

Preliminares

Definition 0.1. The normal distribution $N(\mu, \sigma^2)$ has the probability density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

Definition 0.2. The joint normal distribution $N(\mu, K)$ is defined by probability density function:

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n \det(K)}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^\top K^{-1}(\mathbf{x} - \mu)\right)$$

Theorem 0.1 (Joint normality implies marginal normality). If $\mathbf{X} = (X_1, X_2, \dots, X_n)$ follows a joint normal distribution, then any linear combination of \mathbf{X} follows normal distribution.

1 Entropy, mutual information, divergence

Definition 1.1 (Entropy). The entropy of a random variable $X \in \mathcal{X}$ is defined by

$$H(X) = - \sum_{x \in \mathcal{X}} p_X(x) \log p_X(x)$$

Definition 1.2 (KL Divergence). The **KL Divergence** between two probability distributions p, q on the same support \mathcal{X} is defined by

$$D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

Definition 1.3 (Mutual information). The mutual information between two random variables X, Y is defined by

$$I(X; Y) = H(X) - H(X | Y)$$

It can also be expressed as the KL divergence between the joint distribution $p_{X,Y}$ and the product of marginal distribution $p_X p_Y$:

$$I(X; Y) = D(p_{X,Y} || p_X \cdot p_Y)$$

Proposition 1.1. Conditioning does not increase entropy

$$H(X | Y) \leq H(X)$$

A nice trick for reasoning about the entropy of "sum of random variables":

$$\begin{aligned} H(X + Y | X) &= \sum_{x \in \mathcal{X}} p_X(x) H(X + Y | X = x) \\ &= \sum_{x \in \mathcal{X}} p_X(x) H(x + Y | X = x) \\ &= \sum_{x \in \mathcal{X}} p_X(x) H(Y | X = x) \\ &= H(Y | X) \end{aligned}$$

1.1 Convexity

Definition 1.4. A function f is convex over the interval $x_1 \leq x \leq x_2$ if for all $0 \leq t \leq 1$:

$$(1-t)f(x_1) + tf(x_2) \geq f((1-t)x_1 + tx_2)$$

Some notable functions' convexity:

1. Entropy is a concave function with respect to the probability distribution
2. Mutual information $I(X; Y)$ is concave with respect to p_X and is convex with respect to $p_{Y|X}$ for some fixed p_X
3. \log is a concave function
4. The negative of a convex function is concave
5. The sum or product of a linear function and a concave function is concave
6. The composition of a linear function and a concave function is concave

Theorem 1.1 (Jensen's Inequality). If f is a convex function and X is a random variable, then

$$E[f(X)] \geq f(E[X])$$

1.2 Markov chain

Definition 1.5. Random variables X, Y, Z form a Markov chain (denoted by $X \rightarrow Y \rightarrow Z$) if $p(z | x, y) = p(z | y)$

Theorem 1.2 (Data processing inequality). If $X \rightarrow Y \rightarrow Z$, then

$$I(X; Z) \leq I(Y; Z)$$

Proposition 1.2.

$$I(X; Y | Z) = 0 \Leftrightarrow X \rightarrow Z \rightarrow Y$$

Proposition 1.3. If $X \rightarrow Y \rightarrow Z$, then $Z \rightarrow Y \rightarrow X$

Theorem 1.3. Let X, Y be random variables. $I(X; Y)$ is concave with respect to the probability distribution of X . For a fixed marginal distribution of X , $I(X; Y)$ is convex with respect to $f_{Y|X}$

Theorem 1.4 (Fano's inequality). Let $X \rightarrow Y \rightarrow \hat{X}$ represent an encode-decode process, where $X, \hat{X} \in \mathcal{X}$ have the same support. Let e denote decoding error $\hat{X} \neq X$, then:

$$H(X | Y) \leq h(P_e) + P_e \log(|\mathcal{X}|)$$

2 Entropy rate

Definition 2.1. A stochastic process is stationary if for all integer $n \geq 1$ and integer time shift (possibly negative) l :

$$P[X_1^n] = P[X_{1+l}^{n+l}]$$

Definition 2.2. A Markov chain is time invariant if for all $n \geq 1$:

$$P[X_{n+1} | X_n] = P[X_2 | X_1]$$

Definition 2.3. There are two definitions of entropy rate:

