Information Theory Recap Session

Changyeol Lee (Yonsei University)

Cover and Thomas, Elements of Information Theory (2nd edition)

Entropy

X: a discrete random variable over $\mathcal{X}$ with the PMF(probability mass function) $p(\cdot)$.

The *entropy* of X: a measure of the uncertainty of X

$$H(X) = \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} = \mathbb{E}_{X \sim p}\left[\log \frac{1}{p(X)}\right]$$

Fact. $H(X) \geq 0$.

Conditional Entropy

$$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X = x) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x, y) \log \frac{1}{p(y|x)} = \mathbb{E}_{(X,Y) \sim p}\left[\log \frac{1}{p(Y|X)}\right]$$

Chain Rule

$$H(X_1, X_2, \dots, X_n) = H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1) + \cdots + H(X_n|X_{n-1}, \dots, X_2, X_1)$$

Kullback-Leibler Divergence (Relative Entropy)

Kullback-Leibler divergence between PMFs p and q

$$D(p \parallel q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_{X \sim p}\left[\log \frac{p(X)}{q(X)}\right] = \mathbb{E}_{X \sim p}\left[\log \frac{1}{q(X)}\right] - \mathbb{E}_{X \sim p}\left[\log \frac{1}{p(X)}\right]$$

* $D(p \parallel q) = \infty$ if $\exists\, x \in \mathcal{X}$ s.t. $p(x) > 0$ and $q(x) = 0$.

* $D(p \parallel q) \neq D(q \parallel p)$, i.e., no symmetricity in general

* $D(p \parallel q) + D(q \parallel r) \ngeq D(p \parallel r)$ and $D(p \parallel q) + D(q \parallel r) \nleq D(p \parallel r)$ in general

Chain rule

$$D\big(p(x,y) \parallel q(x,y)\big) = D\big(p(x) \parallel q(x)\big) + D\big(p(y|x) \parallel q(y|x)\big)$$

Mutual Information

Mutual information

- a measure of the amount of information that one RV contains about another RV

$$I(X;Y) = \mathbb{E}_{(X,Y)\sim p}\left[\log \frac{p(X,Y)}{p(X)p(Y)}\right]$$
$$= D\big(p(x,y) \parallel p(x)p(y)\big)$$
$$= H(X) - H(X|Y) = H(Y) - H(Y|X)$$

the reduction in the uncertainty of X (Y) due to the knowledge of Y (X)

Conditional Mutual Information

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

the mutual information of X_1 and X_2, given X_3; not the mutual information of X_1 and $X_2|X_3$.

* $I(X;Y|Z) \nleq I(X;Y)$ and $I(X;Y|Z) \ngeq I(X;Y)$ in general

Chain Rule

$$I(X_1, X_2, \dots, X_n; Y) = I(X_1;Y) + I(X_2;Y|X_1) + I(X_3;Y|X_2,X_1) + \cdots + I(X_n;Y|X_{n-1},\dots,X_2,X_1)$$

Information Inequality

Theorem. $D(p \parallel q) \geq 0$ with equality iff $p = q$.

Corollary. $I(X;Y) \geq 0$ with equality iff X and Y are independent.

Corollary. $H(X|Y) \leq H(X)$, i.e., *conditioning only reduces entropy*.

Corollary. $H(X) \leq \log|\mathcal{X}|$ with equality iff p is the uniform distribution.

Data-processing Inequality

By chain rule,

$$I(X;Z) + I(X;Y|Z) = \boldsymbol{I(X;Y,Z)} = I(X;Y) + I(X;Z|Y)$$

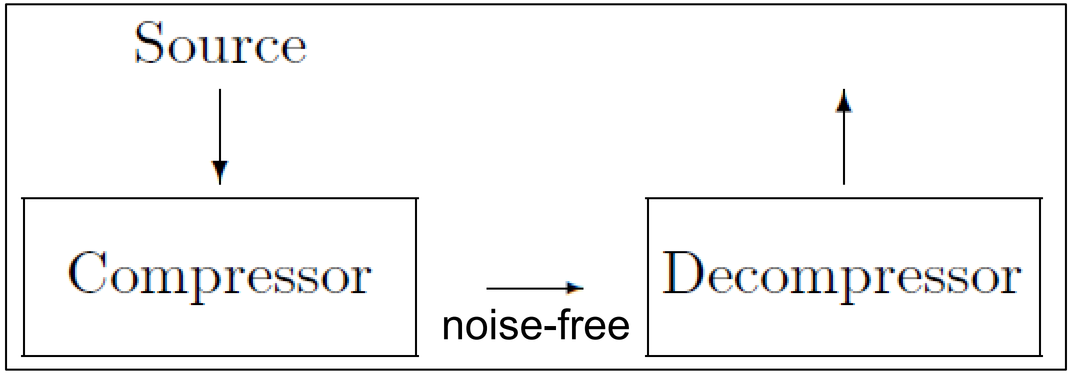
If $X \to Y \to Z$,
by definition,
$I(X;Z|Y) = 0$.

Theorem. If $X \to Y \to Z$, then $I(X;Y) \geq I(X;Z)$.

Theorem. If $X \to Y \to Z$, then $I(X;Y) \geq I(X;Y|Z)$.

Source Coding Theorem

We have n i.i.d. RVs.
What is the min #bits required to send the data only with negligible error?



Upper/Lower bound on Compression

We can construct an instantaneous code given a length function $\ell: \mathcal{X} \to \{0,1\}^*$ if

$$\sum_{x \in \mathcal{X}} 2^{-\ell(x)} \leq 1.$$

Shannon compression $\ell(x) \coloneqq \lceil -\log p(x) \rceil$ gives $\mathbb{E}_{X\sim p}[\ell(X)] < H(X) + 1$.

Theorem. Huffman compression is optimal, i.e., $\mathbb{E}[\ell_{\text{Huffman}}(X)] \leq \mathbb{E}\left[\ell_{\text{Uniquely Decodable Code}}(X)\right]$.

Any uniquely decodable code with a length function $\ell: \mathcal{X} \to \{0,1\}^*$ must satisfies

$$\sum_{x \in \mathcal{X}} 2^{-\ell(x)} \leq 1.$$

Theorem. $\mathbb{E}\left[\ell_{\text{Uniquely Decodable Code}}(X)\right] \geq H(X)$ with equality iff $p(x) = 2^{-\ell(x)}$ for all $x \in \mathcal{X}$.

$$H(X) \leq \mathbb{E}[\ell^*(X)] < H(X) + 1$$

Upper/Lower bound on Compression

Consider a sequence of (possibly dependent) RVs $X_1, X_2, \dots, X_n$ with joint distribution p.
Shannon compression gives $\mathbb{E}_{(X_1,X_2,\dots,X_n)\sim p}[\ell(X_1, X_2, \dots, X_n)] < H(X_1, X_2, \dots, X_n) + 1$.

If $X_1, X_2, \dots, X_n$ are i.i.d., $H(X_1, X_2, \dots, X_n) = H(X)$.
Therefore, the expected length per symbol of an optimal compression is

$$H(X) \leq \frac{1}{n}\mathbb{E}[\ell^*(X_1, X_2, \dots, X_n)] < H(X) + \frac{1}{n}.$$

$\boldsymbol{H(X)}$ is the fundamental limit!

Q. What is the fundamental limit if we allow small error in the compression scheme?

AEP (Asymptotic Equipartition Property)

Consider a sequence of i.i.d. RVs $X_1, X_2, \dots, X_n$.

For any $\epsilon > 0$, for all sufficiently large n,

$$\Pr\left[2^{-n(H(X)+\epsilon)} < p(X_1, X_2, \dots, X_n) < 2^{-n(H(X)-\epsilon)}\right] \geq 1 - \epsilon.$$

The typical set $A_\epsilon^{(n)}$ w.r.t. p is the set of sequence $\mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathcal{X}^n$ such that

$$2^{-n(H(X)+\epsilon)} < p(\mathbf{x}) < 2^{-n(H(X)-\epsilon)}.$$

AEP. For sufficiently large n,

$$\Pr\left[\mathbf{X} \in A_\epsilon^{(n)}\right] \geq 1 - \epsilon \qquad \text{contains most of the probability}$$

$$(1-\epsilon)2^{n(H(X)-\epsilon)} \leq \left|A_\epsilon^{(n)}\right| \leq 2^{n(H(X)+\epsilon)}. \qquad \text{cardinality} \approx 2^{nH(X)}$$

$A_\epsilon^{(n)}$ contains most of the probability and has a cardinality $\approx 2^{nH(X)}$.

Typical Set is a Smallest Set

Let $B_\delta^{(n)} \subseteq \mathcal{X}^n$ be a smallest set with $\Pr\left[\mathbf{X} \in B_\delta^{(n)}\right] \geq 1 - \delta$.

Lemma. $\left|B_\delta^{(n)}\right| \geq (1 - \epsilon - \delta)2^{n(H(X)-\epsilon)} \approx 2^{nH(X)}$.

$\boldsymbol{A_\epsilon^{(n)}}$ **contains most of the probability and *essentially* has a smallest cardinality of $\approx 2^{nH(X)}$.**

Source coding theorem.

