Notes

14 May 2023 10:33

* Entropy: $H_D(X) = -\sum_x p(x)\log_D p(x) \le \log|X|$
  - chain rules: $H(X_1 \dots X_n) = \sum H(X_i | X_{i-1} \dots X_1)$
  - Independence bound: $H(X_1 \dots X_n) \le \sum H(X_i)$
* Theorem of total prob: if X, Y, Z are jointly distributed

$$P(Y|X) = \sum_z P(Y|X,Z).P(Z|X) \quad | \quad E[XZ] = E[E[XZ|Y]]$$

* Markov inequality: $P(X \ge a) \le \frac{E[X]}{a}$; $a > 0$
* Chebyshev inequality: $P(|X - E[X]| < \varepsilon) \ge 1 - \frac{var(X)}{\varepsilon^2}$; $\varepsilon > 0$
* Law of Large Number: $\lim_{n\to\infty} P(|\bar{X}_n - E[X]| < \varepsilon) = 1$
* Mutual information: $I(X;Y) = H(X) - H(X|Y)$

$$= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

  + Conditional MI: $I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$
  + Chain rule: $I(X;YZ) = I(X;Y|Z) + I(X;Z)$
* KL divergence: $D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$
* Markov chain: $X \to Y \to Z(\hat{X})$
  + $P(Z|X,Y) = P(Z|Y)$
  + $P(Z,X|Y) = P(Z|Y)P(X|Y)$
  + $I(X;Y) \ge I(X;Z)$ (DPE)
  + $H(X|Z) \ge H(X|Y)$
  + Fano's inequality: $P_e = P(\hat{X} \ne X)$

$$h(P_e) + P_e \log|X| \ge H(X|\hat{X}) \ge H(X|Y)$$

  + Stationary process: time-shift invariant:

$$P(X_1, \dots X_n) = P(X_{l+1}, \dots X_{n+l}) \ \forall l$$

  + Time-invariant Markov chain: $P(X_{n+1}|X_n) = P(X_2|X_1)$
  + Steady-state distribution: $z = zP \Leftrightarrow z(P - I_m) = 0 \Leftrightarrow zQ = 0 \Leftrightarrow z\tilde{Q} = [1\,0 \dots 0]$

* Channel model:

$$w \in W \to \boxed{W \to C} \xrightarrow{x \in C} \boxed{p(y|x)} \xrightarrow{y} \boxed{Y^n \to W} \to \hat{w}$$

Codebook (M,n)

$x \in X^n$

$M = |W| \le |X^n|$ $= 2^{nR}$

carry nR bits

  + Transmission rate: $R = \frac{1}{n} \log M$ ↑↑ (bits)
  + Conditional prob of error: $\lambda_w = P(g(Y) \ne w | X = x(w))$
  + Avg prob of error: $P_e = \frac{1}{M} \sum_w^M \lambda_w$
  + Max prob of error: $\lambda^{(n)} = \max_w \lambda_w$

Reliable Transmission

  + Channel capacity: $C = \max_{p_X(x)} I(X;Y)$ $= \max_{p_X(x)} H(Y) - H(Y|X)$

BSC: $C = 1 - h(\varepsilon)$

BEC: $C = 1 - \varepsilon$

  + channel coding Theorem: for every rate $R < C$, exists code $(2^{nR}, n)$ with $\lim_{n\to\infty} \lambda^{(n)} = 0$ (reliable decoding is possible)

* Gaussian channel:
  + AWGN power constraint: $\frac{1}{n} \sum_i^n x_i^2 \le P$
  + Noise: $Z_i \sim \mathcal{N}(0, N)$; $N = \int z^2 p_Z(z)\,dz = E[Z^2]$ (power)
  + SNR: $\frac{P}{N}$; $y_i = x_i + z_i$, $E[Y^2] = P + N$
  + Gaussian channel code: a (M,n) gaussian channel code C with power constraint P consists of message index set $\{1, 2, \dots, M\}$ with

$$C = \{x(1), x(2), \dots, x(M)\} \text{ with}$$

$$x(i) = (x_1, x_2, \dots, x_n),\ x_i \in \mathbb{R} \text{ and } \frac{1}{n} \sum_i^n x_i \le P$$

  + Channel capacity: $C = \frac{1}{2} \log(1 + \frac{P}{N})$, $p_X^* \sim \mathcal{N}(0, P)$

* Jensen's inequality: $E[\phi(X)] \ge \phi(E[X])$ for convex funcs.

