TODAY

ENTROPY

- Yesterday: We claimed that $H(X)$ = expected number of bits needed to convey X to receiver who knows P_X.

- Was a lie:
  - Problem: $H(X) + H(Y) \neq H(X,Y)$ when X, Y independent according to definition above
  - Example: $X = 0$ w.p. $.99$
    $= 1$ w.p. $.01$
  - Still need one bit to convey X
  - But $H(X_1 \ldots X_{100})$ where $X_1 \ldots X_{100}$ i.i.d. $\sim P_X$

  is much less (Why?)

Correct "Operational" Definition

- $H(X) \triangleq$ amortized expected # bits to convey ~~of~~ many copies of X

  ie, convey $X_1 \dots X_n$ where $X_1 \dots X_n$ i.i.d. $\sim p_x$.

  amortized $= \frac{1}{n}$ (# bits) ; take $\lim_{n \to \infty}$.

- Let E_n, D_n be functions s.t.

$$E_n : \Omega^n \to \{0,1\}^*$$

$$D_n : \{0,1\}^* \to \Omega^n$$

$\forall\, \underline{x} \in \Omega^n$ $D_n(E_n(\underline{x})) = \underline{x}$; E_n prefix-free

Then
$$H(X) \triangleq \lim_{n \to \infty} \frac{1}{n} \left\{ \mathbb{E}_{\underline{X} = (X_1 \dots X_n) \sim p_x^n} \left[|E_n(\underline{x})| \right] \right\}$$

- Prefix-Free : $\forall\, \underline{x}, \underline{y}$ $E(\underline{x})$ not prefix of $E(\underline{y})$.

-

"Can we compute $H(X)$ given P_X ?" (in finite time)

(- studied in Inf. Theory as "Single-letter Characterization")

Yes ... as we will see ...

$\overline{\qquad\qquad\qquad X \qquad\qquad}$

~~to~~ Suppose $X \sim \text{Bern}(p)$ (think p small)

$$\left\{ \begin{aligned} X &= 0 \quad \text{w.p. } 1-p \\ &= 1 \quad \text{w.p. } p \end{aligned} \right]$$

Potential Compression of $X_1 \dots X_n$

$$E(X_1 \dots X_n) = (k ; i)$$

$k = \sum X_i$; i = index of $X_1 \dots X_n$ among $\binom{n}{k}$ strings of length n with k ones.

Length of compression $= \underbrace{\log n}_{\text{to convey } k} + \log \binom{n}{k}$

But how large is k?

"Chernoff Bounds" : $E[X_i] = p$ $\quad E[\Sigma X_i] = np$

$$\Pr\left[\, \left| \Sigma x_i - np \right| \geq \lambda \cdot \sqrt{n} \,\right] \leq 2 \cdot e^{-\lambda^2/2}$$

- So, ... ~~w.p.~~ with very high prob. $k \approx pn$

$\Big[\, k \geq (p-\epsilon)n$
$k \leq (p+\epsilon)n$
w.p. $1 - 2^{-\epsilon^2 n}$ $\Big]$

- $\log \binom{n}{pn} \cancel{=?} \approx 2^{h(p) \cdot n}$

$$h(p) = -p \log_2 p - (1-p)\log(1-p)$$

Exercise : Prove this using Stirling's approx.

$$n! \approx \frac{1}{\sqrt{2\pi n}} \left(\frac{n}{e}\right)^n$$

<u>So conclusion</u> :

$\forall \epsilon$ for sufficiently large n

$$H(x) \leq \frac{1}{n}\left[\, O(\log n) + (h(p)+\epsilon) \cdot n \,\right]$$

$\Rightarrow \quad H(x) \leq h(p)$. $\quad$ [taking limits on ϵ & n].

<u>Theorem</u>: $H(X) = h(p)$. [for $X = \text{Bern}(p)$]

<u>Proof</u>: Already seen $H(X) \leq h(p)$.

<u>Converse</u>: (1) Whp $k \in ((p-\epsilon)n, (p+\epsilon)n) \approx pn$

(2) Even if k known to receiver

need to distinguish between $\binom{n}{pn}$ possibilities

~~By~~ "~~[illegible]~~" So strings of length $\leq \log\binom{n}{pn} - t$

only occur w.p. $\leq \cdot 2^{-t}$

Conclude that w.p. $\geq (1-2^{-t})$

Encoding length $\geq \log\binom{n}{pn} - t$ in

any valid encoding.

$$\Rightarrow H(X) \geq \lim_{n\to\infty} \frac{h(p)\cdot n - \epsilon n - t}{n}$$

$$\geq h(p).$$

What about non-Bernouli X?

- $X \in \{1 \dots k\} = \Omega$.

$P_1 \dots P_k$ ~~denote~~ $p_i \triangleq \Pr[X = i]$

- typical string $\underline{X} \in \Omega^n$ has $p_1 n$ 1's
$p_2 n$ 2's
$\vdots$
$p_n n$ k's.

- All strings with these #'s 1's, .. k's are equally likely.

$\Rightarrow$ Compression length

$$\approx \log \binom{n}{p_1 n \ p_2 n \ \dots \ p_k n} \pm o(n).$$

$$\approx \cancel{2^{h(P_x) \cdot n}} \ \cancel{\pm} \ h(P_x) \cdot n \pm o(n)$$

$$h(P_x) \triangleq \sum_{i=1}^{k} p_i \log \frac{1}{p_i}.$$

<u>Theorem</u>: $H(X) = \sum_{\omega \in \Omega} P_x(\omega) \cdot \log \frac{1}{P_x(\omega)}$

$Y \in \Omega_Y \quad X \in \Omega_X$

Conditional Entropy

$$H(Y|X) = \mathop{E}_{\omega \sim p_X}\left[H(Y|X=\omega)\right] = \sum_{\omega \in \Omega_X} P_X(\omega) \cdot H(Y|X=\omega)$$

$$= \sum_{\omega \in \Omega_X} \sum_{\gamma \in \Omega_Y} P_X(\omega) \cdot P_{Y|X=\omega}(\gamma) \cdot \log \frac{1}{P_{Y|X=\omega}(\gamma)}$$

$$= \sum_{(\omega,\gamma)} P_{XY}(\omega,\gamma) \cdot \log \frac{P_X(\omega)}{P_{XY}(\omega,\gamma)}$$

Exercise: Verify $H(X,Y) = H(X) + H(Y|X)$.

Exercise: Show that $\underline{X}^* \in \Omega^n$ can be always compressed ~~deterministically~~ to length $\leq H(X) \cdot n + \epsilon n$ so that decompression error $\leq \exp(-f(\epsilon, k) \cdot n)$

where $f(\epsilon, k) > 0$ for every $\epsilon > 0$ & k.

(more on this next.)

- So far we have talked about expected length of compression. Can we get "always"?

  [must allow error - why?]

  [what if we allow error $\to 0$].

- Asymptotic Equipartition Principle:

  - $\forall$ distribution P_X on finite Ω,
  - $\forall \epsilon$ for sufficiently large n

    $\exists\ S \subseteq \Omega^n$ s.t.

    - $\forall\ \underline{\omega} \in S$: $\Pr[X = \omega] \in \underbrace{\left[\frac{1}{|S|^{1+\epsilon}}, \frac{1}{|S|^{1-\epsilon}}\right]}_{\text{nearly uniform on } S}$

    - $\Pr[X \notin S] \le \epsilon$.

  [~~to~~ every distribution looks like uniform]

- Exercise: Prove $|S| \approx 2^{H(X)\cdot n}$
- Exercise: Derive AEP
- derive: "always with small error" compression from above $\boxtimes$