Consider a sequence of n i.i.d. RVs with entropy H.

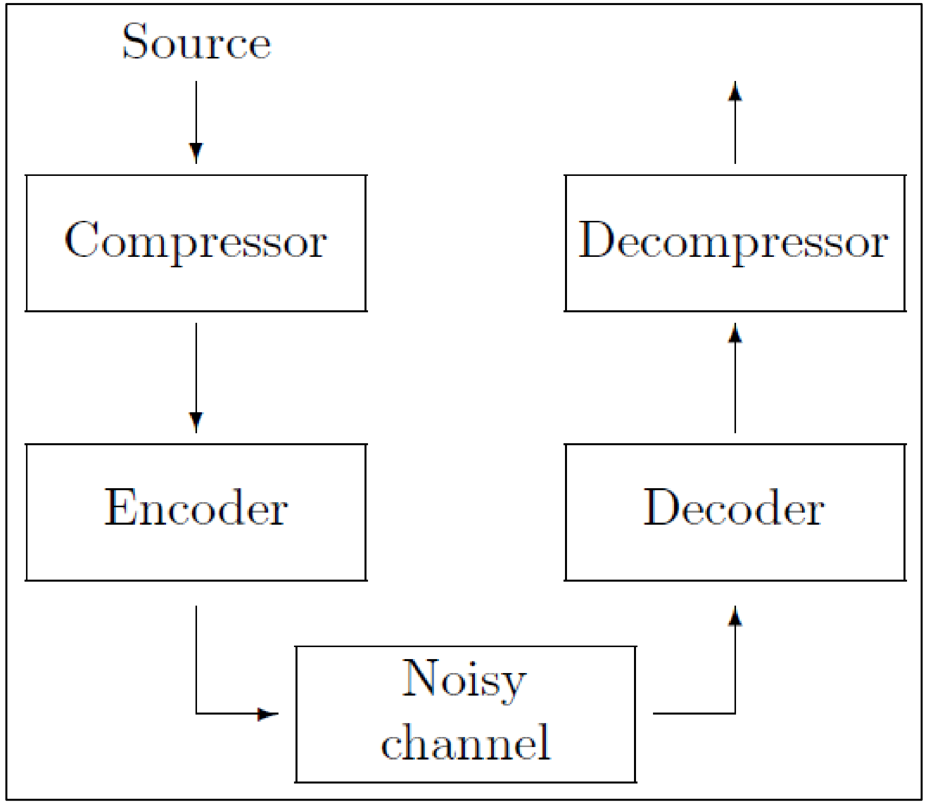
(1-1) Using *slightly more than* nH bits admits a lossless compression.

(1-2) With #bits *very close to* nH, $\Pr[error] \approx 0$ for sufficiently large n

(2) With less than nH bits, $\Pr[error] \approx 1$ for sufficiently large n (i.e., all information is lost).

Channel Coding Theorem

Suppose we use n symbols to encode a date to cope with the channel noise.
What is the max #data we can send only with negligible error?



Motivation

We have a noisy channel $p(y|x)$.

Alice tosses a coin X and send X to Bob (using single bit).

- amount of information before being sent $= H(X)$

Bob receives a bit Y through a noisy channel.

- amount of information that channel decreased $= H(X|Y)$
- amount of information conveyed by the channel $= H(X) - H(X|Y) = I(X, Y)$

Information channel capacity

- Assume for now that we wish to maximize the amount of information conveyed by the channel.
- We do this by choosing a *best* distribution of X.

Channel and Channel Capacity

Discrete memoryless channel Q

$\mathcal{X}$: an input alphabet (a set of input symbols)

$\mathcal{Y}$: an output alphabet (a set of output symbols)

$\{p(y|x)\}_{x\in\mathcal{X}}$: a collection of (conditional) PMFs

Choose a distribution. → Capable of sending $I(X;Y)$ amount of information

Channel capacity (choose a *best* distribution)

$$C = \max_{\text{distribution over } \mathcal{X}} I(X;Y)$$

* maximum is well defined

Code, Decode, Rate, Error

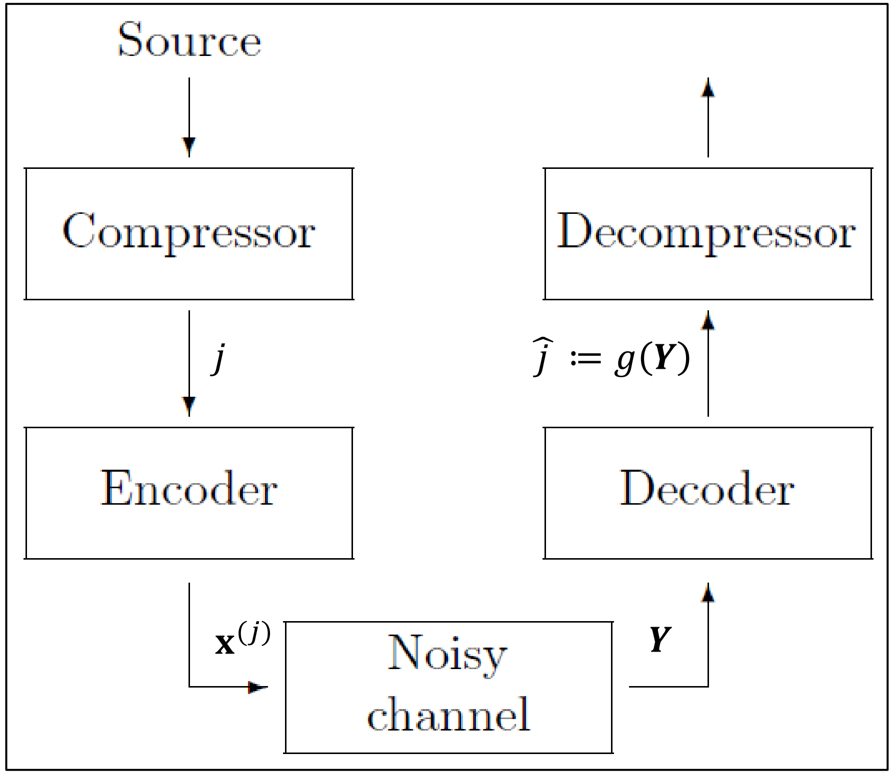
$M, n \in \mathbb{N}$

$(\boldsymbol{M}, \boldsymbol{n})$ **code** for a channel $Q = (\mathcal{X}, p(y|x), \mathcal{Y})$ "encoding scheme"

- a set of indices(data) $\{1, \dots, M\}$
- a set of codewords $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(M)}\}$, where $\mathbf{x}^{(j)} \in \mathcal{X}^n$

A **decoder** g guess an index.

- An optimal decoder g^* chooses a posteriori most likely index.



$\boldsymbol{Y} \sim p(\cdot\,|\mathbf{x}^{(j)}) = \prod_{i=1}^{n} p\left(y_i | x_i^{(j)}\right)$

$(\log M)$-bit

(M, n) code sends an index with n transmissions.

Rate R of (M, n) code for Q

$$R = \frac{size\ of\ data}{\#transmission} = \frac{\log M}{n}$$

Maximal probability of error (for a fixed channel Q and fixed (M, n) code for Q)

$$\lambda_{\max} \coloneqq \max_{j \in \{1, \dots, M\}} \Pr_{\boldsymbol{Y} \sim p(\cdot|\mathbf{x}^{(j)})} [g(\boldsymbol{Y}) \neq j]$$

Achievable Rate

Let us fix n.

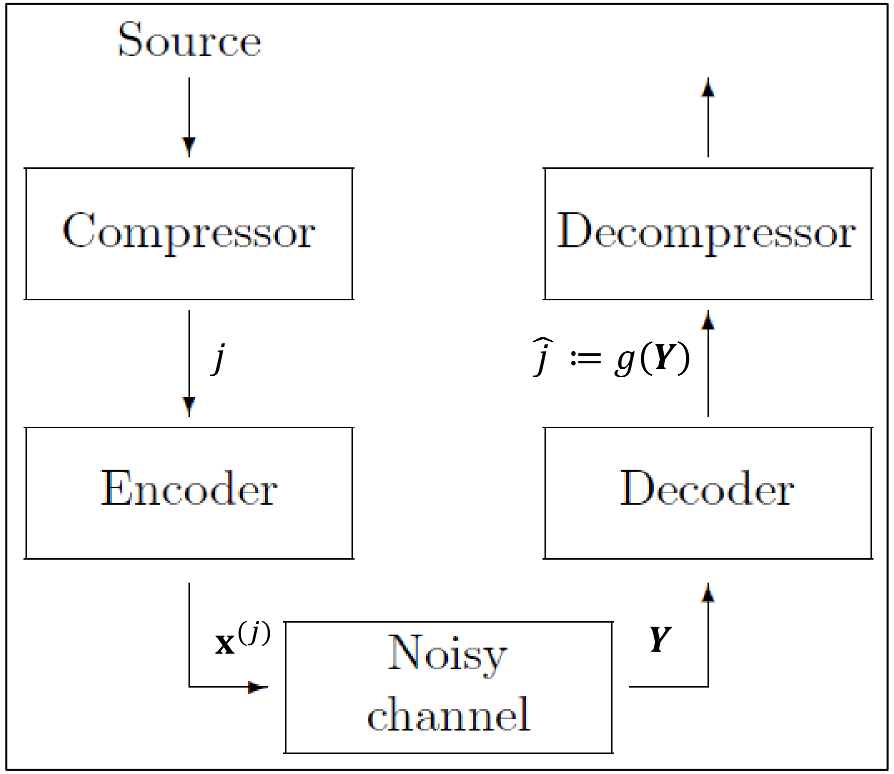
Increase rate = Send more information per transmission

= Cause more error (possibly)

Q. What rate can we prove is *achievable*?

One way to show achievability = show existence of such code

- Show the expected error of a random code is small.
- There must be a code with small error.