* Kraft inequality: $\sum_x D^{-l(x)} \le 1$ for any prefix D-ary code
* $L(C) = \sum_x p(x) l(x)$. A "good code" have $l(x) = \lceil -\log p(x) \rceil$
* Non-binary Huffman code: # dummy symbol $= (1 - |X|) \,\%\, (D - 1)$
* Entropy bound on single variable compression: $H_D(X) \le L^*(C) \le H_D(X) + 1$
* Cost of miscoding: $H(p) + D(p||q)$
* Vector source coding for memoryless source: $x \in X^n \to \boxed{X^n \to C} \to c \in C$
  + Code rate $R = \frac{1}{n} \sum_x p(x) l(x) = \frac{1}{n} L$ (compression rate) ↓↓ (bits/symbol)

$$H(X) \le R \le H(X) + \varepsilon \ (H(X) + \tfrac{1}{n}) \text{ for large } n$$

  + Sample entropy: $-\frac{1}{n} \log p(x)$
  + Typical set: $T_\varepsilon^{(n)} = \{x \in X^n : |-\frac{1}{n} \log p(x) - H(X)| < \varepsilon\}$

$$|T_\varepsilon^{(n)}| 2^{-n(H(X)+\varepsilon)} \le P(X \in T_\varepsilon^{(n)}) \le |T_\varepsilon^{(n)}| 2^{-n(H(X)-\varepsilon)}$$

Lossless compression

  + AEP: $\lim_{n\to\infty} P(X \in T_\varepsilon^{(n)}) = 1$
  + Coding scheme: $X^n = T_\varepsilon^{(n)} \cup \bar{T}_\varepsilon^{(n)}$
    For $x \in T_\varepsilon^{(n)}$: assign binary index from $\{1, 2, \dots, |T_\varepsilon^{(n)}|\}$
    codeword is: 0 + binary index
    For $x \in \bar{T}_\varepsilon^{(n)}$: codeword is: 1 + itself
* Vector source coding for Markov source:
  + Entropy rate: $H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1 \dots X_n) \le H(X_2|X_1)$
  + Conditional entropy rate: $H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n | X_{n-1} \dots X_1)$
  + For stationary process: $H(\mathcal{X}) = H'(\mathcal{X})$
    $\lim_{n\to\infty} R_n^* = H(\mathcal{X})$
  + For stationary Markov chain: $H(\mathcal{X}) = H(X_2|X_1)$

$$= -\sum_i \sum_j z_i p_{ij} \log p_{ij}$$

* Differential Entropy

	discrete	continuous
$H(X)$	$-\sum_{x\in\mathcal{X}} p_X(x) \log p_X(x)$	$-\int_{\mathcal{X}} p_X(x) \log p_X(x) dx$
$H(X\|Y)$	$-\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p_{XY}(x,y) \log p_{X\|Y}(x\|y)$	$-\int_{\mathcal{X}} \int_{\mathcal{Y}} p_{XY}(x,y) \log p_{X\|Y}(x\|y) dy dx$
$D(f(x)\|\|g(x))$	$\sum_{x\in\mathcal{X}} f(x) \log \frac{f(x)}{g(x)}$	$\int_{\mathcal{X}} f(x) \log \frac{f(x)}{g(x)} dx$
$I(X;Y)$	$\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p_{XY}(x,y) \log \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}$	$\int_{\mathcal{X}} \int_{\mathcal{Y}} p_{XY}(x,y) \log \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)} dx dy$