1. $H(\mathcal{X}) = \lim_{n \rightarrow \infty} \frac{1}{n} H(X_1^n)$
2. $H'(\mathcal{X}) = \lim_{n \rightarrow \infty} H(X_n | X_1^{n-1})$

Theorem 2.1. If X_i is a stationary process, then the two entropy rates both exist and are equal

Lemma 2.1.1. If X_i is a stationary process, then $H(X_n | X_1^{n-1})$ is non-increasing with respect to n and has a limit

3 Asymptotic equipartition property

Theorem 3.1 (Weak law of large numbers (WLLN)). *Let $X \sim p_X$ be a random variable, and let $X_i \sim p_X$ be i.i.d. random variables following the same distributions, then:*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n X_i = E[X]$$

In other words, for any $\delta, \epsilon > 0$, there exists $N > 0$ such that for $n > N$:

$$P\left[\left|\frac{1}{n} \sum_{i=1}^n X_i - E[X]\right| < \delta\right] < \epsilon$$

Recall that the entropy of a random variable can be defined as $H(X) = -E[\log(P_X)]$, so the weak law of large numbers also applies: the joint distribution of i.i.d. X_i converges to the probability $2^{-nH(X)}$.

Definition 3.1 (Typical sequence). *For random variable $X \sim p_X$, the set of **typical sequence** A_ϵ^n denotes the set of sequences whose probability is close to the entropy of the random variable*

$$A_\epsilon^{(n)} = \{\mathbf{x} \in \mathcal{X}^n \mid |-\frac{1}{n} \log p_{\mathbf{X}}(\mathbf{x}) - H(X)| < \epsilon\}$$

Theorem 3.2 (Asymptotic equipartition property). *Let $X \sim p_X$ and let $\mathbf{X} \in \mathcal{X}^n$ denote i.i.d. sequence following the same distribution, then:*

1. $\lim_{n \rightarrow \infty} P[\mathbf{X} \in A_\epsilon^{(n)}] = 1$
2. $|A_\epsilon^{(n)}| \leq 2^{n(H-\epsilon)}$
3. For sufficiently large n , $|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H+\epsilon)}$

Proof. A sketch of the proof (1) is the direct consequence of the weak law of large numbers

(2): begin with **the sum of probability of all possible sequence is 1**

(3): begin with **for sufficiently large n , $P[\mathbf{x} \in A_\epsilon^{(n)}] \geq 1 - \epsilon$** □

4 Data compressions

Theorem 4.1 (Kraft inequality). *A D -ary prefix code for a finite set of m symbols \mathcal{X} exists if and only if the lengths of the code words l_1, l_2, \dots, l_m satisfy:*

$$\sum_{x \in \mathcal{X}} D^{-l(x)} \leq 1$$

Using the method of Lagrange multiplier, we can optimize the expected codeword length $E[L]$ under the constraint fo Kraft inequality:

$$l_i = \lceil \log_D \frac{1}{p_i} \rceil$$

Theorem 4.2 (Optimal code length). $H_D(X) \leq E[L] \leq H_D(X) + 1$

Theorem 4.3 (McMillan inequality). *A D -ary uniquely decodable code exists if and only if*

