Shannon entropy & Data Compression

§5.3, §5.1

Last month: QIT formalism. Today: Information Theory proper!

Shannon entropy of $p \in P(\Sigma)$:

$$\boxed{H(p) = -\sum_x p(x) \log p(x)}$$

BASE 2

$0 \cdot \log 0 = 0$

Why do we care? A classical tale... Alice acquired a biased coin:

$$\text{ALICE} \xrightarrow{\text{How many bits?}} \text{BOB}$$

H $p = 75\%$

T $1-p = 25\%$

Clearly: 1 bit (otherwise 25% error)

What if n coin flips? Can we do better than $\boxed{1 \frac{\text{bit}}{\text{coin flip}}}$?

Compression rate

* Consider random seq. $\underbrace{\text{HTTHHTHHTH}}_{k \text{ heads}}$. WHP: $\boxed{\frac{k}{n} \approx p}$?

Isn't H...H more likely? Yes, but...

Law of large numbers implies: $\forall \varepsilon > 0$

$$\Pr\left(\left|\frac{k}{n} - p\right| > \varepsilon\right) \longrightarrow 0 \text{ as } n \to \infty$$

Law of large numbers: $X_1, \dots, X_n$ i.i.d., $V(X_i) < \infty$, $\varepsilon > 0$

$$\Pr\left(\left|\frac{X_1 + \dots + X_n}{n} - \mathbb{E}[X_i]\right| > \varepsilon\right) = O\left(\frac{1}{n}\right) \longrightarrow 0$$

$\Pr(X_i = 1) = p$

$\Pr(X_i = 0) = 1-p$

$\Longrightarrow \sum_i X_i = \#\text{heads}$

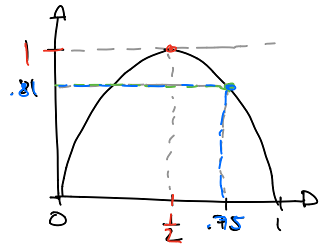
* NB: Also good method to estimate p!
* How many seq. with k heads? $\binom{n}{k}$

Asymptotics?

$$1 = (x + (1-x))^n = \sum_{l=0}^{n} \binom{n}{l} x^l (1-x)^{n-l} \geq \binom{n}{k} x^k (1-x)^{n-k}$$

$$\Longrightarrow \binom{n}{k} \leq x^{-k} (1-x)^{-(n-k)} \overset{x = \frac{k}{n}}{=} \left(\frac{k}{n}\right)^{-k} \left(\frac{n-k}{n}\right)^{-(n-k)} = 2^{n h\left(\frac{k}{n}\right)}$$

With binary Shannon entropy



$$h(p) := H(\{p, 1-p\}) = -p \cdot \log p - (1-p) \cdot \log(1-p)$$

* If $\left|\frac{k}{n} - p\right| < \varepsilon$:

$$\binom{n}{k} = 2^{n \cdot h(p \pm \varepsilon)} \leq 2^{n(\boxed{h(p)+\varepsilon'})}$$

bits needed

e.g. $p = 75\%$: $h(p) \approx 81\%$

$p = 50\%$: $h(p) = 100\%$

Compression protocol: Fix $\varepsilon > 0$.

1. If $\left|\frac{k}{n} - p\right| > \varepsilon$: FAIL (send over arbitrary string)
2. Send $k \in \{0, \ldots, n\}$ to Bob ← $\lceil \log(n+1) \rceil$ bits
3. Send index in list of all coin flip sequences w/ k heads. ← $\lceil n(h(p)+\varepsilon') \rceil$ bits

Analysis: * $\Pr(\text{FAIL}) \to 0$ as $n \to \infty$

* Rate: $R = \frac{\#\text{bits}}{\#\text{coin flips}} = \underbrace{\frac{\log(n+1)+1}{n}}_{\to 0} + h(p) + \boxed{\varepsilon'} + \underbrace{\cancel{\frac{1}{n}}}_{\to 0}$ ☺

as small as we like

* entropy = optimal asymptotic compression rate (for binary source)
* instead of failing, can also send uncompressed string ("lossless" vs. "lossy" compression) ⇝ $E[\text{length}] \leq R + \Pr(\text{FAIL})$

EX CLASS

Rest of today: Generalize to arbitrary alphabets Σ. Properties of entropy.

Let $p \in \mathcal{P}(\Sigma)$. A general compression scheme looks as follows:

(n, R, δ)-code for p: $\underbrace{\mathcal{E}: \Sigma^n \to \{0,1\}^{\lfloor nR \rfloor}}_{\text{encoder}}$, $\underbrace{\mathcal{D}: \{0,1\}^{\lfloor nR \rfloor} \to \Sigma^n}_{\text{decoder}}$

s.th. $\underbrace{\sum_{x \in \Sigma^n, \mathcal{D}(\mathcal{E}(x)) = x} p(x_1) \cdots p(x_n)}_{\text{prob. of success}} \geq 1 - \delta$

Thm (Shannon source coding): Let $\delta \in (0,1)$.