	discrete	continuous
Non-negativity of $H(X)$?	Yes: $H(X) \ge 0$	No: $H(X) < 0$ possible
Maximum entropy	$H(X) \le \log \|\mathcal{X}\|$	$H(X) \le \log 2\pi e \sigma^2 \cdot \frac{1}{2}$
Entropy maximizing $p_X(x)$	uniform $p_X(x) = \frac{1}{\|\mathcal{X}\|}$	Gaussian $\frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-m)^2}{2\sigma^2})$
Shift by constant c	$H(X + c) = H(X)$	
Multiply by constant a	$H(aX) = H(X)$	$H(aX) = H(X) + \log\|a\|$
Cond. reduces entropy	$H(X\|Y) \le H(X)$	
Chain rule	$H(X,Y) = H(X\|Y) + H(Y)$	
Mutual Information	$I(X;Y) = D(p_{XY}(x,y)\|\|p_X(x)p_Y(y))$	
Non-negativity	$D(p_{XY}(x,y)\|\|p_X(x)p_Y(y)) \ge 0 \Rightarrow I(X;Y) \ge 0$	

* Multivariate Gaussian:

$$p_X(x) = (1 / (2\pi)^{n/2} |\det K|^{\frac{1}{2}}) \exp(-\tfrac{1}{2}(x - m)^T K^{-1} (x - m))$$

$$m = E[X_1], \dots, E[X_n]$$

$$K = E[(X - m)^T (X - m)];\ k_{ij} = cov(X_i, X_j) = E[(X_i - m_i)^T (X_j - m_j)]$$

$$H(X) \le \tfrac{1}{2} \log (2\pi e)^n |\det K|$$

* Rate-distortion code: (M,n) codes:

$$x \in X^n \to \boxed{X^n \to W} \xrightarrow{w \in W} \boxed{W \to \hat{X}^n} \to \hat{x}$$

Lossy compression

  + Code rate: $R = \frac{1}{n} \log M$ (compression rate) ↓↓
  + Hamming distortion: $d(x_i, \hat{x}_i) = \begin{cases} 0 & x_i = \hat{x}_i \\ 1 & \text{otherwise} \end{cases}$. $d(x, \hat{x}) = \frac{1}{n} \sum_i^n d(x_i, \hat{x}_i)$
  + Expected distortion: $D = \sum_{x \in X^n} p_X(x)\, d(x, g \circ f(x))$
  + Achievable (R, D): a $(2^{nR}, n)$ code with f, g s.t. $\lim_{n\to\infty} E[d(X, g \circ f(X))] \le D$
  + Rate distortion function: $R(D) = \min_{p_{\hat{X}|X}(\cdot): Ed(x,\hat{x}) \le D} I(X; \hat{X})$

$$\text{with } E[d(X, \hat{X})] = \sum_{x \in X} \sum_{\hat{x} \in \hat{X}} p_{X,\hat{X}}(x, \hat{x})\, d(x, \hat{x}) \quad (*)$$

  + $R(D)$ for discrete sources: use test channel $\hat{X} \xrightarrow{p_{X|\hat{X}}(\cdot)} X$
    + Parameterize $p_{\hat{X}|X}(\cdot)$ w.r.t the distortion ($d(x, \hat{x}) = \infty \Rightarrow p_{X|\hat{X}}(x|\hat{x}) = 0$)

+ Channel capacity : $C = \frac{1}{2} \log(1 + \frac{P}{N})$, $p_X \sim \mathcal{N}(0, P)$

* Jensen's inequality : $E[\phi(X)] \geq \phi(E[X])$ for convex funcs.
* Correlation coefficient:

$$Cov(X, Y) = E[(X - \bar{X})^T (Y - \bar{Y})]$$

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)\, Var(Y)}}$$

* Convolution: express the sum of 2 distributions:

$$p_Z(z) = p_Z(X+Y) = \int_x p_X(x)\, p_Y(z - x)\, dx$$

with $E[d(X, \hat{X})] = \sum_{x \in \mathcal{X}} \sum_{\hat{x} \in \hat{\mathcal{X}}} p_{X,\hat{X}}(x, \hat{x})\, d(x, \hat{x})$

