{X}^{nk}\to\{0,1\}^*$, $(x_1,\cdots,x_k)\to c(x_1)\cdots c(x_k)\in\{0,1\}^*$ injective and $\frac{1}{n}\mathbb{E}\left(|c\left(X_1,\cdots,X_n\right)|\right)\leq H(X)+\epsilon$

---

§3. Optimal Codes

$\lceil A\rceil - 1 < A \leq \lceil A\rceil$ $\qquad A\leq\lceil A\rceil < A+1$

Kraft-McMillan: (i) $c:\mathcal{X}\to\mathcal{Y}^*$ uniquely decodable and $l_x \coloneqq |c(x)| \to \sum_{x\in\mathcal{X}}|\mathcal{Y}|^{-l_x}\leq 1$

Pf) $(\sum_x d^{-|c(x)|})^n = \sum_{k=nl_{\min}}^{nl_{\max}} a(k)d^{-k}$, $a(k) < d^k$ by uniq decod. Take n^{th} root and $n\to\infty$.

(ii) Given $(l_x)_{x\in\mathcal{X}}\subset\mathbb{N}$ and $\sum_{x\in\mathcal{X}}|\mathcal{Y}|^{-l_x}\leq 1$, then $\exists$ prefix code $c:\mathcal{X}\to\mathcal{Y}^*$ s.t. $|c(x)| = l_x$

Pf) Relabel $\mathcal{X} = \{1,\cdots,|\mathcal{X}|\}$, $l_1\leq\cdots\leq l_{|\mathcal{X}|}$. $r_m \coloneqq \sum_{i=1}^{m-1}|\mathcal{Y}|^{-l_i}$. $c(m)$: first m digits of r_m

X be RV in finite $\mathcal{X}$ and c (uniq decod., d-ary), then $H_d(X)\leq\mathbb{E}(|c(X)|)$ ($=$ iff $|c(x)| = -\log_x p_X(x)$, Lower bd on length) $\quad$ Pf) Let $l_x = c(x)$, $q(x) = \frac{d^{-l_x}}{\sum_{x\in\mathcal{X}} d^{-l_x}}$, consider $\mathbb{E}(|c(X)|) - H_d(X)$

<u>Existence of Optimal Code</u> $H_d(X)\leq\mathbb{E}(|c^*(X)|) < H_d(X)+1$ $\quad$ Pf) $l_x = \lceil -\log_d(p(x))\rceil$ Kraft-McMillan

Shannon's Code (i) Order $p_1 \geq \cdots \geq p_m$. (ii) $c(x_r) =$ first $l_r := \lceil -\log_{|\mathcal{Y}|}(p_r) \rceil$ digits of $\sum_{i=1}^{r-1} p_i$

Shannon with distrib. estimation p, q (estimation) pmf on $\mathcal{X}$,

$H_d(X) + D_d(p||q) \leq \mathbb{E}(|c_q(X)|) < H_d(X) + D_d(p||q) + 1$

Pf: Bound $\mathbb{E}(|c_q(X)|) = \sum_x p(x) \lceil -\log_d(q(x)) \rceil$

Elias' code First $\lceil -\log_d(p_i) \rceil + 1$ digits of $\sum_{i<r} p_i + \frac{p_r}{2} \implies H_d(X) + 1 \leq \mathbb{E}(|c_E(X)|) \leq H_d(X) + 2$

Bijection between d-ary prefix codes and d-ary rooted trees.

Huffman code is optimal Pf) if $p_1 \geq \cdots \geq p_m$, (i) $p_j > p_k \implies |c(x_j)| \leq |c(x_k)|$, (ii) two longest codewords have same len (iii) two longest codewords only differ in last digit. p,$p' \implies L(c^p) - L(c^{p'}) = p_{m-1} + p_m$ Also for e^p,$e^{p'}$, $L(e^p) - L(e^{p'}) = p_{m-1} + p_m$. Subtract each other, $L(e^p) = L(c^p)$

---

§4 Channel Coding DMC $(\mathcal{X}, M, \mathcal{Y})$ where $\mathcal{X}$ (input alphabet), $\mathcal{Y}$ (output alphabet), M ($|\mathcal{X}| \times |\mathcal{Y}|$ stochastic matrix).

