ECE253/CSE208 Introduction to Information Theory

Lecture 16: Summary

Dr. Yu Zhang

ECE Department
University of California, Santa Cruz

- Chap 1–10 of *Elements of Information Theory (2nd Edition) by Thomas Cover & Joy Thomas*.

Key Concepts

- Entropy and differential entropy; KL divergence; mutual information.
- DPI; sufficient statistics; convex functions; Jensen's ineq; Fano's ineq.
- AEP; typical set; entropy rate; Markov chain; stationary stochastic process.
- Source coding theorem; Kraft-McMillan inequality; competitive optimality.
- Lossless source coding: Huffman, Shannon, SF, SFE, and arithmetic coding.
- Horse race; log wealth relative; double rate; financial value of side information.
- Channel coding theorem and its proof; channel capacity of "simple" channels.
- Channel capacity of AWGN/parallel AWGN/feedback; water-filling power allocation.
- Rate-distortion theory; Blahut–Arimoto algorithm.

For Discrete Distributions

PMF $p(x)$	Entropy
Definition	$H(X) = -\sum_x p(x) \log p(x)$
Bounds	$[0, \log \lvert\mathcal{X}\rvert]$
$H_{\max}$ distribution	Uniform
Translation	$H(X + c) = H(X)$
Scaling	$H(cX) = H(X)$
Joint entropy	$H(X, Y) = -\mathrm{E}_{(X,Y)} \log p(X, Y)$
Conditional entropy	$H(Y \mid X) = -\mathrm{E}_{(X,Y)} \log p(Y \mid X)$
Relative entropy	$D(p \| q) = \sum p \log \frac{p}{q}$
Mutual information	$I(X; Y) = D\big(p(X, Y) \| p(X)p(Y)\big)$
Chain rule	$H(X^n) = \sum_{i=1}^n H(X_i \mid X^{i-1})$
AEP (i.i.d. X^n)	$-\frac{1}{n} \log p(X^n) \xrightarrow{\text{i.p.}} H(X)$
Typical set	$A_\epsilon^{(n)} = \big\{x^n : \lvert -\frac{1}{n} \log p(x^n) - H(X)\rvert \le \epsilon\big\}$

For Continuous Distributions

PDF $p(x)$	Differential Entropy
Definition	$h(X) = -\int_{\mathcal{S}} f(x) \log f(x)\, dx$
Bounds	$(-\infty, \frac{1}{2}\log(2\pi e \sigma^2)]$
$H_{\max}$ distribution	Gaussian
Translation	$h(X+c) = h(X)$
Scaling	$h(cX) = h(X) + \log\lvert c\rvert$
Joint entropy	$h(X,Y) = -\mathrm{E}_{(X,Y)} \log f(X,Y)$
Conditional entropy	$h(Y\lvert X) = -\mathrm{E}_{(X,Y)} \log f(Y\lvert X)$
Relative entropy	$D(f\lvert\lvert g) = \int f \log(\frac{f}{g})$
Mutual information	$I(X;Y) = D\big(f(x,y)\lvert\lvert f(x)f(y)\big)$
Chain rule	$h(X^n) = \sum_{i=1}^{n} h(X_i\lvert X^{i-1})$
AEP (i.i.d. X^n)	$-\frac{1}{n}\log f(X^n) \xrightarrow{\text{i.p.}} h(X)$
Typical set	$A_\epsilon^{(n)} = \left\{x^n : \lvert -\frac{1}{n}\log f(x^n) - h(X)\rvert \le \epsilon\right\}$

Typical Sequences and Set

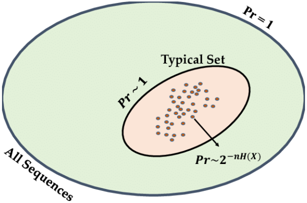

Figure: For n sufficiently large, all typical sequences have about the same probability $2^{nH(X)}$ (*asymptotic equipartition*). Everything outside the typical set has a negligible probability.