+ $R(D)$ for discrete sources : use Test channel $\hat{X} \xrightarrow{p_{X|\hat{X}}(\cdot)} X$
  + Parameterize $p_{\hat{X}|X}(\cdot)$ w.r.t the distortion ($d(x, \hat{x}) = \infty \Rightarrow p_{X|\hat{X}}(x|\hat{x}) = 0$)
  + " $p_{\hat{X}}(\cdot)$
  + Derive parameters using (*) and $\sum_{\hat{x} \in \hat{\mathcal{X}}} p_{X,\hat{X}}(x, \hat{x}) = p_X(x)$
  + Compute $H(X|\hat{X})$

$$R(D) = \min I(X; \hat{X}) = H(X) - \max H(X|\hat{X})$$

+ $R(D)$ for continuous sources :
  + Expected distortion: $D = E[(X - \hat{X})^2]$; Euclidean distance : $d(x, \hat{x}) = (x - \hat{x})^2$
  + $R(D)$ for gaussian source X with euclidean distance : $R(D) = \begin{cases} \frac{1}{2} \log \frac{\sigma^2}{D} & 0 \leq D \leq \sigma^2 \\ 0 & D \geq \sigma^2 \end{cases}$

$$\int \log_a x\, dx = x \log_a x - \frac{x}{\ln a} = \frac{x \ln x - x}{\ln a}$$

$$\int \ln(ax)\, dx = x \ln(ax) - x$$

$$\int \ln(ax + b)\, dx = \frac{(ax + b) \ln(ax + b) - (ax + b)}{a}$$

$\frac{d}{dx}(c) = 0$	$\frac{d}{dx}(cx) = c$	$\frac{d}{dx}(x^c) = cx^{c-1}$
$\frac{d}{dx}(c^x) = c^x \ln(c) \quad c > 0$	$\frac{d}{dx}(x^x) = x^x(1 + \ln x)$	$\frac{d}{dx}(e^x) = e^x$
$\frac{d}{dx}\left(\frac{1}{x}\right) = \frac{-1}{x^2}$	$\frac{d}{dx}\left(\frac{1}{x^2}\right) = \frac{-2}{x^3}$	$\frac{d}{dx}\left(\frac{1}{x^n}\right) = \frac{-n}{x^{n+1}}$
$\frac{d}{dx}(\sqrt{x}) = \frac{1}{2\sqrt{x}} \quad x > 0$	$\frac{d}{dx}(\sqrt[3]{x}) = \frac{1}{3 \cdot \sqrt[3]{x^2}}$	$\frac{d}{dx}(\sqrt[n]{x}) = \frac{1}{n \cdot \sqrt[n]{x^{n-1}}}$
$\frac{d}{dx}\left(\frac{1}{\sqrt{x}}\right) = \frac{-1}{2\sqrt{x^3}}$	$\frac{d}{dx}\left(\frac{1}{\sqrt[3]{x}}\right) = \frac{-1}{3 \cdot \sqrt[3]{x^4}}$	$\frac{d}{dx}\left(\frac{1}{\sqrt[n]{x}}\right) = \frac{-1}{n \cdot \sqrt[n]{x^{n+1}}}$
$\frac{d}{dx}(\ln x) = \frac{1}{x} \quad x > 0$	$\frac{d}{dx}(x \cdot \ln x) = \ln x + 1$	$\frac{d}{dx}(\log_c x) = \frac{1}{x \ln c} \quad c > 0 \quad c \neq 1$
$\frac{d}{dx}\left(\frac{1}{\ln x}\right) = \frac{-1}{x(\ln x)^2}$	$\frac{d}{dx}\left(\frac{1}{x \cdot \ln x}\right) = \frac{-(\ln x + 1)}{(x \cdot \ln x)^2}$	$\frac{d}{dx}\left(\frac{1}{\log_c x}\right) = \frac{-1}{x \cdot \ln c \cdot (\log_c x)^2}$
$\frac{d}{dx}\left(\frac{1}{x+1}\right) = \frac{-1}{(x+1)^2}$	$\frac{d}{dx}\left(\frac{1}{(x+1)^2}\right) = \frac{-2}{(x+1)^3}$	$\frac{d}{dx}\left(\frac{1}{(x+1)^n}\right) = \frac{-n}{(x+1)^{n+1}}$
$\frac{d}{dx}\left(\frac{1}{\sqrt{x+1}}\right) = \frac{-1}{2 \cdot \sqrt{(x+1)^3}}$	$\frac{d}{dx}\left(\frac{1}{\sqrt[3]{x+1}}\right) = \frac{-1}{3 \cdot \sqrt[3]{(x+1)^4}}$	$\frac{d}{dx}\left(\frac{1}{\sqrt[n]{x+1}}\right) = \frac{-1}{n \cdot \sqrt[n]{(x+1)^{n+1}}}$