$$\sum_i D^{-l_i} \leq 1$$

Requires knowledge of source probabilities.

```

graph LR
    Root(( )) ---|0| Node0_9((0.9))
    Root ---|1| Node1_0_2(( ))
    Node1_0_2 ---|0| Node1_0_2_0_2((0.2))
    Node1_0_2 ---|1| Node1_0_2_1_0_15(( ))
    Node1_0_2_1_0_15 ---|0| Node1_0_2_1_0_15_0_15((0.15))
    Node1_0_2_1_0_15 ---|1| Node1_0_2_1_0_15_1_0_15_0_15(( ))
    Node1_0_2_1_0_15_1_0_15_0_15 ---|0| Node1_0_2_1_0_15_1_0_15_0_15_0_15((0.15))
    Node1_0_2_1_0_15_1_0_15_0_15 ---|1| Node1_0_2_1_0_15_1_0_15_0_15_1_0_1(( ))
    Node1_0_2_1_0_15_1_0_15_0_15_1_0_1 ---|0| Node1_0_2_1_0_15_1_0_15_0_15_1_0_1_0_1((0.1))
    Node1_0_2_1_0_15_1_0_15_0_15_1_0_1 ---|1| Node1_0_2_1_0_15_1_0_15_0_15_1_0_1_1_0((0.1))
    Node1_0_2_1_0_15_1_0_15_0_15_1_0_1_0_1 ---> Output((1.0))
    Node1_0_2_1_0_15_1_0_15_0_15_1_0_1_1_0 ---> Output
  
```

5 Channel capacity

Definition 5.1. A discrete channel consists of a discrete set of words \mathcal{W} , two sets of symbols \mathcal{X}, \mathcal{Y} , an encoder $\mathcal{W} \rightarrow \mathcal{X}$, a decoder $\mathcal{Y} \rightarrow \mathcal{W}$, and a channel described by the conditional distribution $p_{Y|X}$:

For this section we are only concerned with **memoryless** channel: given X_n , Y_n is independent of all other variables:

For a fixed $p_{Y|X}$, **information channel capacity** is defined by

5.1 Channel rate

Definition 5.3 (Decoding error). *For a single word $w \in \mathcal{W}$, the **conditional probability of error** is defined by*

Denote the maximal probability of error across all words by

Denote the average probability of error across all words by

The **rate of an (M, n) -code** denotes the number of bits of information transmitted per channel use:

$$R = \frac{\log(M)}{n}$$

Definition 5.4 (Achievable rate). A rate $R > 0$ is **achievable** if there exists a sequence of $(2^{nR}, n)$ -code such that:

$$\lim_{n \rightarrow \infty} \lambda^{(n)} = 0$$

The **capacity** of a channel is the supremum of all achievable rates

5.2 Jointly typical sequence

Let $(\mathbf{X}, \mathbf{Y}) \in \mathcal{X}^n \times \mathcal{Y}^n$ be i.i.d. sequences drawn according to the joint distribution $P_{X,Y}$, then by the weak law of large numbers, $\lim_{n \rightarrow \infty} -\frac{1}{n} \log(p_{X,Y}(\mathbf{x}, \mathbf{y})) = H(X, Y)$. We can define notion of typicality on the joint sequence similar to the notion of typicality on individual sequences

Definition 5.5 (Jointly typical sequence). The set of jointly typical sequence $A_\epsilon^{(n)}$ is the set of (\mathbf{x}, \mathbf{y}) such that:

1. $|\frac{1}{n} \log(p_X(\mathbf{x})) - H(X)| \leq \epsilon$
2. $|\frac{1}{n} \log(p_Y(\mathbf{y})) - H(Y)| \leq \epsilon$
3. $|\frac{1}{n} \log(p_{X,Y}(\mathbf{x}, \mathbf{y})) - H(X, Y)| \leq \epsilon$

In other words, (\mathbf{x}, \mathbf{y}) is jointly typical if \mathbf{x} and \mathbf{y} are each individually typical according to their marginal distribution, and (\mathbf{x}, \mathbf{y}) is typical according to the joint distribution

Theorem 5.1 (Joint asymptotic equipartition property). Let $(\mathbf{X}, \mathbf{Y}) \in \mathcal{X}^n \times \mathcal{Y}^n$ be i.i.d. sequences drawn according to $p_{X,Y}$, then