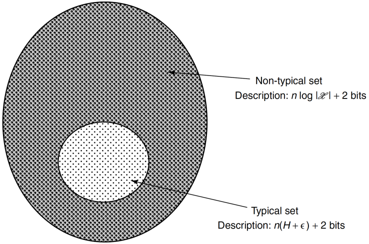

Figure: Encoding for the typical set: On average $H(X)$ bits are needed to encode X^n per symbol.

Lossless Source Coding

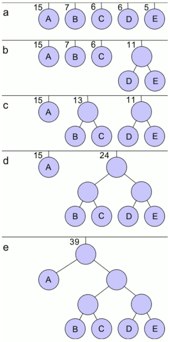

Fig 1. Huffman tree (bottom-up)

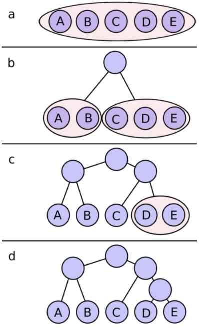

Fig 2. SF tree (top-down)

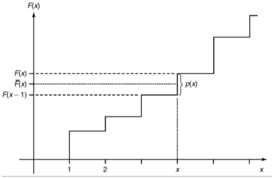

Fig 3. SFE coding

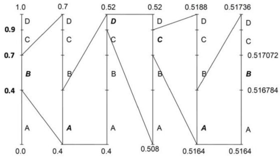

Fig 4. Arithmetic coding

Channel Coding Theorem

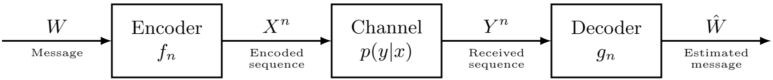

Figure: Channel capacity C is the sharp threshold between reliable and unreliable communication.

- *Information* channel capacity: $C = \max\limits_{p(x)} I(X;Y)$
- *Operational* channel capacity: number of bits that are transmitted reliably (with an arbitrarily small probability of error) per channel usage.
- Information channel capacity $=$ operational channel capacity.
- **Achievability**: All rates below capacity $R < C$ are achievable.
- **Converse**: $(2^{nR}, n)$ code with probability of error $\lambda^{(n)} \xrightarrow{n\to\infty} 0$ must have $R \leq C$.

Shannon Limit of Channel Capacity

$$C_{\mathsf{AWGN}} = W \log_2 \left(1 + \frac{P}{N_0 W}\right)$$

$$\max_{P_i \geq 0} \quad \sum_{i=1}^{N} \frac{1}{2} \log \left(1 + \frac{P_i}{N_i}\right) \tag{1a}$$

$$\text{s.t.} \quad \sum^{N} P_i = P \tag{1b}$$

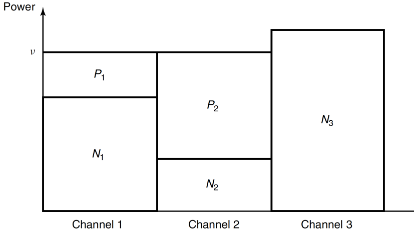

Figure: Water-filling optimal power allocation: **allocate more power in less noisy channels**. The optimal solution is obtained as: $P_i^* = (\nu - N_i)^+, \; i = 1, 2, \ldots k$, where the water level ν satisfies $\sum_i (\nu - N_i)^+ = P$.

Rate Distortion Coding Theorem

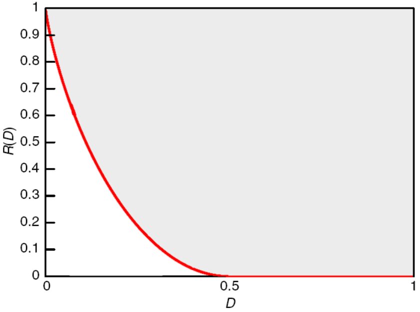

Figure: The rate-distortion function of a Bernoulli random variable with Hamming distortion.

Theorem (Rate Distortion Coding Theorem)

The rate distortion function for an i.i.d. source X with distribution $p(x)$ and bounded distortion function $d(x, \hat{x})$ is equal to the associated information rate distortion function. That is, $R(D) = R^{(I)}(D)$.

Thank You!

Email: <zhangy@ucsc.edu>
Homepage: https://people.ucsc.edu/~yzhan419/