$\boldsymbol{Y} \sim p(\cdot \mid \mathbf{x}^{(j)}) = \prod_{i=1}^{n} p\left(y_i \mid x_i^{(j)}\right)$

One (naïve) way to construct a random code

- Fix a distribution p_x over $\mathcal{X}$. Sample i.i.d. n symbols from p_x for each codeword, independently.

A. $R \lesssim I$ is achievable. i.e., sending 2^{nI} is possible

If p_x is a *best* distribution, $R \lesssim C$ is achievable.

Theorem (part 1)

The following holds for any discrete memoryless channel $Q = (\mathcal{X}, p(y|x), \mathcal{Y})$.

For any $\epsilon > 0$ and $R < C \coloneqq \max_{p_x} I(X;Y)$, there exists a $(M \coloneqq [2^{nR}], n)$ code for Q such that $\lambda_{\max} < \epsilon$ for all sufficiently large n.

Showing the existence

- Fix any p_x. Generate a random (M', n) code according to p_x where $M' = \lceil 2^{n(R+1/n)} \rceil$:
  - For each $j \in [M']$, independently, $\boldsymbol{X}^{(j)} = X_1^{(j)} X_2^{(j)} \cdots X_n^{(j)}$ where $X_i^{(j)} \sim p_x$ independently.
- Sample $J \in [M']$ at random. Consider $\boldsymbol{Y} \sim p\left(\cdot \,|\boldsymbol{X}^{(J)}\right)$ and a *jointly typical* decoder g.
  Sample a codeword $\mathbf{X}^{(J)}$ at random
- **Claim.** $\Pr_{\substack{(M,n)\text{ code,} \\ J, \boldsymbol{Y} \sim p(\cdot|\boldsymbol{X}^{(J)})}} [g(\boldsymbol{Y}) \neq J]$ is small. $\rightarrow \exists$ a (M, n) code with small $\Pr_{J, \boldsymbol{Y} \sim p(\cdot|\boldsymbol{X}^{(J)})} [g(\boldsymbol{Y}) \neq J]$.
- Removing the worst half of codewords ensures $\lambda_{\max} = \max_{j \in [M]} \Pr_{\boldsymbol{Y} \sim p(\cdot|\mathbf{x}^{(j)})} [g(\boldsymbol{Y}) \neq j]$ is also small.
  Rate decreases by $1/n$.

Intuitive Idea

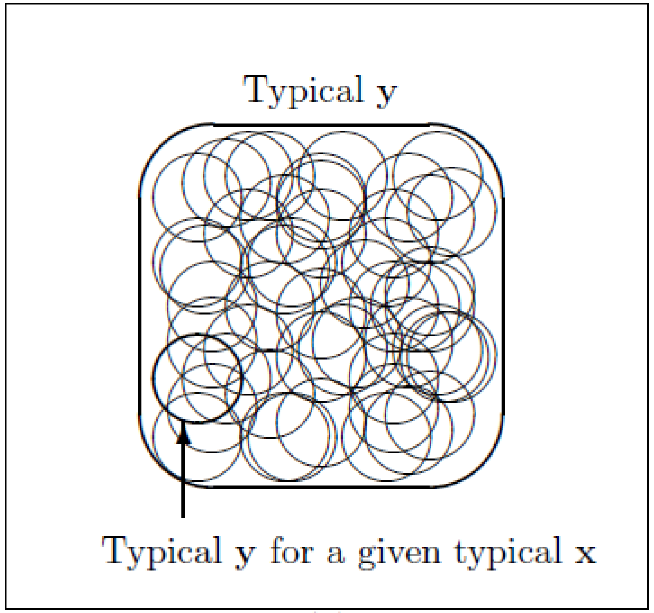
Consider sufficiently large n.

By AEP,

$$\left|\left\{\mathbf{y} \mid p(\mathbf{y}) \approx 2^{-nH(Y)}\right\}\right| \approx 2^{nH(Y)}$$

Similarly, given typical $\mathbf{x}$,

$$\left|\left\{\mathbf{y} \mid p(\mathbf{y}|\mathbf{x}) \approx 2^{-nH(Y|X)}\right\}\right| \approx 2^{nH(Y|X)}$$



Intuitive Idea

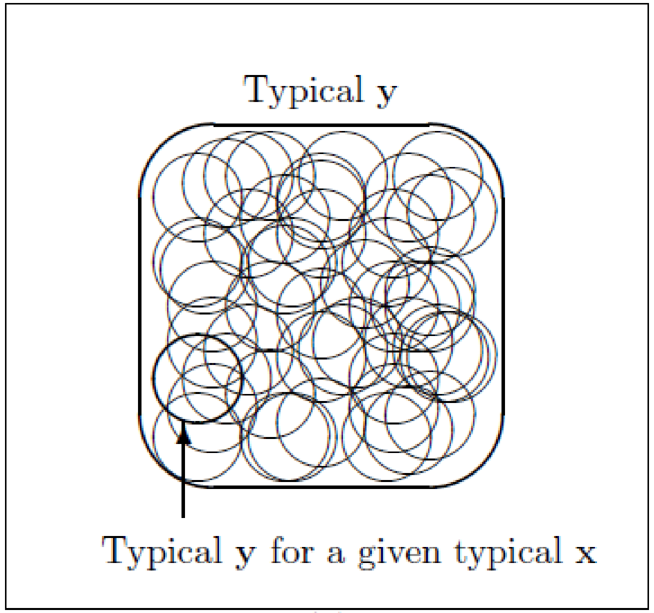
Consider sufficiently large n.

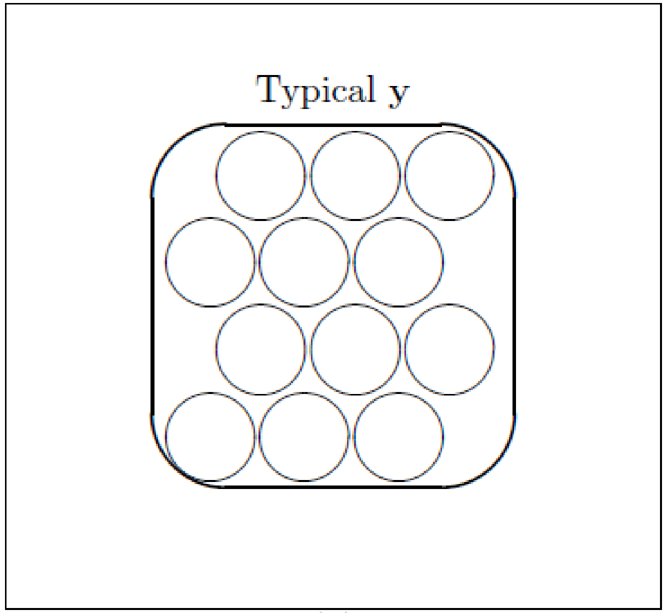
By AEP,

$$\left|\left\{\mathbf{y} \mid p(\mathbf{y}) \approx 2^{-nH(Y)}\right\}\right| \approx 2^{nH(Y)}$$

Similarly, given typical $\mathbf{x}$,

$$\left|\left\{\mathbf{y} \mid p(\mathbf{y}|\mathbf{x}) \approx 2^{-nH(Y|X)}\right\}\right| \approx 2^{nH(Y|X)}$$





Ideally, we "pack" $2^{nH(Y)}/2^{nH(Y|X)} = 2^{nI(X;Y)}$ number of non-confusable typical $\mathbf{y}$ for a given typical $\mathbf{x}$.

Joint AEP **Independently sampling (about) $2^{nI(X;Y)}$ #codewords suffices for non-confusability.**

Jointly-typical decoder g (which is sub-optimal compared to g^*)

$g(\mathbf{y}) = j$ if $\left(\mathbf{x}^{(j)}, \mathbf{y}\right)$ is *jointly typical* and no other j' exists such that $\left(\mathbf{x}^{(j')}, \mathbf{y}\right)$ is *jointly typical*.

Otherwise, outputs arbitrary index.