1. $\lim_{n \rightarrow \infty} P[(\mathbf{X}, \mathbf{Y}) \in A_\epsilon^{(n)}] = 1$
2. $|A_\epsilon^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)}$
3. For sufficiently large n :

$$|A_\epsilon^{(n)}| \geq (1 - \epsilon)2^{n(H(X,Y)-\epsilon)}$$

4. Let $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ be drawn independently from p_X and p_Y , then

$$P[(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) \in A_\epsilon^n] \leq 2^{-n(I(X;Y)-3\epsilon)}$$

5. Let $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ be drawn independently from p_X and p_Y , then for sufficiently large n :

$$P[(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) \in A_\epsilon^n] \geq (1 - \epsilon)2^{-n(I(X;Y)+3\epsilon)}$$

Proof. (1) is the direct consequence of the weak law of large numbers

(2):

$$\begin{aligned} 1 &= \sum_{(\mathbf{x}, \mathbf{y}) \in \mathcal{X}^n \times \mathcal{Y}^n} p(\mathbf{x}, \mathbf{y}) \\ &\geq \sum_{(\mathbf{x}, \mathbf{y}) \in A_\epsilon^{(n)}} p(\mathbf{x}, \mathbf{y}) \\ &\geq \sum_{(\mathbf{x}, \mathbf{y}) \in A_\epsilon^{(n)}} 2^{-n(H(X,Y)+\epsilon)} \\ &= |A_\epsilon^{(n)}| 2^{-n(H(X,Y)+\epsilon)} \end{aligned}$$

(3): for sufficiently large n , $P[(\mathbf{X}, \mathbf{Y}) \in A_\epsilon^{(n)}] \geq 1 - \epsilon$. On the other hand:

$$\begin{aligned} P[(\mathbf{X}, \mathbf{Y}) \in A_\epsilon^{(n)}] &= \sum_{(\mathbf{x}, \mathbf{y}) \in A_\epsilon} P[(\mathbf{X}, \mathbf{Y}) = (\mathbf{x}, \mathbf{y})] \\ &\leq \sum_{(\mathbf{x}, \mathbf{y}) \in A_\epsilon} 2^{-n(H_{X,Y} - \epsilon)} \\ &= |A_\epsilon| \cdot 2^{-n(H_{X,Y} - \epsilon)} \end{aligned}$$

Putting the two inequalities above together:

$$|A_\epsilon| \cdot 2^{-n(H_{X,Y} - \epsilon)} \geq 1 - \epsilon$$

(4): consider the probability of $(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}})$ falling into the jointly typical set:

$$\begin{aligned} P[(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) \in A_\epsilon^{(n)}] &= \sum_{(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \in A_\epsilon^{(n)}} P[\tilde{\mathbf{X}} = \tilde{\mathbf{x}}, \tilde{\mathbf{Y}} = \tilde{\mathbf{y}}] \\ &= \sum_{(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \in A_\epsilon^{(n)}} P[\tilde{\mathbf{X}} = \tilde{\mathbf{x}}] \cdot P[\tilde{\mathbf{Y}} = \tilde{\mathbf{y}}] \\ &\leq \sum_{(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \in A_\epsilon^{(n)}} 2^{-n(H_X - \epsilon)} \cdot 2^{-n(H_Y - \epsilon)} \\ &= |A| \cdot 2^{-n(H_X - \epsilon)} \cdot 2^{-n(H_Y - \epsilon)} \\ &\leq 2^{n(H(X,Y) + \epsilon)} \cdot 2^{-n(H_X - \epsilon)} \cdot 2^{-n(H_Y - \epsilon)} \end{aligned}$$

