EECS 126 — 2.23 LECTURE 10

INFORMATION THEORY

Traces back to one paper (Shannon 1948). This paper kicked off the information age by addressing the following questions:

1. how reliably and quickly can I communicate a message over a noisy channel?
2. how many bits do I need to losslessly represent an observation?

SOURCE CODING / COMPRESSION

For a discrete RV $X \sim P_X$, we define the Shannon entropy:

$$H(X) := \sum_x P_X(x) \log_2 \frac{1}{P_X(x)} = E\left[\log_2 \frac{1}{P_X(x)}\right]$$

This is the uncertainty/randomness of X on average (similar to $\mathrm{Var}(X)$). The interpretation of $H(X)$ as uncertainty is justified by the Source Coding Theorem: $H(X)$ is # bits needed to describe X on average.

THEOREM For any $\varepsilon > 0$, discrete $X_1, X_2, \ldots, X_n \sim_{IID} P_X$ can be losslessly represented using $\leq n(H(X)+\varepsilon)$ bits for all n sufficiently large. Any representation using $< nH(X)$ bits must lose information.

This means that descriptions $\leq n(H(X)+\varepsilon)$ bits are possible, and descriptions $< nH(X)$ are not.

For an example, see today's 170 notes (Huffman Codes).

If we flip a coin with $p = 0.11$ n times, we can describe the outcome in $n/2$ bits. How?

CONCENTRATION

From a lot of randomness comes determinism.

For a sequence $X_1, X_2, \ldots, X_n$, let probability of observation be $P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^n P_X(x_i)$.

THEOREM Asymptotic Equipartition Theorem: if $(X_i)_{i \geq 1} \sim_{IID} P_X$, $-\frac{1}{n} \log P(X_1, X_2 \ldots) \to H(X)$ in probability; as in, with overwhelming probability, $P(X_1 \ldots X_n) = 2^{-nH(X)}$

PROOF WLLN: $-\frac{1}{n} \log P(X_1 \ldots X_n) = \frac{1}{n} \sum \log \frac{1}{P_X(x_i)} \to \underbrace{E\left[\log \frac{1}{P_X(x)}\right]}_{H(X)}$ in probability.