Joint AEP and Error Bound

Jointly typical set $A_\epsilon^{(n)} = \{ (\mathbf{x}, \mathbf{y}) \mid p(\mathbf{x}) \approx 2^{-nH(X)},\ p(\mathbf{y}) \approx 2^{-nH(Y)},\ p(\mathbf{x}, \mathbf{y}) \approx 2^{-nH(X,Y)} \}$

Joint AEP. For sufficiently large n,

$$\Pr_{(\boldsymbol{X},\boldsymbol{Y})\sim p(\mathbf{x},\mathbf{y})}\left[(\boldsymbol{X},\boldsymbol{Y}) \in A_\epsilon^{(n)}\right] \geq 1-\epsilon$$

contains most of the probability

$$\left|A_\epsilon^{(n)}\right| \leq 2^{n(H(X,Y)+\epsilon)}$$

cardinality $\lesssim 2^{nH(X)}$

$$(1-\epsilon)2^{-n(I(X;Y)+3\epsilon)} \leq \Pr_{\substack{\boldsymbol{X}\sim p(\mathbf{x})\\ \boldsymbol{Y}\sim p(\mathbf{y})}}\left[(\boldsymbol{X},\boldsymbol{Y}) \in A_\epsilon^{(n)}\right] \leq 2^{-n(I(X;Y)-3\epsilon)}$$

Independent $\boldsymbol{X}, \boldsymbol{Y}$ being jointly typical is *exponentially* small

Error Bound. Fix any j.

- $\Pr[(\boldsymbol{X}^{(j)}, \boldsymbol{Y})$ is not jointly typical$] < \epsilon$
- For fixed $j' \neq j$, $\Pr[(\boldsymbol{X}^{(j')}, \boldsymbol{Y})$ is jointly typical$] \lesssim 2^{-nI(X;Y)}$ ($\boldsymbol{X}^{(j')}$ and $\boldsymbol{Y}$ are independent)

$\rightarrow \Pr[\exists\, j' \neq j : (\boldsymbol{X}^{(j')}, \boldsymbol{Y})$ is jointly typical$] \lesssim (M'-1)2^{-nI(X;Y)} \lesssim \epsilon$

Union bound — Holds if $R \lesssim I(X;Y)$

$$\Pr_{\substack{(M,n)\text{ code},\\ J,\boldsymbol{Y}\sim p(\cdot|\boldsymbol{X}^{(J)})}}[g(\boldsymbol{Y}) \neq J] \lesssim 2\epsilon$$

Theorem (part 2)

The following holds for any discrete memoryless channel $Q = (\mathcal{X}, p(y|x), \mathcal{Y})$.

Any $([2^{nR}], n)$ code with $R > C \coloneqq \max_{p_x} I(X;Y)$ has λ_{avg} bounded away from 0 for all n.

Proof sketch)

- For fixed encoder and decoder, we have $J \to \boldsymbol{X}^{(J)} \to \boldsymbol{Y} \to g(\boldsymbol{Y})$.

$$H(J) = H(J|g(\boldsymbol{Y})) + I(J; g(\boldsymbol{Y}))$$

$$\leq H(J|g(\boldsymbol{Y})) + I\left(\boldsymbol{X}^{(J)}; \boldsymbol{Y}\right) \quad \text{Data processing}$$

$$\leq 1 + \Pr[J \neq g(\boldsymbol{Y})] \log(M-1) + I\left(\boldsymbol{X}^{(J)}; \boldsymbol{Y}\right) \quad \text{Fano's inequality}$$

$$\leq 1 + \Pr[J \neq g(\boldsymbol{Y})]\, nR + nC$$

Doing n transmissions; $I(X;Y)$ per transmission.

Assuming J is sampled from a uniform distribution, $H(J) = \log[2^{nR}] \approx nR$ and thus

$$\Pr[J \neq g(\boldsymbol{Y})] \geq 1 - \frac{C}{R} - \frac{1}{nR}.$$

Stronger Theorem (part 2)

The following holds for any discrete memoryless channel $Q = (\mathcal{X}, p(y|x), \mathcal{Y})$.

Any $([2^{nR}], n)$ code with $R > C \coloneqq \max_{p_x} I(X;Y)$ has $\lambda_{\text{avg}} \approx 1$ for all sufficiently large n.

Rate Distortion Theory

We have n i.i.d. RVs.

Source coding theorem says

with a distortion function $d(\mathbf{x}, \hat{\mathbf{x}}) \coloneqq \mathbb{I}_{\mathbf{x} \neq \hat{\mathbf{x}}}$,

- if $R < H$, then $\mathbb{E}\left[d\left(\boldsymbol{X}, \widehat{\boldsymbol{X}}\right)\right] \leq 1 - \epsilon$ is not possible (when n is large);
- if $R > H$, then $\mathbb{E}\left[d\left(\boldsymbol{X}, \widehat{\boldsymbol{X}}\right)\right] \leq \epsilon$ is possible (when n is large).

Rate distortion theory says

with a separable distortion function $d(\cdot,\cdot)$, i.e., $d(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{n}\sum_{i=1}^{n} d(x_i, \widehat{x_i})$,

- if $R <$?, then $\mathbb{E}\left[d\left(\boldsymbol{X}, \widehat{\boldsymbol{X}}\right)\right] \leq D$ is not possible;
- if $R >$?, then $\mathbb{E}\left[d\left(\boldsymbol{X}, \widehat{\boldsymbol{X}}\right)\right] \leq D$ is possible.

Rate and Distortion

Consider a (per symbol) distortion function $d(x, \hat{x})$, e.g., $d(x, \hat{x}) = \mathbb{I}_{\mathrm{x} \neq \hat{\mathrm{x}}}$. and $d(x, \hat{x}) < \infty$.

A compression scheme $(f, g)_{n,R}$ compresses n symbols with R bits per symbol. i.e., $nR (\in \mathbb{N})$ bits in total.

- Codebook (for decoding): $\left\{j, \hat{\mathbf{x}}^{(j)}\right\}_{j=1}^{2^{nR}}$

Distortion of $(f, g)_{n,R}$= $\mathbb{E}_{\boldsymbol{X}}\left[d\left(\boldsymbol{X}, g\big(f(\boldsymbol{X})\big)\right)\right]$.

(R, D) is **achievable** if for any $\delta > 0$,

there exist a scheme $(f, g)_{n,R}$ that compresses n i.i.d. symbols with distortion $\leq D + \delta$.

Rate distortion function $R(D) = \min_{(R,D) \text{ achievable}} R$

Assume, given a distribution $p(x)$ over $\mathcal{X}$, we wish to find a "test channel" $(\mathcal{X}, p(\hat{x}|x), \mathcal{X})$ with distortion at most D such that it conveys a minimum amount of information.

$$R^{(I)}(D) \coloneqq \min_{p(\hat{x}|x) : \underset{\substack{X \sim p(x) \\ \hat{X} \sim p(\cdot|X)}}{\mathbb{E}} [d(X, \hat{X})] \leq D} I\big(X; \hat{X}\big)$$

Rate Distortion Theory

Theorem. $R(D) = R^{(I)}(D)$.

proof sketch)

- Any $(f, g)_{n,R}$ with distortion $\leq D$ must satisfy $R \geq R^{(I)}(D)$.
  - $R^{(I)}(D)$ is nonincreasing and convex in D.
  - d is separable.
  - Data inequalities.
- $\left(R^{(I)}(D), D\right)$ is achievable.
  - Ideally, construct a codebook $\left\{j, \hat{\mathbf{x}}^{(j)}\right\}_{j=1}^{2^{nR^{(I)}(D)}}$ that "covers" all typical $\mathbf{x}$ within distortion D.
  - (Joint) Distortion AEP

Independently sampling 2^{nR} number of typical $\hat{\mathbf{x}}$ suffices for covering all typical $\mathbf{x}$.

Achievability

One way to show achievability = show existence of such scheme

- Construct a random scheme $(f,g)_{n,R}$ where $R > R^{(I)}(D)$. Show the expected distortion $\leq D$.

One way to construct a random scheme

- Fix a distribution $p(\hat{x}|x)$ with distortion $= D \rightarrow p(\hat{x})$ is fixed.
  Fix a test channel.
- For each $j \in [2^{nR}]$, independently, $\boldsymbol{X}^{(j)} = X_1^{(j)} X_2^{(j)} \cdots X_n^{(j)}$ where $X_i^{(j)} \sim p(\hat{x})$ independently.
  - Let $g(j) = \hat{\mathbf{x}}^{(j)}$ for each $j \in [2^{nR}]$.
- Consider a (*jointly*) *distortion typical* f.
  - $f(\mathbf{x}) =$ any j such that $(\mathbf{x}, \hat{\mathbf{x}}^{(j)})$ is distortion typical.
  - If there is no j such that $(\mathbf{x}, \hat{\mathbf{x}}^{(j)})$ is distortion typical, outputs arbitrary j.

Claim. Random scheme has distortion $\leq D + \delta$. $\rightarrow$ There is a scheme with distortion $\leq D + \delta$.

(Joint) Distortion AEP

(Jointly) Distortion typical set

$$A_{d(\cdot,\cdot),\epsilon}^{(n)} = \left\{ (\mathbf{x}, \mathbf{y}) \mid p(\mathbf{x}) \approx 2^{-nH(X)},\ p(\mathbf{y}) \approx 2^{-nH(Y)},\ p(\mathbf{x}, \mathbf{y}) \approx 2^{-nH(X,Y)}, d(\mathbf{x}, \mathbf{y}) \approx \mathbb{E}[d(\boldsymbol{X}, \boldsymbol{Y})] \right\}$$

Since $d(\mathbf{x}, \mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} d(x_i, y_i)$ is separable, the law of large number can be applied.