Karush-Kuhn-Tucker (KKT) conditions

Theorem

Let $f, \boldsymbol{h}, \boldsymbol{g} \in C^1$ and $\boldsymbol{x}^*$ be a regular local minimizer for

$$\text{minimize } f(\boldsymbol{x}) \qquad \text{subject to } \boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{0},\ \boldsymbol{g}(\boldsymbol{x}) \geq \boldsymbol{0}.$$

Then, there exists $\boldsymbol{\lambda}^* \in \mathbb{R}^n$ and $\boldsymbol{\mu}^* \in \mathbb{R}^p$ such that

1. $\nabla f(\boldsymbol{x}^*) + D\boldsymbol{h}(\boldsymbol{x}^*)^T \boldsymbol{\lambda}^* + D\boldsymbol{g}(\boldsymbol{x}^*)^T \boldsymbol{\mu}^* = \boldsymbol{0}$;
2. $\boldsymbol{\mu}^* \geq \boldsymbol{0}$;
3. $\boldsymbol{g}(\boldsymbol{x}^*)^T \boldsymbol{\mu}^* = 0$.

The KKT conditions also conclude

4. $\boldsymbol{h}(\boldsymbol{x}^*) = \boldsymbol{0}$;
5. $\boldsymbol{g}(\boldsymbol{x}^*) \leq \boldsymbol{0}$.

Let $Z \sim \phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{x^2}{2\sigma^2})$ and let $X \sim p_X(x)$ with equal variance

$$\int_{\mathcal{X}} p_X(x) \ln \phi(x) dx = \int_{\mathcal{X}} \phi(x) \ln \phi(x) dx$$

$$\int_{\mathcal{X}} p_X(x) \left(\frac{x^2}{2\sigma^2} + \ln \sqrt{2\pi\sigma^2}\right) dx = \int_{\mathcal{X}} \phi(x) \left(\frac{x^2}{2\sigma^2} + \ln \sqrt{2\pi\sigma^2}\right) dx$$

$$p_{Y|X}(y|x) = \begin{bmatrix} p & 1-p & 0 & 0 \\ 1-p & p & 0 & 0 \\ 0 & 0 & q & 1-q \\ 0 & 0 & 1-q & q \end{bmatrix}$$

This is two parallel BSCs, the p channel and the q channel At any time, one of the two channels is used. Create a new random variable Z:

$$p_Z(z) = \begin{cases} \alpha & z = \text{"use p channel"} \\ 1 - \alpha & z = \text{"use q channel"} \end{cases}$$