(A) If $R > H(p)$ then $\exists n_0: \forall n \geq n_0: \exists (n, R, \delta)$-code

(B) If $R < H(p)$ then $\exists n_0: \forall n \geq n_0: \nexists (n, R, \delta)$-code

entropy = "optimal" rate

Ⓐ "achievability", Ⓑ "converse" $\to$ $\boxed{\text{HW}}$

$$\boxed{\begin{array}{l} x \in \Sigma^n \text{ } \varepsilon\text{-typical for } p: \quad 2^{-n(H(p)+\varepsilon)} \leq p(x_1)\cdots p(x_n) \leq 2^{-n(H(p)-\varepsilon)} \\ T_{n,\varepsilon}(p) = \{x \in \Sigma^n \text{ } \varepsilon\text{-typical for } p\} \end{array}}$$

① $|T_{n,\varepsilon}| \leq 2^{n(H(p)+\varepsilon)}$

Could also look at frequencies, i.e. $\frac{\#\{k : x_k = x\}}{n} \approx p(x)$

Pf: $1 \geq \sum_{x \in T_{n,\varepsilon}} p(x_1)\cdots p(x_n) \geq |T_{n,\varepsilon}| \cdot 2^{-n(H(p)+\varepsilon)}$. □

② $\sum_{x \in T_{n,\varepsilon}} p(x_1)\cdots p(x_n) \longrightarrow 1$ as $n \to \infty$

Pf: Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p$ and $L_k = \begin{cases} -\log p(X_k) & \text{if } p(X_k) > 0 \\ 0 & \text{if } p(X_k) = 0. \end{cases}$

$$\mathbb{E}[L_k] = \sum_x p(x)(-\log p(x)) = H(p)$$

$$\Rightarrow \sum_{x \in T_{n,\varepsilon}} p(x_1)\cdots p(x_n) = \Pr(X \in T_{n,\varepsilon}) = \Pr\left(\left|\frac{L_1+\ldots+L_n}{n} - H(p)\right| > \varepsilon\right) \xrightarrow{\text{LLN}} 0$$

□

Proof of Shannon's thm, part Ⓐ: Choose $\varepsilon = \frac{R - H(p)}{2} > 0$. Then:

$$n(H(p)+\varepsilon) = n(R-\varepsilon) \leq \lfloor nR \rfloor \text{ if } n \geq \frac{1}{\varepsilon}$$

① $\leadsto \exists$ injective map $\mathcal{E}_n : T_{n,\varepsilon} \to \{0,1\}^{\lfloor nR \rfloor}$ w/ left inverse $\mathcal{D}_n$

Extend $\mathcal{E}_n$ arbitrarily to Σ^n. Then:

$$\sum_{x : \mathcal{D}_n(\mathcal{E}_n(x)) = x} p(x_1)\cdots p(x_n) \geq \sum_{x \in T_{n,\varepsilon}} p(x_1)\cdots p(x_n) \xrightarrow{②} 1 \text{ as } n \to \infty.$$

$\leadsto \geq 1 - \delta$ for n sufficiently large.

□

Properties of Shannon entropy

Shannon entropy: $H(p) = -\sum_{x\in\Sigma} p(x) \log p(x)$ for $p \in \mathcal{P}(\Sigma)$

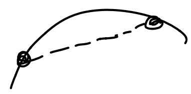
* $0 \le H(p) \le \log|\Sigma|$, $= 0$ iff deterministic (all but one $p_x = 0$)

  $= \log|\Sigma|$ iff uniform

  apply Jensen to $\sum_x p(x) \log \frac{1}{p(x)}$

Jensen's inequality: $p \in \mathcal{P}(\Sigma)$, $a \in \mathbb{R}^\Sigma$, f concave

$$\sum_x p(x) f(a(x)) \le f\Big(\sum_x p(x) a(x)\Big)$$



* Concave in p

  $\forall p, q \in \mathcal{P}(\Sigma)$, $\lambda \in [0,1]$:

  $\lambda H(p) + (1-\lambda) H(q) \le H(\lambda p + (1-\lambda) q)$ $\leftarrow$ follows from concavity of $f(q) = -q \cdot \log(q)$ on $[0,\infty)$

* optimal rate for compression

Entropies of subsystems:

$p_{XY} \in P(\Sigma_X \times \Sigma_Y)$ $\rightsquigarrow$ $p_X \in P(\Sigma_X)$ marginal distribution

$H(XY) = H(p_{XY})$ $\qquad$ $H(X) = H(p_X)$

* **Subadditivity:** $H(XY) \le H(X) + H(Y)$

  Pf: Can compress at rate $H(X) + H(Y) + \varepsilon$, but not nec. optimal $\square$

* **Monotonicity:** $H(XY) \ge H(X)$ $\lightning$ WRONG FOR Q. STATES

  Pf: Given $X_1, \dots, X_n$, generate $Y_k \sim p_{Y|X=X_k} = \frac{p_{XY}(X_k, \cdot)}{p_X(X_k)}$

  $\Rightarrow (X_k, Y_k) \overset{\text{iid}}{\sim} p_{XY} \Rightarrow$ Can compress at rate $H(XY)$. $\square$