Joint (Distortion) AEP. For sufficiently large n,

$$\Pr_{(\boldsymbol{X},\boldsymbol{Y})\sim p(\mathbf{x},\mathbf{y})}\left[(\boldsymbol{X}, \boldsymbol{Y}) \in A_{d,\epsilon}^{(n)}\right] \geq 1 - \epsilon$$

contains most of the probability

Distortion Bound.

- $\boldsymbol{X}$ is distortion typical with some $\widehat{\boldsymbol{X}}$ → Contribution to distortion $\approx D$
- $\Pr[\boldsymbol{X}$ is not distortion typical with any $\widehat{\boldsymbol{X}}] \cdot \max d(\mathbf{x}, \hat{\mathbf{x}})$ → Contribution to distortion ≈ 0

This probability goes to zero *exponentially* fast if $R > I(X; \hat{X})$.

Since $d(x, \hat{x})$ is bounded.

* Stronger sense of typicality upper-bounds the distortion w.h.p., i.e., $\Pr\left[d\left(\mathbf{X}, \widehat{\mathbf{X}}\right) > D + \delta\right] \approx 0$

Characterization of $R(D)$

We showed $R(D) = \min_{\substack{p(\hat{x}|x)\,:\, \mathbb{E}_{\substack{X\sim p(x)\\ \hat{X}\sim p(\hat{x}|X)}}[d(X,\hat{X})]\leq D}} I(X;\hat{X})$.

Solve the minimization problem of a convex function over the convex set of some distributions.

$\rightarrow$ We obtain an optimal $p(\hat{x}|x)$.

Blahut-Arimoto algorithm computes two alternating minimization iteratively.

- Also converges to an optimal $p(\hat{x}|x)$.

Channel Coding Theorem (part 3)

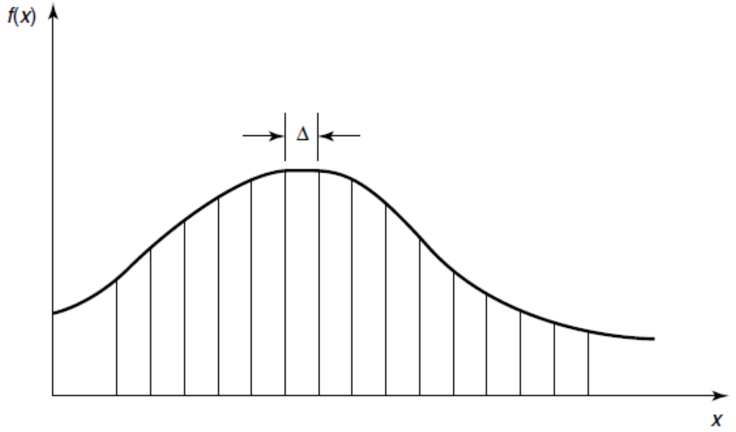
When given a discrete memoryless channel with a bounded separable distortion function,

the distortion D is achievable iff $C > R(D)$.

Differential Entropy

Continuous Random variable

Differential Entropy



Consider a continuous random variable X with density f.

Divide range of X into bins of length Δ.

Let x_i be a value such that $f(x_i) \cdot \Delta = \int_{i\Delta}^{(i+1)\Delta} f(x)dx$.

Consider a discrete random variable X^Δ where $X^\Delta = x_i$ with probability $f(x_i)\Delta$.

$$H(X^\Delta) = -\sum_{i=-\infty}^{\infty} f(x_i)\Delta \log(f(x_i)\Delta) = -\sum_{i=-\infty}^{\infty} f(x_i)\Delta \log f(x_i) - \log \Delta$$

As $\Delta \to 0$, , $H(X^\Delta) + \log \Delta \to -\int_{-\infty}^{\infty} f(x) \log f(x)\, dx$. if $f(x) \log f(x)$ is Riemann integrable

Differential Entropy

$$h(X) = h(f) = -\int_{-\infty}^{\infty} f(x) \log f(x)\, dx$$

Properties of Differential Entropy

- h ≠ amount of information (or entropy of quantized continuous RV)
  - $h(f) < 0$ is possible. Consider a density function f that corresponds to $U[0,1/4]$.

$$h(f) = -\int_0^{1/4} 4 \log 4 \, dx = -2$$

- Translation does not change h. $h(X) = h(X + c)$.
- Scaling does change h. $h(aX) = h(X) + \log|a|$
- Differential entropy $h(X)$ of gaussian RV $X \sim \mathrm{N}(0, \sigma^2)$ = $\frac{1}{2}\log(2\pi\sigma^2 e)$

$$X \sim f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$$

$$h(f) = -\int_{-\infty}^{\infty} f(x) \log f(x) \, dx = -\int_{-\infty}^{\infty} f(x) \left(\log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{x^2}{2\sigma^2} \log e \right) dx$$

$$= \frac{1}{2}\log(2\pi\sigma^2) + \frac{\mathbb{E}_{X\sim N(0,\sigma^2)}[X^2]}{2\sigma^2} \log e = \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2}\log e = \frac{1}{2}\log(2\pi\sigma^2 e)$$

KL Divergence and Mutual Information

Kullback-Leibler divergence between PDFs f and g

$$D(f \parallel g) = \int\int f(x_1)\log\frac{f(x_1)}{g(x_2)}dx_1\,dx_2$$

Mutual information between X and Y with joint density $f(x,y)$

$$\begin{aligned} I(X;Y) &= \int\int f(x,y)\log\frac{f(x,y)}{f(x)f(y)}dx\,dy \\ &= D\big(f(x,y) \parallel f(x)f(y)\big) \\ &= h(X) - h(X|Y) = h(Y) - h(Y|X) \\ &\approx I\big(X^{\Delta};Y^{\Delta}\big) \end{aligned}$$

Theorem. $D(f \parallel g) \geq 0$ with = iff $f = g$.

Corollary. $I(X;Y) \geq 0$ with = iff X and Y are independent.

Corollary. $h(X|Y) \leq h(X)$ with = iff X and Y are independent.

Corollary. If X be a RV with support $[-a,a]$, $h(X) \leq h(\mathrm{U}[-a,a])$ with equality iff $X\sim\mathrm{U}[-a,a]$.

Corollary. If X be a RV with a variance σ^2, $h(X) \leq h(\mathrm{N}(0,\sigma^2))$ with equality iff $X\sim\mathrm{N}(0,\sigma^2)$.

AEP

Consider a sequence of i.i.d. RVs $X_1, X_2, \dots, X_n$.

For any $\epsilon > 0$, for all sufficiently large n,

$$\Pr\left[2^{-n(h(X)+\epsilon)} < f(X_1, X_2, \dots, X_n) < 2^{-n(h(X)-\epsilon)}\right] \geq 1 - \epsilon.$$

The typical set $A_\epsilon^{(n)}$ w.r.t. f is the set of sequence $\mathbf{x} = (x_1, x_2, \dots, x_n) \in S^n$ such that

S: support of f

$$2^{-n(h(X)+\epsilon)} < f(\mathbf{x}) < 2^{-n(h(X)-\epsilon)}.$$

AEP. For sufficiently large n,

$$\Pr\left[\mathbf{X} \in A_\epsilon^{(n)}\right] \geq 1 - \epsilon \quad \text{and} \quad (1-\epsilon)2^{n(h(X)-\epsilon)} \leq \mathrm{Vol}\left(A_\epsilon^{(n)}\right) \leq 2^{n(h(X)+\epsilon)}.$$

contains most of the probability — volume $\approx 2^{nh(X)}$

Let $B_\delta^{(n)} \subseteq S^n$ be a smallest set with $\Pr\left[\mathbf{X} \in B_\delta^{(n)}\right] \geq 1 - \delta$.

Lemma. $\mathrm{Vol}\left(B_\delta^{(n)}\right) \geq (1-\epsilon-\delta)2^{n(h(X)-\epsilon)} \approx 2^{nh(X)}$.

$A_\epsilon^{(n)}$ contains most of the probability and *essentially* has a smallest volume of $\approx 2^{nh(X)}$.

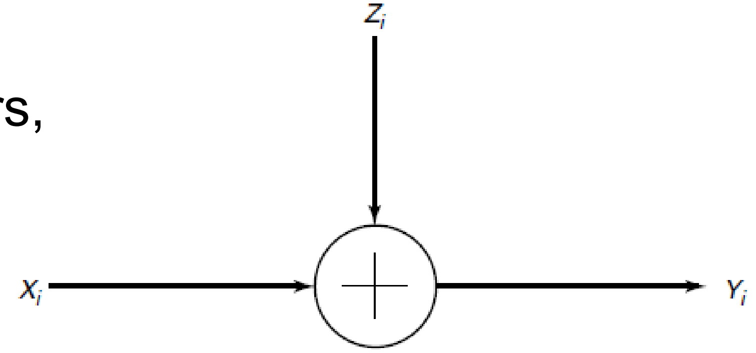
Gaussian Channel

Previously, we considered discrete-time discrete-space input channel.

