

Review

I-Hsiang Wang

Department of Electrical Engineering
National Taiwan University

ihwang@ntu.edu.tw

November 22, 2023

What we have covered so far

- Representing information (losslessly, or with a fidelity criterion)
- Delivering information reliably (with or without cost constraints)
- Learning a bit of information
- Various information measures

1 Information measures

2 Coding theorems: setups and results

3 Tools for establishing fundamental limits

Entropy

Entropy

The entropy of a discrete RV $X \sim p_X \in \mathcal{P}(\mathcal{X})$ is the expectation of the self information

$$H(X) \equiv H(p_X) = \mathbb{E}_{X \sim p_X} \left[\log \frac{1}{p_X(X)} \right]$$

Conditional entropy

The conditional entropy of X given Y with conditional PMF $p_{X|Y}$ is

$$H(X|Y) = \mathbb{E}_{X,Y} \left[\log \frac{1}{p_{X|Y}(X|Y)} \right] = \mathbb{E}_Y [H(p_{X|Y}(\cdot|Y))]$$

Differential entropy

Differential entropy

The differential entropy of a continuous RV X with density f_X is

$$h(X) \equiv h(f_X) = \mathbb{E}_{X \sim f_X} \left[\log \frac{1}{f_X(X)} \right]$$

Conditional differential entropy

The conditional differential entropy of X given Y with conditional density $f_{X|Y}$ is

$$h(X|Y) = \mathbb{E}_{X,Y} \left[\log \frac{1}{f_{X|Y}(X|Y)} \right] = \mathbb{E}_Y [h(f_{X|Y}(\cdot|Y))]$$

Relative entropy/Information divergence

Relative entropy/KL divergence

The relative entropy (KL divergence, information divergence) of P_X from Q_X is

$$D(P_X \| Q_X) = E_{X \sim P_X} \left[\log \frac{P_X(X)}{Q_X(X)} \right]$$

Conditional relative entropy/KL divergence

The conditional relative entropy (KL divergence, information divergence) of $P_{Y|X}$ from $Q_{Y|X}$ given P_X is

$$D(P_{Y|X} \| Q_{Y|X} | P_X) = E_{X \sim P_X} [D(P_{Y|X}(\cdot|X) \| Q_{Y|X}(\cdot|X))]$$

Mutual information

Mutual information

The mutual information between two RVs $(X, Y) \sim P_{X,Y}$ is

$$\begin{aligned} I(X; Y) &\equiv I(P_X, P_{Y|X}) = E_{X,Y \sim P_{X,Y}} \left[\log \frac{P_{X,Y}(X, Y)}{P_X(X)P_Y(Y)} \right] \\ &= D(P_{X,Y} \| P_X \times P_Y) = D(P_{Y|X} \| P_Y | P_X) \end{aligned}$$

Conditional mutual information

For $(X, Y, Z) \sim P_{X,Y,Z}$, the conditional MI between X, Y given Z is

$$\begin{aligned} I(X; Y|Z) &= E_{X,Y,Z \sim P_{X,Y,Z}} \left[\log \frac{P_{X,Y|Z}(X, Y|Z)}{P_{X|Z}(X|Z)P_{Y|Z}(Y|Z)} \right] \\ &= E_{Z \sim P_Z} [I(P_{X|Z}, P_{Y|X,Z})] \end{aligned}$$

Properties

Nonnegativity

$$\begin{aligned} H(X|Y) &\geq 0 & I(X; Y|Z) &\geq 0 \\ D(P_{Y|X} \| Q_{Y|X} | P_X) &\geq 0 & \text{but } h(X|Y) &\geq 0 \end{aligned}$$

Convexity

$$\begin{aligned} H(P_X) &: \text{concave in } P_X \\ I(P_X, P_{Y|X}) &: \text{concave in } P_X \text{ when } P_{Y|X} \text{ is fixed} \\ &\quad \text{convex in } P_{Y|X} \text{ when } P_X \text{ is fixed} \\ D(P \| Q) &: \text{convex in } (P, Q) \end{aligned}$$

Chain rule

$$H(X, Y|Z) = H(X|Y, Z) + H(Y|Z)$$

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

$$D(P_{X,Y|Z} \| Q_{X,Y|Z} | P_Z) = D(P_{X|Y,Z} \| Q_{X|Y,Z} | P_{Y,Z}) + D(P_{Y|Z} \| Q_{Y|Z} | P_Z)$$

Conditioning

$$H(X|Z) \geq H(X|Y, Z)$$

$$I(X; Z|W) \geq I(X; Z|Y, W) \quad \text{if } Y - (X, W) - Z$$

$$D(P_Y \| Q_Y) \leq D(P_{Y|X} \| Q_{Y|X} | P_X)$$

Data processing

$$I(X; Y) \geq I(X; Z) \quad \text{if } X - Y - Z$$

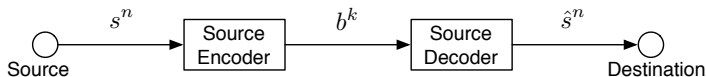
$$D(P_X \| Q_X) \geq D(P_Y \| Q_Y) \quad \begin{array}{l} \text{if } P_Y(\cdot) = E_{X \sim P_X} [W_{Y|X}(\cdot|X)] \\ \text{and } Q_Y(\cdot) = E_{X \sim Q_X} [W_{Y|X}(\cdot|X)] \end{array}$$

