Review

I-Hsiang Wang

Department of Electrical Engineering
National Taiwan University

ihwang@ntu.edu.tw

November 27, 2024

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

Entropy

Entropy

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

$$\mathrm{H}(X) \equiv \mathrm{H}(\mathsf{p}_X) = \mathsf{E}_{X \sim \mathsf{p}_X}\left[\log \frac{1}{\mathsf{p}_X(X)}\right]$$

Conditional entropy

The conditional entropy of X given Y with conditional PMF $\mathsf{p}_{X|Y}$ is

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

Differential entropy

Differential entropy

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

$$\mathrm{h}(X) \equiv \mathrm{h}(\mathsf{f}_X) = \mathsf{E}_{X \sim \mathsf{f}_X}\left[\log \frac{1}{\mathsf{f}_X(X)}\right]$$

Conditional differential entropy

The conditional differential entropy of X given Y with conditional density $\mathsf{f}_{X|Y}$ is

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

Relative entropy/Information divergence

Relative entropy/KL divergence

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

$$\mathrm{D}(\mathsf{P}_X \| \mathsf{Q}_X) = \mathsf{E}_{X \sim \mathsf{P}_X}\left[\log \frac{\mathsf{P}_X(X)}{\mathsf{Q}_X(X)}\right]$$

Conditional relative entropy/KL divergence

The conditional relative entropy (KL divergence, information divergence) of $\mathsf{P}_{Y|X}$ from $\mathsf{Q}_{Y|X}$ given P_X is

$$\mathrm{D}\left(\mathsf{P}_{Y|X} \middle\| \mathsf{Q}_{Y|X} \middle| \mathsf{P}_X\right) = \mathsf{E}_{X \sim \mathsf{P}_X}\left[\mathrm{D}\left(\mathsf{P}_{Y|X}(\cdot|X) \middle\| \mathsf{Q}_{Y|X}(\cdot|X)\right)\right]$$

Mutual information

Mutual information

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

$$\begin{aligned}\mathrm{I}(X;Y) \equiv \mathsf{I}(\mathsf{P}_X, \mathsf{P}_{Y|X}) &= \mathsf{E}_{X,Y\sim \mathsf{P}_{X,Y}}\left[\log \frac{\mathsf{P}_{X,Y}(X,Y)}{\mathsf{P}_X(X)\mathsf{P}_Y(Y)}\right] \\ &= \mathrm{D}\left(\mathsf{P}_{X,Y} \| \mathsf{P}_X \times \mathsf{P}_Y\right) = \mathrm{D}\left(\mathsf{P}_{Y|X} \middle\| \mathsf{P}_Y \middle| \mathsf{P}_X\right)\end{aligned}$$

Conditional mutual information

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

$$\begin{aligned}\mathrm{I}(X;Y|Z) &= \mathsf{E}_{X,Y,Z\sim \mathsf{P}_{X,Y,Z}}\left[\log \frac{\mathsf{P}_{X,Y|Z}(X,Y|Z)}{\mathsf{P}_{X|Z}(X|Z)\mathsf{P}_{Y|Z}(Y|Z)}\right] \\ &= \mathsf{E}_{Z\sim \mathsf{P}_Z}\left[\mathsf{I}(\mathsf{P}_{X|Z}, \mathsf{P}_{Y|X,Z})\right]\end{aligned}$$

Properties

Nonnegativity

$$\mathrm{H}(X|Y) \geq 0 \qquad \mathrm{I}(X;Y|Z) \geq 0$$

$$\mathrm{D}\left(\mathsf{P}_{Y|X} \middle\| \mathsf{Q}_{Y|X} \middle| \mathsf{P}_X\right) \geq 0 \quad \text{but} \quad \mathrm{h}(X|Y) \gtreqless 0$$

Convexity

$$\begin{aligned} \mathrm{H}(\mathsf{P}_X) &: \text{concave in } \mathsf{P}_X \\ \mathsf{I}\left(\mathsf{P}_X, \mathsf{P}_{Y|X}\right) &: \text{concave in } \mathsf{P}_X \text{ when } \mathsf{P}_{Y|X} \text{ is fixed} \\ &\ \text{convex in } \mathsf{P}_{Y|X} \text{ when } \mathsf{P}_X \text{ is fixed} \\ \mathrm{D}(\mathsf{P}\|\mathsf{Q}) &: \text{convex in } (\mathsf{P}, \mathsf{Q}) \end{aligned}$$