(5): Using similar logic as shown above:

$$\begin{aligned} P[(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) \in A_\epsilon^{(n)}] &= \sum_{(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \in A_\epsilon^{(n)}} P[\tilde{\mathbf{X}} = \tilde{\mathbf{x}}, \tilde{\mathbf{Y}} = \tilde{\mathbf{y}}] \\ &= \sum_{(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \in A_\epsilon^{(n)}} P[\tilde{\mathbf{X}} = \tilde{\mathbf{x}}] \cdot P[\tilde{\mathbf{Y}} = \tilde{\mathbf{y}}] \\ &\geq \sum_{(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \in A_\epsilon^{(n)}} 2^{-n(H_X + \epsilon)} \cdot 2^{-n(H_Y + \epsilon)} \\ &= |A_\epsilon| 2^{-n(H_X + \epsilon)} \cdot 2^{-n(H_Y + \epsilon)} \\ &\geq (1 - \epsilon) 2^{n(H_{X,Y} - \epsilon)} \cdot 2^{-n(H_X + \epsilon)} \cdot 2^{-n(H_Y + \epsilon)} \end{aligned}$$

□

5.3 Channel coding theorem

Theorem 5.2. *A rate R is achievable if and only if it is below the information channel capacity*

Proof. This is a sketch of proof

First we show that rate below information capacity is achievable: Let R be a rate such that $R < C$, then we can construct a $(2^{nR}, n)$ -code through the following procedure:

1. Find some distribution of p_X and $\epsilon > 0$ such that $R = I(p_X; p_{Y|X}) - 4\epsilon$
2. For each $w \in \{1, \dots, 2^{nR}\}$, the encoding is $f(w) = (x_1(w), x_2(w), \dots, x_n(w))$ where $x_i(w)$ is i.i.d. sampled from p_X
3. Transmit each symbol $x_i(w)$ according to $p_{Y|X}$
4. Upon receiving \mathbf{y} , find some $\hat{w} \in \{1, \dots, 2^{nR}\}$ such that (\mathbf{x}, \mathbf{y}) is jointly typical according to p_X and $p_{Y|X}$ (which can be used to compute $p_{X,Y}$ and p_Y)

For each word w , decoding error occurs if and only if one of two scenarios occurs:

1. $(\mathbf{x}(w), \mathbf{y})$ is not jointly typical, in which case no appropriate \hat{w} can be found in the decoding step
2. there is another $w' \neq w$ such that $(\mathbf{x}(w'), \mathbf{y})$ is also jointly typical

The probability of (1) converges to 0 because $\lim_{n \rightarrow \infty} P[(\mathbf{x}, \mathbf{y}) \notin A_\epsilon^{(n)}] = 0$.

The probability of (2) converges to 0 because \mathbf{y} is independent of $\mathbf{x}(w')$. By joint AEP, the probability of independently sampled sequences being jointly typical is at most $2^{-n(I(X;Y)-3\epsilon)}$, hence:

$$\begin{aligned} \sum_{w' \neq w} P[(\mathbf{x}(w'), \mathbf{y}) \in A_\epsilon^{(n)}] &= (2^{nR} - 1)P[(\mathbf{x}(w'), \mathbf{y}) \in A_\epsilon^{(n)}] \\ &\leq (2^{nR} - 1)2^{-n(I(X;Y)-3\epsilon)} \\ &\leq 2^{nR-n(I(X;Y)-3\epsilon)} \\ &= 2^{n(I(X;Y)-4\epsilon)-n(I(X;Y)-3\epsilon)} \\ &= 2^{-n\epsilon} \end{aligned}$$

Which also converges to 0 as $n \rightarrow \infty$

Converse: achievable rates are below information capacity

Recall that the channel represents a Markov chain $W \rightarrow \mathbf{X} \rightarrow \mathbf{Y} \rightarrow \hat{W}$. By Fano's inequality, we know that:

$$H(W | \hat{W}) \leq h(p_e) + p_e \cdot \log(|\mathcal{W}|)$$

Consider the mutual information between W and \hat{W} :

$$I(W; \hat{W}) = H(W) - H(W | \hat{W})$$

Assuming that W is uniformly distributed, $H(W) = \log(|\mathcal{W}|) = nR$.