1 Information measures

2 Coding theorems: setups and results

3 Tools for establishing fundamental limits

Representing information losslessly



Fixed-to-fixed source coding: setup

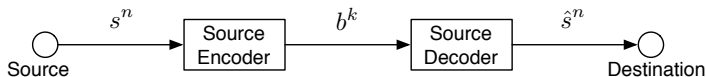
For a sequence of $(n, \lfloor nR \rfloor)$ -codes (indexed by $n = 1, 2, \dots$), $k = \lfloor nR \rfloor$, we are interested in the rate R .

R is achievable if there exist a sequence of $(n, \lfloor nR \rfloor)$ -codes such that

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

Minimum compression rate: $R^* := \inf\{R \mid R \text{ is achievable}\}.$

Representing information with a fidelity criterion



Fixed-to-fixed source coding: setup

For a sequence of $(n, \lfloor nR \rfloor, D)$ -codes (indexed by $n = 1, 2, \dots$), $k = \lfloor nR \rfloor$, we are interested in the rate and the asymptotic distortion (R, D) .

(R, D) is achievable if there exist a sequence of $(n, \lfloor nR \rfloor, D)$ -codes such that

$$\limsup_{n \rightarrow \infty} D^{(n)} \leq D, \text{ where } D^{(n)} := E[d(S^n, \hat{S}^n)] = \frac{1}{n} \sum_{i=1}^n E[d(S_i, \hat{S}_i)].$$

Rate distortion function: $R(D) := \inf\{R \mid (R, D) \text{ is achievable}\}.$

Fixed-to-fixed source coding: fundamental limits

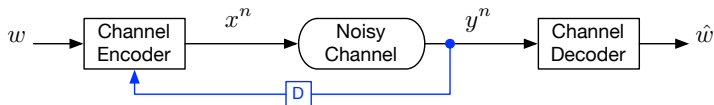
Lossless: for a discrete memoryless source (DMS) $S_i \stackrel{\text{i.i.d.}}{\sim} P$,

$$R^* = H(S) \equiv H(P)$$

Lossy: for a memoryless source $S_i \stackrel{\text{i.i.d.}}{\sim} P$,

$$R(D) = \inf_{(\hat{S}, S)} \left\{ I(S; \hat{S}) \mid S \sim P, E[d(S, \hat{S})] \leq D \right\}.$$

Delivering information reliably



Fixed-to-fixed channel coding: setup

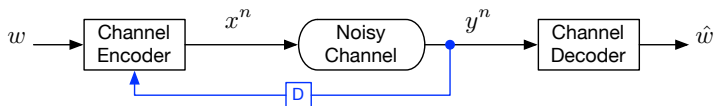
For a sequence of $(n, \lceil nR \rceil)$ -codes (indexed by $n = 1, 2, \dots$), $k = \lceil nR \rceil$, we are interested in the rate R .

R is achievable if there exist a sequence of $(n, \lceil nR \rceil)$ -codes such that

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

Channel capacity: $C := \sup \{R \mid R \text{ is achievable}\}.$

Delivering information reliably with input cost



Fixed-to-fixed channel coding: setup

For a sequence of $(n, \lceil nR \rceil, B)$ -codes (indexed by $n = 1, 2, \dots$), $k = \lceil nR \rceil$, we are interested in the rate R and the input cost B .

(R, B) is achievable if there exist a sequence of $(n, \lceil nR \rceil, B)$ -codes such that

$$\lim_{n \rightarrow \infty} P_e^{(n)} = 0 \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^n b(x_i) \leq B \quad \forall n \in \mathbb{N}.$$

Channel capacity: $C(B) := \sup \{R \mid (R, B) \text{ is achievable}\}.$

Fixed-to-fixed channel coding: fundamental limits

For a DMC $(\mathcal{X}, P_{Y|X}, \mathcal{Y})$ with or without feedback,

$$C = \max_X I(X; Y)$$

For a memoryless channel $(\mathcal{X}, P_{Y|X}, \mathcal{Y})$ with or without feedback,

$$C(B) = \sup_X \{I(X; Y) \mid E[b(X)] \leq B\}$$

Remark:

For channels with memory, feedback may increase the capacity.