Chain rule

$$
\begin{aligned}
\mathrm{H}(X, Y|Z) &= \mathrm{H}(X|Y, Z) + \mathrm{H}(Y|Z) \\
\mathrm{I}(X, Y; Z|W) &= \mathrm{I}(X; Z|Y, W) + \mathrm{I}(Y; Z|W) \\
\mathrm{D}\left(\mathsf{P}_{X,Y|Z} \middle\| \mathsf{Q}_{X,Y|Z} \middle| \mathsf{P}_Z\right) &= \mathrm{D}\left(\mathsf{P}_{X|Y,Z} \middle\| \mathsf{Q}_{X|Y,Z} \middle| \mathsf{P}_{Y,Z}\right) + \mathrm{D}\left(\mathsf{P}_{Y|Z} \middle\| \mathsf{Q}_{Y|Z} \middle| \mathsf{P}_Z\right)
\end{aligned}
$$

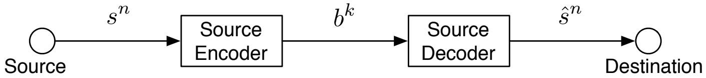
Conditioning

$$
\begin{aligned}
\mathrm{H}(X|Z) &\ge \mathrm{H}(X|Y, Z) \\
\mathrm{I}(X; Z|W) &\ge \mathrm{I}(X; Z|Y, W) \quad \text{if } Y - (X, W) - Z \\
\mathrm{D}(\mathsf{P}_Y \| \mathsf{Q}_Y) &\le \mathrm{D}\left(\mathsf{P}_{Y|X} \middle\| \mathsf{Q}_{Y|X} \middle| \mathsf{P}_X\right)
\end{aligned}
$$

Data processing

$$
\begin{aligned}
\mathrm{I}(X;Y) \ge \mathrm{I}(X;Z) \quad & \text{if } X - Y - Z \\
\mathrm{D}(\mathsf{P}_X \| \mathsf{Q}_X) \ge \mathrm{D}(\mathsf{P}_Y \| \mathsf{Q}_Y) \quad & \text{if } \mathsf{P}_Y(\cdot) = \mathsf{E}_{X \sim \mathsf{P}_X}\left[\mathsf{W}_{Y|X}(\cdot|X)\right] \\
& \text{and } \mathsf{Q}_Y(\cdot) = \mathsf{E}_{X \sim \mathsf{Q}_X}\left[\mathsf{W}_{Y|X}(\cdot|X)\right]
\end{aligned}
$$

Representing information losslessly



Fixed-to-fixed source coding: setup

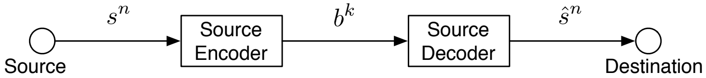
For a sequence of $(n, \lfloor n\mathrm{R} \rfloor)$-codes (indexed by $n = 1, 2, ...$), $k = \lfloor n\mathrm{R} \rfloor$, we are interested in the rate R.

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

$$\lim_{n\to\infty} \mathsf{P}_{\mathsf{e}}^{(n)} = 0.$$

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

Representing information with a fidelity criterion



Fixed-to-fixed source coding: setup

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

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

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

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

Fixed-to-fixed source coding: fundamental limits

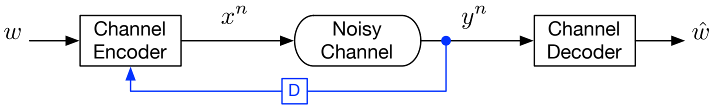
Lossless: for a discrete memoryless source (DMS) $S_i \overset{\text{i.i.d.}}{\sim} \mathsf{P}$,

$$\mathrm{R}^* = \mathrm{H}(S) \equiv \mathrm{H}(\mathsf{P})$$

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

$$\mathrm{R}(\mathsf{D}) = \inf_{(\hat{S},S)} \left\{ \mathrm{I}\left(S;\hat{S}\right) \,\middle|\, S \sim \mathsf{P},\ \mathsf{E}\left[d(S,\hat{S})\right] \le \mathsf{D} \right\}.$$

Delivering information reliably



Fixed-to-fixed channel coding: setup

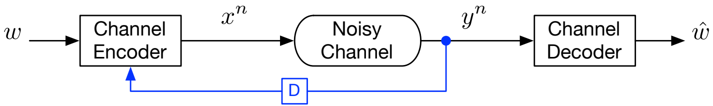
For a sequence of $(n, \lceil n\mathrm{R} \rceil)$-codes (indexed by $n = 1, 2, ...$), $k = \lceil n\mathrm{R} \rceil$, we are interested in the rate R.