By the data-processing inequality $I(W; \hat{W}) \leq I(\mathbf{X}; \mathbf{Y})$. By the chain rule of mutual information: $I(\mathbf{X}; \mathbf{Y}) \leq nI(X; Y)$. By the definition of information channel capacity $I(X; Y) \leq C$. Putting everything together, we have

$$nR - nC \leq p_e$$

Assuming the rate is achievable, then $\lim_{n \rightarrow \infty} p_e = 0$, which implies $R \leq C$. \square

6 Differential entropy

Definition 6.1. Let $X \in \mathbb{R}$ be a random variable with probability density function f_X . The **differential entropy** of X is defined by

$$h(X) = - \int_{x \in \mathbb{R}} f_X(x) \ln(f_X(x)) dx = -E[\ln(f_X(X))]$$

Some notable results:

- For uniform distribution over $[0, a]$, $f_X(x) = \frac{1}{a}$, $h(X) = \ln(a)$
- For $X \sim N(0, \sigma^2)$, $h(X) = \frac{1}{2} \ln(2\pi e \sigma^2)$
- For multivariate normal $\mathbf{X} \sim N(\mathbf{0}, K)$, $h(\mathbf{X}) = \frac{1}{2} \log((2\pi e)^n \det K)$
- For some constant a , $h(X + a) = h(X)$
- For some constant a , $h(aX) = \ln(|a|) + h(X)$

Theorem 6.1. Let X be continuous random variable with 0 mean and σ^2 variance, then

$$h(X) \leq \frac{1}{2} \ln(2\pi e \sigma^2)$$

Equality is reached if and only if X is Gaussian

7 Gaussian channel

Definition 7.1 (Information channel capacity). Let $Y = X + Z$, where $Z \stackrel{s}{\leftarrow} N(0, \sigma^2)$ and $E[X^2] \leq P$ for some power level constraint P . The **information channel capacity** is defined by

$$C^I = \max_{f_X: E[X^2] \leq P} I(X; Y)$$

Theorem 7.1. The information channel capacity of a Gaussian channel is

$$\max_{f_X: E[X^2] \leq P} I(X; Y) = \frac{1}{2} \log \left(1 + \frac{P}{\sigma^2} \right) \quad (1)$$

Where P is the power constraint, and σ^2 is the variance of the Gaussian noise. The maximum is achieved when X follows Gaussian distribution $X \sim N(0, P)$

7.1 Parallel Gaussian Channel

Suppose for $1 \leq l \leq n$, $Y_l = X_l + Z_l$, where $Z_l \sim N(0, N_l)$ is independent Gaussian noise, then $\mathbf{Y} = (Y_l)_{l=1}^n$ can be seen as the output of n parallel Gaussian channels combined into a single channel. The capacity of the combined channel is

$$C = \max_{f_{\mathbf{X}}} I(\mathbf{X}; \mathbf{Y})$$

The optimal strategy for maximizing the capacity is to make X_l independent and individually Gaussian. If there is a combined power constraint $\sum_{l=1}^n E[X_l^2] \leq P$, then *power should be first given to the channel with the lowest amount of noise until the combined noise + power exceeds the next lowest noise (waterfilling)*.

8 Rate distortion theory

Definition 8.1. For some fixed p_X , the **information rate distortion** is defined by

$$R^I(D) = \min_{p_{\hat{X}|X}: E[d(X, \hat{X})] \leq D} I(X; \hat{X})$$

Theorem 8.1. Let X follow Bernoulli(p), then

$$R(D) = \begin{cases} h(p) - h(D) & \text{When } D < \min(p, 1-p) \\ 0 & \text{otherwise} \end{cases}$$

Theorem 8.2. For $X \stackrel{s}{\leftarrow} N(0, \sigma^2)$:

$$R(D) = \begin{cases} \frac{1}{2} \log\left(\frac{\sigma^2}{D}\right) & (D \leq \sigma^2) \\ 0 & \text{otherwise} \end{cases}$$