Learning a bit of information

$$\mathcal{H}_0 : X_i \stackrel{\text{i.i.d.}}{\sim} P_0, i = 1, 2, \dots, n \quad \equiv \quad X^n \sim P_0^{\otimes n}$$

$$\mathcal{H}_1 : X_i \stackrel{\text{i.i.d.}}{\sim} P_1, i = 1, 2, \dots, n \quad \equiv \quad X^n \sim P_1^{\otimes n}$$

Binary hypothesis testing: Stein's asymptotic regime

For a sequence of $(n, \epsilon, e_{0|1})$ -tests $\{\phi_n\}$ (indexed by $n = 1, 2, \dots$), we are interested in the type-II error exponent $e_{0|1}$.

$(\epsilon, e_{0|1})$ is achievable if there exist a sequence of $(n, \epsilon, e_{0|1})$ -tests such that

$$\pi_{1|0}^{(n)} \leq \epsilon, \quad \liminf_{n \rightarrow \infty} \frac{1}{n} \log \frac{1}{\pi_{0|1}^{(n)}} \geq e_{0|1}.$$

Maximum error exponent: $e_{0|1}^{*,\text{St}}(\epsilon) := \sup\{e_{0|1} \mid (\epsilon, e_{0|1}) \text{ is achievable}\}.$

Binary hypothesis testing: fundamental limits

For arbitrary sample sizes, randomized likelihood ratio tests attain the optimal trade-off between the two types of error probabilities.

In Stein's asymptotic regime,

$$\forall \epsilon \in (0, 1), e_{0|1}^{*,\text{St}}(\epsilon) = D(P_0 \| P_1)$$

1 Information measures

2 Coding theorems: setups and results

3 Tools for establishing fundamental limits

Typicality

Weakly typical sequence

For $\delta > 0$, a sequence x^n is called **δ -weakly-typical** with respect to $X \sim p_X$ if

$$\left| \frac{1}{n} \log \frac{1}{p_X^{\otimes n}(x^n)} - H(X) \right| \leq \delta,$$

The **δ -typical set**

$$\mathcal{A}_\delta^{(n)}(X) \equiv \mathcal{A}_\delta^{(n)}(p_X) := \{x^n \in \mathcal{X}^n \mid x^n \text{ is } \delta\text{-weakly typical w.r.t. } X \sim p_X\}.$$

Asymptotic equipartition property (AEP)

AEP, a consequence of LLN, is useful for establishing coding theorems.

Weak typicality:

$$1 \quad \forall x^n \in \mathcal{A}_\delta^{(n)}(p_X), 2^{-n(H(p_X)+\delta)} \leq p_X^{\otimes n}(x^n) \leq 2^{-n(H(p_X)-\delta)}.$$

$$2 \quad p_X^{\otimes n} \left\{ \mathcal{A}_\delta^{(n)}(p_X) \right\} \geq 1 - \epsilon \text{ for } n \text{ large enough.}$$

$$3 \quad |\mathcal{A}_\delta^{(n)}(p_X)| \leq 2^{n(H(p_X)+\delta)}.$$

$$4 \quad |\mathcal{A}_\delta^{(n)}(p_X)| \geq (1 - \epsilon)2^{n(H(p_X)-\delta)} \text{ for } n \text{ large enough.}$$

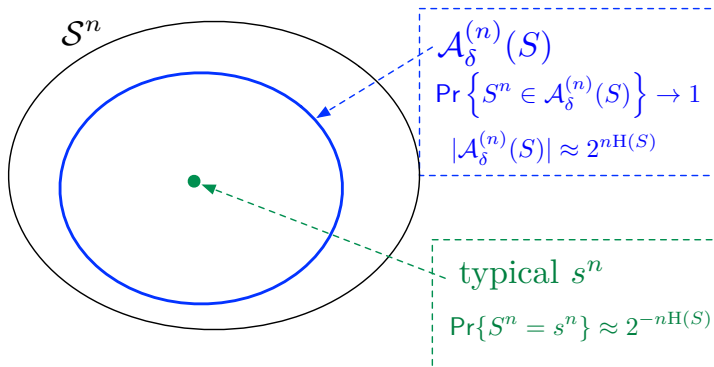


Figure: A simple illustration of AEP

Robust typical sequence

For $\varepsilon \in (0, 1)$, a sequence x^n is called ε -robust typical with respect to a discrete $X \sim p_X$ if

$$|\hat{p}_{x^n}(a) - p_X(a)| \leq \varepsilon p_X(a), \forall a \in \mathcal{X}.$$

The ε -typical set

$$\mathcal{T}_\varepsilon^{(n)}(X) \equiv \mathcal{T}_\varepsilon^{(n)}(p_X) := \{x^n \in \mathcal{X}^n \mid x^n \text{ is } \varepsilon\text{-typical w.r.t. } X \sim p_X\}.$$

Remark:

- Weak typicality is more general, works beyond i.i.d., and can be extended to continuous data with density.
- Robust typicality works for discrete memoryless data only, but it has more useful properties such as conditional typicality, typical average lemma, etc..