Because z can be determined from the channel output y, $X \to Y \to Z$. Write mutual information $I(X; Y, Z)$ two ways:

$$I(X; Y, Z) = I(X; Z) + I(X; Y|Z)$$
$$I(X; Y, Z) = I(X; Y) + I(X; Z|Y)$$

Since $X \to Y \to Z$, $I(X; Z|Y) = 0$, so:

$$\begin{aligned} I(X; Y) &= I(X; Z) + I(X; Y|Z) \\ &= H(Z) - H(Z|X) + \alpha I(X_p; Y_p) + (1 - \alpha) I(X_q; Y_q) \\ &= h(\alpha) + \alpha I(X_p; Y_p) + (1 - \alpha) I(X_q; Y_q) \\ C &= \max_{\alpha} h(\alpha) + \alpha C_p + (1 - \alpha) C_q, \end{aligned}$$

where $C_p = 1 - h(p)$ and $C_q = 1 - h(q)$ are the capacities of the two BSCs. The solution, found by taking the derivative with respect to α, is:

$$\alpha^* = \frac{2^{C_p}}{2^{C_p} + 2^{C_q}}.$$

Then, the capacity is:

$$C = \log\left(2^{1-h(p)} + 2^{1-h(q)}\right).$$

Proof To find the optimal power allocation $P_1^*, \ldots, P_k^*$; Lagrange multipliers will be used. Since Z_j are independent:

$$\begin{aligned} I(X_1, \ldots, X_k; Y_1, \ldots, Y_k) &= H(Y_1, \ldots, Y_k) - H(Y_1, \ldots, Y_k | X_1, \ldots, X_k) \\ &= H(Y_1, \ldots, Y_k) - H(Z_1, \ldots, Z_k | X_1, \ldots, X_k) \\ &= H(Y_1, \ldots, Y_k) - H(Z_1, \ldots, Z_k) \\ &= H(Y_1, \ldots, Y_k) - \sum_i H(Z_i) \qquad \text{independence of } Z_i \\ &\leq \sum_i H(Y_i) - \sum_i H(Z_i) \\ &\leq \sum_i \frac{1}{2} \log\left(1 + \frac{P_i}{N_i}\right) \end{aligned}$$

Using Lagrange multipliers, the Largrangian J is:

$$J(P_1, \ldots, P_k) = \sum_j \frac{1}{2} \log(1 + \frac{P_j}{N_j}) + \lambda \sum_j P_j$$

and λ is the Lagrange multiplier. For $i = 1, \ldots, k$:

$$\begin{aligned} \frac{\partial}{\partial P_i} J(P_1, \ldots, P_k) &= \frac{1}{2} \frac{1}{P_i + N_i} + \lambda = 0 \\ P_i &= \underbrace{\left(-\frac{1}{2\lambda}\right)}_{\stackrel{\text{def}}{=} \nu} - N_i \\ P_i &= \nu - N_i \end{aligned}$$

Equality is achieved by:

$$(X_1, \ldots, X_k) \sim \mathcal{N}(0, \begin{bmatrix} P_1 & 0 & \cdots & 0 \\ 0 & P_2 & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & P_k \end{bmatrix}) \qquad (9.100)$$

Achieve the goal by solving an optimization problem:

$$\max_{P_1, \ldots, P_k} \left(\frac{1}{2} \log(1 + \frac{P_1}{N_1}) + \cdots + \frac{1}{2} \log(1 + \frac{P_k}{N_k})\right) \quad \text{such that} \quad \sum P_i = P$$

Alternatively, note that $\lambda = -\frac{1}{2} \frac{1}{P_i + N_i}$ so another solution is:

$$-\frac{1}{2} \frac{1}{P_1 + N_1} = -\frac{1}{2} \frac{1}{P_2 + N_2} = \cdots = -\frac{1}{2} \frac{1}{P_k + N_k}, \qquad (9.105)$$

that is, $P_i + N_i$ should be a constant, defined to be ν.

But, P_i must be non-negative. Let P_i^* be the optimal input powers. Use the Kuhn-Tucker conditions to verify the solution is:

$$P_i^* = (\nu - N_i)^+. \quad \square \qquad (9.106)$$