Gaussian channel is a discrete-time continuous-space input channel.

We also consider a continuous-time continuous-space input channel with bandlimit.

(Discrete-time) Gaussian Channel



In a Gaussian channel, the input space is continuous, e.g., real numbers, and Gaussian noise is added to the input.

Note that it is a discrete-time channel.

- If no constraint on the input, it has infinite capacity.
    - Even if the noise variance > 0, it can transmit infinitely many numbers almost perfectly.
- Common constraint: upper bound on the average power of the input $(x_1, x_2, \ldots, x_n)$

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 \leq P$$

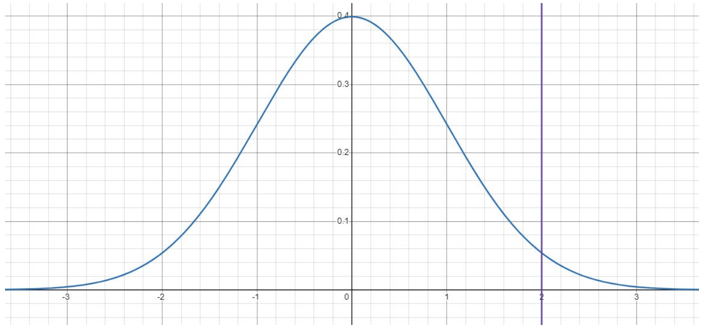
Example

There is a Gaussian channel Q with input-space as $\mathbb{R}$, a Gaussian noise $\mathrm{N}(0,1)$ and power constraint $P = 4$.

Alice want to send the result of a random coin with $x \in \mathbb{R}$ (i.e., length 1 codeword) through Q.
Alice encode the result of the coin to a codeword X where $X = 2$ if HEAD, $X = -2$, otherwise.
Bob receives a noisy number $Y(= X + Z)$. Bob decodes Y to HEAD if $Y > 0$, TAIL, otherwise.

Error-probability

$$\begin{aligned}\Pr[error] &= \Pr[Y \leq 0, X = 2] + \Pr[Y > 0, X = -2] \\ &= \Pr[Z \leq -2, X = 2] + \Pr[Z > 2, X = -2] \\ &= \Pr[Z > 2] \approx 0.022\end{aligned}$$



Channel Capacity

The information capacity of the Gaussian channel $\mathrm{N}(0, \sigma^2)$ with power constraint P

$$C = \max_{f(x): \mathbb{E}_{X \sim f}[X^2] \leq P} I(X; Y)$$

We have $I(X; Y) = h(Y) - h(Y \mid X) = h(Y) - h(Z \mid X) = h(Y) - h(Z)$.

Since $Z \sim \mathrm{N}(0, \sigma^2)$, we have $h(Z) = \frac{1}{2} \log(2\pi\sigma^2 e)$

Moreover, the variance of Y is $\mathbb{E}[Y^2] = \mathbb{E}[X^2 + 2XZ + Z^2] = P + \sigma^2$.

$$h(Y) \leq h\left(\mathrm{N}(0, P + \sigma^2)\right) = \frac{1}{2} \log(2\pi(P + \sigma^2)e)$$

Therefore,

$$C = \frac{1}{2} \log\left(1 + \frac{P}{\sigma^2}\right)$$

Code, Decode, Rate, Error

$(\boldsymbol{M}, \boldsymbol{n})$ **code** for a channel $\mathrm{N}(0, \sigma^2)$ with power constraint P

- a set of indices(data) $\{1, \dots, M\}$
- a set of codewords $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(M)}\}$
  - $\mathbf{x}^{(j)} = x_1^{(j)} x_1^{(j)} \cdots x_n^{(j)}$ such that ${x_1^{(j)}}^2 + {x_2^{(j)}}^2 + \cdots + {x_n^{(j)}}^2 \leq nP$

A **decoder** g guess an index among $[M]$.

Rate R of (M, n) code for Q

$$R = \frac{size\ of\ data}{\#transmission} = \frac{\log M}{n}$$

Maximal probability of error (for a fixed channel Q and fixed (M, n) code for Q)

$$\lambda_{\max} \coloneqq \max_{j \in \{1, \dots, M\}} \Pr_{\boldsymbol{Z} \sim \mathrm{N}^n(0, \sigma^2)} \left[g\left(\boldsymbol{Z} + \mathbf{x}^{(j)}\right) \neq j \right]$$

Theorem

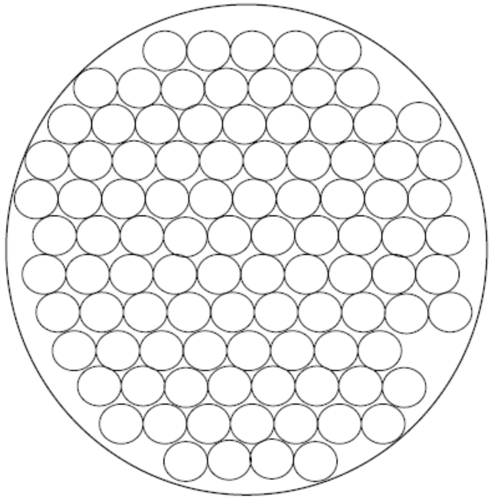
The following holds for any Gaussian channel Q with $\mathrm{N}(0, \sigma^2)$ with power constraint P.

1) For any $\epsilon > 0$ and $R < C$, there exists a $(M := \lceil 2^{nR} \rceil, n)$ code for Q such that $\lambda_{\max} < \epsilon$ for all sufficiently large n.
2) Any $(\lceil 2^{nR} \rceil, n)$ code with $R > C$ has λ_{avg} bounded away from 0 for all n.

For part 1, it uses Joint AEP.

Ideally, we “pack” non-confusable typical $\mathbf{y}$ for a given typical $\mathbf{x}$.

Joint AEP says independent sampling suffices to achieve this.



For part 2, it utilizes data processing and Fano’s inequality.

Continuous-time Gaussian Channel and Bandlimit

Previously, we considered discrete-time channels. (n usage of given channel)

Now, consider a continuous-time Gaussian channel.

- an input *signal* $x(t)$
- *additive white Gaussian noise* $Z(t)$
- power constraint P (defined in a continuous manner)

Consider a **bandlimited** (continuous-time) Gaussian channel

- Channel cuts out all frequencies greater than W (e.g., by applying a bandpass filter)

Nyquist-Shannon's Theorem

Sampling a signal that is bandlimited to W at a sampling rate $\frac{1}{2W}$ is sufficient for the reconstruction.

Discretize input signal. → Send through discrete-time channel for multiple times.

Capacity of Bandlimited Gaussian Channel

Consider a continuous-time Gaussian channel Q with

- bandwidth W Hz, power P, and *power spectral density* of noise $N_0/2$ W/Hz.

By Nyquist-Shannon's theorem, it is equivalent to $2W$ usage (per sec) of a discrete-time Gaussian channel Q' with power constraint $P/2W$, and noise $\mathrm{N}(0, N_0/2)$ of

Note the capacity of Q' is $C' = \frac{1}{2}\log\left(1 + \frac{P}{WN_0}\right)$.

Then the capacity of Q is

$$C = 2W \cdot C' = W \log\left(1 + \frac{P}{WN_0}\right)$$

Statistics

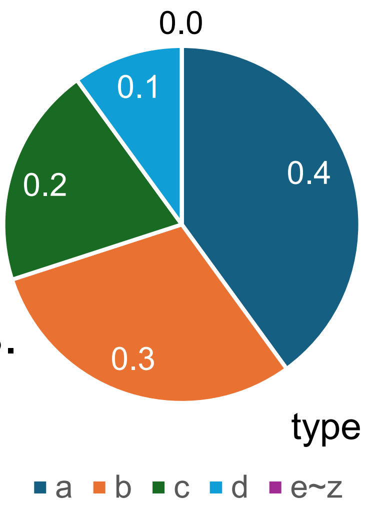
Type gives a stronger sense of AEP.

Type

$\text{type}(\mathbf{x})$: **type** of a sequence $\mathbf{x}$, i.e., the frequency of each symbol in $\mathcal{X}$

$\text{typeclass}(t)$: **type class of type t**, i.e., a set of sequence (of length n) whose type is t

type of *acbdaabcba*



type is a distribution over $\mathcal{X}$

Observation.

Exponential #sequence of length n ($= |\mathcal{X}|^n$).

Polynomial #types of a sequence of length n ($\leq (n+1)^{|\mathcal{X}|}$).

$\rightarrow \exists$ a type whose type class contains *exponentially* many sequences.

In fact, every type class contains

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{nH(t)} \leq |\text{typeclass}(t)| \leq 2^{nH(t)}$$

type class of the above type

aaaabbbccd
aaaabbbcdc
…
dccbbbaaaa

"AEP" for Type and Universal Compression

Consider a sequence of i.i.d. RVs $\boldsymbol{X} = X_1 X_2 \cdots X_n$ where $X_i \sim p$.

Note that $\mathrm{type}(\boldsymbol{X})$ is a random distribution over $\mathcal{X}$.