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

$$\lim_{n \to \infty} \mathsf{P}_{\mathsf{e}}^{(n)} = 0.$$

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

Delivering information reliably with input cost



Fixed-to-fixed channel coding: setup

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

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

$$\lim_{n \to \infty} \mathsf{P}_{\mathsf{e}}^{(n)} = 0 \text{ and } \frac{1}{n} \sum_{i=1}^{n} b(x_i) \le \mathsf{B} \quad \forall\, n \in \mathbb{N}.$$

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

Fixed-to-fixed channel coding: fundamental limits

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

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

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

$$\mathrm{C}(\mathsf{B}) = \sup_{X} \{\mathrm{I}(X;Y) \,|\, \mathsf{E}[b(X)] \le \mathsf{B}\}$$

Remark:

For channels with memory, feedback may increase the capacity.

Learning a bit of information

$$\mathcal{H}_0 : X_i \overset{\text{i.i.d.}}{\sim} \mathsf{P}_0,\ i = 1, 2, \ldots, n \quad \equiv \quad X^n \sim \mathsf{P}_0^{\otimes n}$$
$$\mathcal{H}_1 : X_i \overset{\text{i.i.d.}}{\sim} \mathsf{P}_1,\ i = 1, 2, \ldots, n \quad \equiv \quad X^n \sim \mathsf{P}_1^{\otimes n}$$

Binary hypothesis testing: Stein's asymptotic regime

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

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

$$\pi_{1|0}^{(n)} \le \epsilon, \quad \liminf_{n\to\infty} \frac{1}{n} \log \frac{1}{\pi_{0|1}^{(n)}} \ge \mathrm{e}_{0|1}.$$

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

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,

$$\forall\, \epsilon \in (0,1),\; \mathrm{e}_{0|1}^{*,\mathsf{St}}(\epsilon) = \mathrm{D}\left(\mathsf{P}_0 \| \mathsf{P}_1\right)$$

Typicality

Weakly typical sequence

For $\delta > 0$, a sequence x^n is called δ-weakly-typical with respect to $X \sim \mathsf{p}_X$ if

$$\left| \frac{1}{n} \log \frac{1}{\mathsf{p}_X^{\otimes n}(x^n)} - \mathrm{H}(X) \right| \le \delta,$$

The δ-typical set

$$\mathcal{A}_\delta^{(n)}(X) \equiv \mathcal{A}_\delta^{(n)}(\mathsf{p}_X) := \left\{ x^n \in \mathcal{X}^n \,\middle|\, x^n \text{ is } \delta\text{-weakly typical w.r.t. } X \sim \mathsf{p}_X \right\}.$$

Asymptotic equipartition property (AEP)

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

Weak typicality:

1. $\forall\, x^n \in \mathcal{A}_\delta^{(n)}(\mathsf{p}_X)$, $2^{-n(\mathrm{H}(\mathsf{p}_X)+\delta)} \le \mathsf{p}_X^{\otimes n}(x^n) \le 2^{-n(\mathrm{H}(\mathsf{p}_X)-\delta)}$.
2. $\mathsf{p}_X^{\otimes n}\left\{\mathcal{A}_\delta^{(n)}(\mathsf{p}_X)\right\} \ge 1-\epsilon$ for n large enough.
3. $|\mathcal{A}_\delta^{(n)}(\mathsf{p}_X)| \le 2^{n(\mathrm{H}(\mathsf{p}_X)+\delta)}$.
4. $|\mathcal{A}_\delta^{(n)}(\mathsf{p}_X)| \ge (1-\epsilon)2^{n(\mathrm{H}(\mathsf{p}_X)-\delta)}$ for n large enough.

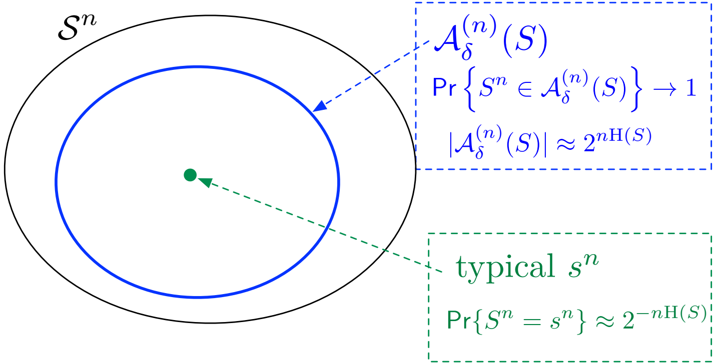


Figure: A simple illustration of AEP

Robust typical sequence

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

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

The ε-typical set

$$\mathcal{T}_\varepsilon^{(n)}(X) \equiv \mathcal{T}_\varepsilon^{(n)}(\mathsf{p}_X) := \left\{x^n \in \mathcal{X}^n \,|\, x^n \text{ is } \varepsilon\text{-typical w.r.t. } X \sim \mathsf{p}_X\right\}.$$

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..