Channel Capacity: $C := \sup I(x;y) = H(Y) - H(Y|X)$

Tip: $H(Y|X) = \sum_x H(Y|X=x) p_X(x)$, Use $I(X,Y) = H(Y) - H(Y|X)$

(m,n)-channel code for DMC $(\mathcal{X}, M, \mathcal{Y})$: tuple (c,d) where $c : \{1, \cdots, m\} \to \mathcal{X}^n$ (Encoder) and $d : \mathcal{Y}^n \to \{1, \cdots, m\}$ (Decoder)

Rate of (m,n)-code (c,d): $\rho(c,d) := \frac{1}{n} \log_{|\mathcal{X}|}(m)$

$\epsilon_i = \mathbb{P}\left(d(\mathbf{Y} \neq i | c(i) = \mathbf{X}\right)$ for $i = 1, \cdots, m$, $\epsilon_{\max} := \max_{i \in \{1, \cdots, m\}} \epsilon_i$, $\bar{\epsilon} := \frac{1}{m} \sum_{i=1}^m \epsilon_i$

Rate R achievable if $\forall \epsilon > 0$, $\exists$ suff. large. m,n and (m,n)-channel code (c,d) with $\rho(c,d) > R - \epsilon$ and $\epsilon_{\max} < \epsilon$

Shannon's 2^{nd} theorem DMC $(\mathcal{X}, M, \mathcal{Y})$ with capacity C, then $R > 0$ achievable iff $R \leq C$

Pf)

$\mathcal{J}_\epsilon^{(n)} =$

$\left\{(x,y) \in \mathcal{X}^n \times \mathcal{Y}^n : \max\left(|\frac{-\log p_{\mathbf{X},\mathbf{Y}}(x,y)}{n} - H(X,Y)|, |\frac{-\log(p_{\mathbf{X}}(x))}{n} - H(X)|, |\frac{-\log(p_{\mathbf{Y}}(y))}{n} - H(Y)|\right)\right\}$

Joint AEP: $\mathbf{X} = (X_1, \cdots, X_n)$, $\mathbf{Y} = (Y_1, \cdots, Y_n)$

(i) $\lim_{n\to\infty} \mathbb{P}\left((\mathbf{X}, \mathbf{Y}) \in \mathcal{J}_\epsilon^{(n)}\right) = 1$ (ii) $|\mathcal{J}_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)}$

(iii) $\exists n_0$ s.t. $\forall n \geq n_0$, $(1-\epsilon) 2^{-n(I(X;Y)+3\epsilon)} \leq \mathbb{P}\left((\mathbf{X}', \mathbf{Y}') \in \mathcal{J}_\epsilon^{(n)}\right) \leq 2^{-n(I(X;Y)-3\epsilon)}$

Channel Coding w/ non-iid Input

Stationary stochastic process: $\mathbb{P}(X_1 = x_1, \cdots, X_n = x_n) = \mathbb{P}(X_{1+j} = x_1, \cdots, X_{n+j} = x_n)$ for all $n, j \in \mathbb{Z}$

Entropy rate of stochastic process: $\mathcal{H}(X) = \lim_{n\to\infty} \frac{1}{n} H(X_1, \cdots, X_n)$

Lemma (1): Stationary Stochastic process X, $n \to H(X_n | X_{n-1}, \cdots, X_1)$ is non-increasing and $\lim_{n\to\infty} H(X_n | X_{n-1}, \cdots, X_1)$ exists.

Lemma (2): $\lim_{n\to\infty} a_n = a \implies \frac{1}{n} \sum_{i=1}^n a_i = a$

Thrm: Stationary stoch. process $X \implies \mathcal{H}(X) = \lim_{n\to\infty} H(X_n | X_{n-1}, \cdots, X_1)$ PF) Lemma (i),(ii)

Lemma (3): $H(Y_n | Y_{n-1}, \cdots, Y_1)$

Lemma (4): $H(Y_n | Y_{n-1}, \cdots, Y_2, Y_1) \leq \lim_k H(Y_{n+k+1} | Y_{n+k}, \cdots, Y_1) = \mathcal{H}(Y)$

Thrm: $X(X_i)_{i\geq 1}$ stationary Markov. $\phi : \mathcal{X} \to \mathcal{Y}$, $Y_i := \phi(X_i)$, then $H(Y_n | Y_{n-1}, \cdots, Y_1, X_1) \leq \mathcal{H}(Y) \leq H(Y_n | Y_{n-1}, \cdots, Y_1)$, $\mathcal{H}(Y) = \lim_{n\to\infty} H(Y_n | Y_{n-1}, \cdots, Y_1, X_1) = \lim_{n\to\infty} H(Y_n | Y_{n-1}, \cdots, Y_1)$

PF) Lemma (3),(4)

---

§Appendix

Markov Inequality: $\mathbb{P}(X \geq x) \leq \frac{\mathbb{E}(X)}{x}$ for all $x > 0$

Chebyshev Inequality: $\mathbb{P}(|X - \mu| > \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$

Optimal $\neq$ Huffman: $0.3 = 00, 0.3 = 10, 0.2 = 01, 0.2 = 11$