Let typical sequence be a sequence $\mathbf{x} \in \mathcal{X}^n$ such that $D(\mathrm{type}(\mathbf{x}) \parallel p) \approx 0$.

"AEP" (for type). If n is sufficiently large, $\Pr[D(\mathrm{type}(\boldsymbol{X}) \parallel p) \approx 0] \approx 1$.

For almost every sequence, the sample frequencies are close to the true probability.

Corollary (Universal Codes).

Even if p is unknown, we can compress an i.i.d. source with (very close to) $H(p)$-bit per symbol.

Sanov's Theorem

Consider a sequence of i.i.d. dices $\boldsymbol{X} = X_1 X_2 \cdots X_n$ where $X_i \sim p$.

Q. What is the probability being $\sum_i X_i \geq 4n$?

A1. Central limit theorem, i.e., the distribution of the sample mean → a normal distribution.

- Poor approximation...

A2. Let $\mathcal{T}$ be the set of types of sequence whose sum is at least $4n$, i.e.,

$$\mathcal{T} = \{\mathrm{type}(\mathbf{x}) \mid \mathrm{sum}(\mathbf{x}) \geq 4n\}.$$

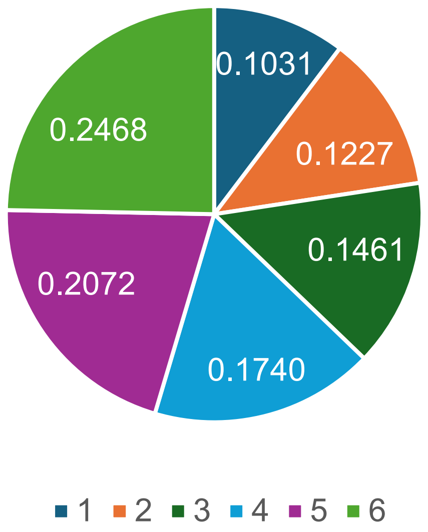
Sanov's theorem says

$$\Pr[\mathrm{type}(\boldsymbol{X}) \in \mathcal{T}] \leq |\mathcal{T}| \cdot 2^{-nD(t^* \| p)} \leq (n+1)^{|\mathcal{X}|} \cdot 2^{-nD(t^* \| p)},$$

where $t^* \in \mathrm{argmin}_{t \in \mathcal{T}} D(t \parallel p)$.

t^*

when p is uniform



The probability measure of $\mathcal{T}$ is essentially determined by t^*.

= probability of large deviation

Conditional Limit Theorem

Consider a sequence of i.i.d. dices $\boldsymbol{X} = X_1 X_2 \cdots X_n$ where $X_i \sim p$.

Q. Suppose $\sum_i X_i \geq 4n$. What can we say about the marginal probability distribution?

A. Let $\mathcal{T}$ be the set of types of sequence whose sum is at least $4n$, i.e.,

$$\mathcal{T} = \{\mathrm{type}(\mathbf{x}) \mid \mathrm{sum}(\mathbf{x}) \geq 4n\},$$

and let $t^* \in \underset{t \in \mathcal{T}}{\operatorname{argmin}} D(t \parallel p)$ be a "closest" distribution in $\mathcal{T}$ to p.

Conditional limit theorem says, for any $x \in \mathcal{X}$, if $\mathcal{T}$ is a closed convex set.

$$\Pr[X_1 = x \mid \mathrm{type}(\boldsymbol{X}) \in \mathcal{T}] \to t^*(x).$$

The probability measure of $\mathcal{T}$ is not only determined by $\boldsymbol{t}^*$

but also concentrated near $\boldsymbol{t}^*$ i.e., the *conditional type* is close to $\boldsymbol{t}^*$.

Another Proof of Joint AEP

product of marginal distributions of p

Consider a sequence of pairs of i.i.d. RVs $(\boldsymbol{X}, \boldsymbol{Y}) = (X_1 Y_1)(X_2 Y_2) \cdots (X_n Y_n)$ where $(X_i Y_i) \sim p_x p_y$

Recall a jointly typical set $A_\epsilon^{(n)} = \{ (\mathbf{x}, \mathbf{y}) \mid p_x(\mathbf{x}) \approx 2^{-nH(X)},\ p_y(\mathbf{y}) \approx 2^{-nH(Y)},\ p(\mathbf{x}, \mathbf{y}) \approx 2^{-nH(X,Y)} \}$.

Let $\mathcal{T}$ be the set of types of typical sequence, i.e., $\mathcal{T} = \left\{ \text{type}(\mathbf{x}, \mathbf{y}) \mid (\mathbf{x}, \mathbf{y}) \in A_\epsilon^{(n)} \right\}$.

By applying Sanov's theorem,

$$\Pr_{(\boldsymbol{X}, \boldsymbol{Y}) \sim \prod p_x p_y} [\text{type}(\boldsymbol{X}, \boldsymbol{Y}) \in \mathcal{T}] \le (n+1)^{|\mathcal{X}|} \cdot 2^{-nD\left(t^* \| p_x p_y\right)},$$

where $t^* \in \underset{t \in \mathcal{T}}{\operatorname{argmin}}\, D\left(t \parallel p_x p_y\right)$.

In fact, $t^* = p$ and thus $\Pr_{(\boldsymbol{X}, \boldsymbol{Y}) \sim \prod p_x p_y} [\text{type}(\boldsymbol{X}, \boldsymbol{Y}) \in \mathcal{T}] \le 2^{-n(I(X;Y)+\epsilon)}$. when n is sufficiently large.

Also, by conditional limit theorem, its conditional type is likely to be close to p. when n is sufficiently large.

Two Hypothesis Testing

Trade-off between $\alpha \coloneqq \Pr[\text{choosed } P_2 \text{ but } P_1 \text{ is true}]$ and $\beta \coloneqq \Pr[\text{choosed } P_1 \text{ but } P_2 \text{ is true}]$.

Neyman-Pearson lemma says the optimum test is a (log-)likelihood ratio test.

$$\frac{P_1(\boldsymbol{X})}{P_2(\boldsymbol{X})} \geq \tau \iff D(\text{type}(\boldsymbol{X}) \parallel P_2) - D(\text{type}(\boldsymbol{X}) \parallel P_1) \geq \frac{1}{n}\log\tau$$

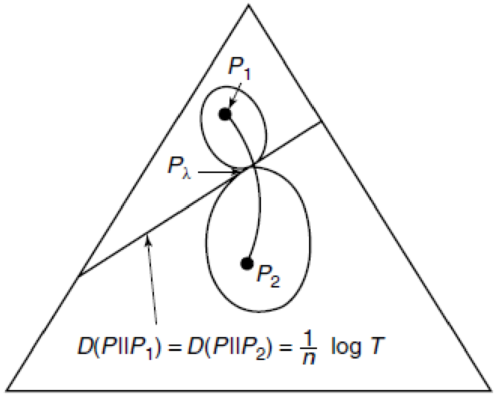
Let $\mathcal{T}$ be the set of types that satisfies the above, i.e., $t \in \mathcal{T}$ accepts P_1 and $t \notin \mathcal{T}$ accepts P_2.

Suppose P_1 was the true distribution, i.e., $\boldsymbol{X} \sim \prod P_1$. Sanov's theorem gives

$$\alpha = \Pr[\text{type}(\boldsymbol{X}) \notin \mathcal{T}] = 2^{-n(D(t_1^* \| P_1) - \epsilon_1)}.$$

Similarly, if P_2 was the true distribution, we have

$$\beta = \Pr[\text{type}(\boldsymbol{X}) \in \mathcal{T}] = 2^{-n(D(t_2^* \| P_2) - \epsilon_2)}.$$



We can also show $t_1^* = \underset{t \notin \mathcal{T}}{\operatorname{argmin}} D(t \parallel P_1) = \underset{t \in \mathcal{T}}{\operatorname{argmin}} D(t \parallel P_2) = t_2^*$.

Chernoff-Stein lemma says if $\alpha < \epsilon$ is small, $\beta \approx 2^{-n(D(P_1 \| P_2) - \epsilon)}$ is best possible. by AEP for KL-divergence.

Kolmogorov Complexity

Source coding says $\approx H(X)$ bits are required to describe X.

What is the shortest length of a program that describes (or outputs) X?

“Approximately equal to its entropy”

Kolmogorov Complexity

Consider any universal Turing machine $\mathcal{U}$.

$K_{\mathcal{U}}(\mathbf{x}) \coloneqq \min_{p:\mathcal{U}(p)=\mathbf{x}} \ell(p)$ is the length of a shortest program that prints $\boldsymbol{x}$ and halts (w.r.t. $\mathcal{U}$).

Consider another Turing machine $\mathcal{A}$ and let $p_{\mathcal{A}}$ be a program for $\mathcal{A}$ that prints $\mathbf{x}$ and halts.

Consider a program $s_{\mathcal{A}}$ for $\mathcal{U}$ that simulates $\mathcal{A}$ on $\mathcal{U}$.

Consider an input $s_{\mathcal{A}}p_{\mathcal{A}}$ to $\mathcal{U}$; it prints x and halts. Therefore,

$$K_{\mathcal{U}}(\mathbf{x}) \leq K_{\mathcal{A}}(\mathbf{x}) + c_{\mathcal{A}}$$

where $c_{\mathcal{A}} \coloneqq \ell(s_{\mathcal{A}})$ is a constant.

$K(\mathbf{x})$ differs by a constant for any two universal Turing machines.

Kolmogorov Complexity

Consider a Kolmogorov complexity when the length of x is additionally given.

$K(\mathbf{x} \mid \ell(\mathbf{x}))$ is the length of a shortest program such that when given $n \coloneqq \ell(\mathbf{x})$, prints $\mathbf{x}$ and halts.

Consider a program p: "print the first n-bit $x_1 x_2 \cdots x_n$"; its length $\ell(p)$ is $n + c$.

$$K(\mathbf{x} \mid n) \leq n + c$$

However, we cannot say $K(\mathbf{x}) \leq n + c$, since if n is unknown, p does not know when to stop.

Consider a program p': "read the first $2\lceil \log n \rceil + 2$ bits and decide n; print the next n-bit."

if $n = 5 = 101_{(2)}$, we can describe n as $110011\mathbf{01}$ with $\mathbf{01}$ meaning ','

We can upper bound $K(\mathbf{x}) \leq K(\mathbf{x} \mid n) + 2 \log n + c'$.

Kolmogorov Complexity (Information Theory)

Consider a sequence of i.i.d. (binary) RVs $\boldsymbol{X} = X_1 X_2 \cdots X_n$.

Source coding theorem says $\frac{1}{n}\mathbb{E}[K(\boldsymbol{X} \mid n)] \geq H(X)$. A shortest program is a compression of $\boldsymbol{X}$.

Consider any type t and its type class $\mathrm{typeclass}(t)$. We index each $\mathbf{x} \in \mathrm{typeclass}(t)$.

Consider a program p: "print i-th string $\mathbf{x}$ of $\mathrm{typeclass}(t)$"

Recall $|\mathrm{typeclass}(t)| \leq 2^{nH(t)}$

- To describe a type, $|\mathcal{X}| \log n$ bits suffice. To describe an index, $nH(t)$ bits suffices.

Since the sample frequencies are close to the true probability, we have

$$\frac{1}{n}\mathbb{E}[K(\boldsymbol{X} \mid n)] \leq H(X) + \frac{|\mathcal{X}| \log n}{n} + \frac{c}{n}.$$

Since $K(\mathbf{x}) \leq K(\mathbf{x} \mid n) + 2 \log n + c'$,

$$H(X) \leq \frac{1}{n}\mathbb{E}[K(\boldsymbol{X})] \leq H(X) + \frac{1}{n}(\log|\mathcal{X}| + c')$$

* also holds without expectation

Kolmogorov Complexity (Theory of Computation)

There is no program that decides $K(\mathbf{x}) = k$ (for input $\mathbf{x}, k$).

- Suppose t.c. there is such program p.
- Fix large k such that it satisfies $k > \ell(p) + \log k + c$.
- Consider a program q: "iterates until it finds a string $\mathbf{y}$ where $K(\mathbf{y}) > k$; print $\mathbf{y}$" (by running p) (e.g., in lexicographical order)
- $\ell(q) = \ell(p) + \log k + c$
- However, $k < K(\mathbf{y}) \leq \ell(q) = \ell(p) + \log k + c$, which is a contradiction. "Berry paradox"

Shares the essential spirit with the noncomputability of the Halting problem (Chaitin's number) and Gödel's incompleteness theorem.

Side Notes

Sufficient Statistics

$\{f_\theta(x)\}$: a family of PMFs indexed by θ

X: a sample from a distribution in $\{f_\theta(x)\}$.

$T(X)$: any statistics (such as sample mean or sample variance).

Then $\theta \to X \to T(X)$, and by the data-processing inequality, for any distribution on θ,

$$I(\theta; X) \geq I\big(\theta; T(X)\big).$$

If it holds with equality, i.e., $\theta \to T(X) \to X$, no information is lost.

We say $T(X)$ is a *sufficient statistics* for θ.

Once we know $T(X)$, the remaining randomness in X does not depend on θ.

Example)

$\boldsymbol{X} = X_1, X_2, \ldots, X_{10}$ be an i.i.d. sequence of coin w.p. θ (chosen randomly).

Let $T(\boldsymbol{X}) = X_1 + \cdots + X_{10}$ be the #1's.

$T(\boldsymbol{X})$ is a sufficient statistics for θ.

$$I(\theta; X) = H(\theta) - H(\theta|\boldsymbol{X}) = H(\theta) - H(\theta|T(\boldsymbol{X})) = I\big(\theta; T(\boldsymbol{X})\big)$$

Sufficient Statistics

$\{f_\theta(x)\}$: a family of PMFs indexed by θ

X: a sample from a distribution in $\{f_\theta(x)\}$.

$T(X)$: any statistics (such as sample mean or sample variance).

Then $\theta \to X \to T(X)$, and by the data-processing inequality, for any distribution on θ,

$$I(\theta; X) \geq I\big(\theta; T(X)\big).$$

If it holds with equality, i.e., $\theta \to T(X) \to X$, no information is lost.

We say $T(X)$ is a *sufficient statistics* for θ.

Once we know $T(X)$,
the remaining randomness in X
does not depend on θ.

If T is a function of every other sufficient statistic U, i.e., $\theta \to X \to U(X) \to T(X)$,

$$I(\theta; X) \Big(\geq I\big(\theta; U(X)\big)\Big) \geq I\big(\theta; T(X)\big).$$

If it holds with equality, i.e., $\theta \to T(X) \to U(X) \to X$,

we say $T(X)$ is a *minimal sufficient statistics* for θ.

Fano's Inequality

Theorem. For any estimator $\hat{X}$ s.t. $X \to Y \to \hat{X}$, we have

$$H(X|Y) \le H\left(X|\hat{X}\right) \le H(1_{\hat{X} \ne X}) + \Pr\left[\hat{X} \ne X\right] \log|\mathcal{X}| \le 1 + \Pr\left[\hat{X} \ne X\right] \log|\mathcal{X}|\,.$$

Probability of Error and Entropy

Lemma. If X and X' are i.i.d., $\Pr[X = X'] \ge 2^{-H(X)}$ with equality iff X has a uniform distribution.

Corollary. If $X \sim p$ and $X' \sim q$ are independent and $\mathcal{X} = \mathcal{X}'$,

$$\Pr[X = X'] \ge 2^{-H(p) - D(p \| q)}$$

$$\Pr[X = X'] \ge 2^{-H(q) - D(q \| p)}$$

Stochastic Process and Entropy Rate

Stochastic process $\{X_i\}$: an indexed sequence of RVs with arbitrary dependence

Stationary stochastic process: joint distribution of any subset is invariant w.r.t. shifts in index

$$\Pr[(X_1, X_2, \ldots, X_n) = (x_1, x_2, \ldots, x_n)] = \Pr[(X_{1+\ell}, X_{2+\ell}, \ldots, X_{n+\ell}) = (x_1, x_2, \ldots, x_n)]$$

Entropy rate

Definition 1 (entropy per symbol).

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n)$$ when the limit exists

Definition 2 (conditional entropy of the last).

$$H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n \mid X_{n-1}, X_{n-2}, \ldots, X_1)$$ when the limit exists

Theorem. For a stationary stochastic process, $H(\mathcal{X}) = H'(\mathcal{X})$.

General AEP

AEP

For any i.i.d. process, in probability,

$$-\frac{1}{n}\log p(X_1, \dots, X_n) \to H(X)$$

General AEP (chapter 16)

For any stationary *ergodic* process, with probability 1,

$$-\frac{1}{n}\log p(X_1, \dots, X_n) \to H(\mathcal{X})$$

Entropy Rate of Stationary Markov Chain

With initial dist. as stationary dist. μ, Markov chain is a stationary process.

$$H(\mathcal{X}) = H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n \mid X_{n-1}, X_{n-2}, \dots, X_1) \underset{\text{Markovity}}{=} \lim_{n\to\infty} H(X_n \mid X_{n-1}) \underset{\text{stationarity}}{=} H(X_2 \mid X_1)$$

Entropy Rate of Functions of Markov Chain

Let $\{X_i\}$ be a stationary Markov chain. Consider $\{Y_i\}$ where $Y_i = \phi(X_i)$. $\{Y_i\}$ does not necessarily form a Markov chain.

How to know $H(Y_n \mid Y_{n-1}, Y_{n-2}, \dots, Y_1) \approx H(\mathcal{Y})$ for any n?

$$H(Y_n \mid Y_{n-1}, Y_{n-2}, \dots, Y_1, X_1) \le H(\mathcal{Y}) \le H(Y_n \mid Y_{n-1}, Y_{n-2}, \dots, Y_1)$$

$$\lim_{n\to\infty} H(Y_n \mid Y_{n-1}, Y_{n-2}, \dots, Y_1, X_1) = H(\mathcal{Y}) = \lim_{n\to\infty} H(Y_n \mid Y_{n-1}, Y_{n-2}, \dots, Y_1)$$

Relates to a *hidden Markov model* (HMM)

Thank You