① Shannon info / entropy

- should I email you that class is canceled?
- $X = 1$ class canceled
  $= 0$ not canceled
- $P(X=1) = \frac{1}{1024}$ (rarely canceled)
- what is the "information content" of that event?
- $h(X=1) = \log_2 \frac{1}{P(X=1)} = 10$ bits (information)
- $h(X=0) = \log_2 \frac{1}{P(X=0)} \approx 0.0015$ bits
- Not much info in $X=0$, so only email if $X=1$
- information = "surprise"
- should you email me to ask if class is canceled?
- How much info do you expect to gain?
- $$\mathbb{E}_{X \sim P(X)}[h(X)] = \sum_X P(X) \log_2 \frac{1}{P(X)} \equiv H(X) \quad \text{(entropy)}$$
- in this case, the binary entropy function
  $$H_2(p) \triangleq p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p}$$
  $$\equiv H(X), \text{ for } X \sim \text{Bernoulli}(p).$$
- $H_2(\frac{1}{1024}) = \frac{1}{1024} \times 10 + \frac{1023}{1024} \times 0.0015 \approx 0.01$ bits
- so, not much expected info from asking.
- $\frac{\partial}{\partial p} H_2(p) = 0 \longleftrightarrow p = 1-p = 0.5$
- maximum expected info = maximum uncertainty
- entropy = "uncertainty"
- $H_2(0.5) = 1$ bit ; binary variable is worth at most 1 bit
- more generally, $X \in \{1 \dots K\}$, $p = p_1 \cdots p_K$
  $$H(X) \equiv H(p) = \sum_k p_k \log_2 \frac{1}{p_k}$$
  $$\max_p H(p) = \log_2 K \text{ bits} = \text{"raw bit content of } X\text{"} \equiv H_0(X)$$

② "Game of Submarine" (Mackay Chap 4.1)

- sub is somewhere in 64-cell grid
- $P(Z=k) = \frac{1}{64}$
- $P(\text{guess correctly on 1st try}) \equiv P(X_1=1) = \frac{1}{64}$
- info gained : $h(X_1=1) = h(Z=k) = \log_2 64 = 6$ bits
- if we miss : $h(X_1=0) = h(Z\neq k) = \log_2 \frac{64}{63} \approx 0.023$ bits
- $P(\text{guess correctly on 2nd try}) \equiv P(X_2=1 \mid X_0=1) = \frac{1}{63}$
- total info gained after 32 misses

$$h(X_1=0 \ldots X_{32}=0) = \log_2 \frac{64}{\cancel{63}} + \log_2 \frac{\cancel{63}}{\cancel{62}} + \ldots + \log \frac{\cancel{33}}{32}$$

$$= \log_2 \frac{64}{32} = \log_2 2 = 1 \text{ bit}$$

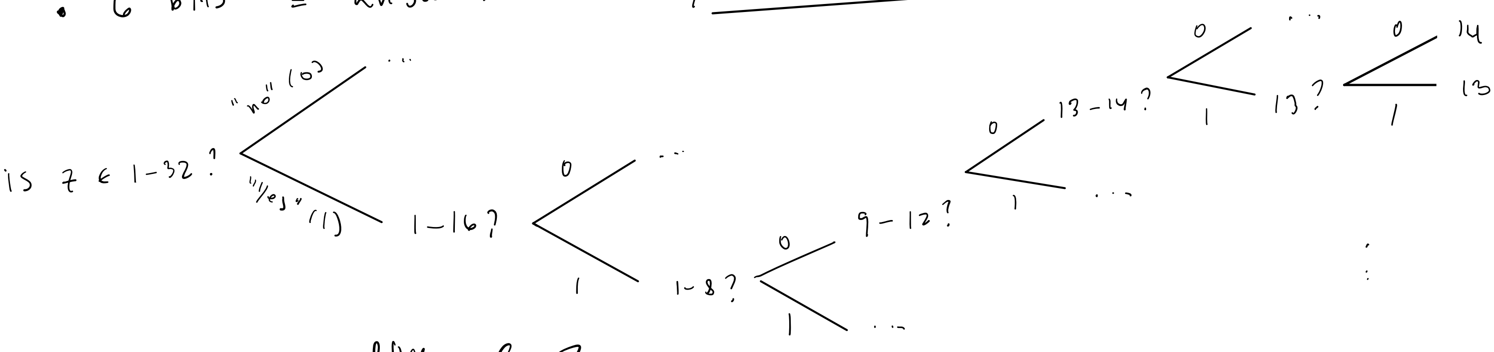
- total info gained after hit on 33rd attempt

$$h(X_1=0, \ldots X_{32}=0, X_{33}=1) = 1 + \log_2 32 = 6 \text{ bits}$$

- what about hitting after 48 misses?

$$= \log_2 \frac{64}{\cancel{32}} + \ldots + \log_2 \frac{\cancel{17}}{16} + \log_2 16 = 6 \text{ bits}$$

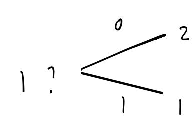
- Z is always worth 6 bits. why?
- 6 bits = answers to 6 yes-no questions



- Binary encoding of Z

Z	code
1	111 111
⋮	...
13	110 011
14	110 010
⋮	...
64	000 000

← 6-bit fixed length codes



(3) Lossy compression of fixed length codes

- Should not be surprising that Z can be encoded into 6 bits
- $H_0(Z) = \log_2 64 = 6$ bits
- in this case $H(Z) = H_0(Z)$ (max. uncertainty)
- can we do better if Z is more predictable? $\rightarrow H(Z) < H_0(Z)$
- Historical context: Shannon, Bell Labs, WW2
- Encoding / decoding messages x
- "Alphabet" $x \in A_x$
- Error tolerance δ
- say $A_x = \{1 \dots 64\}$ and $P\left(x \in \underbrace{\{1 \dots 32\}}_{\doteq S_\delta}\right) \geq 1 - \delta$
- Then only code for S_δ:

x	code	
1	1111	4-bit codes
2	1110	
⋮	⋮	
32	0000	
33	$\emptyset$	no code available
⋮	⋮	
64	$\emptyset$	

- raw bit content: $H_\delta(X) < H_0(X)$ compression
  $\log_2 32 < \log_2 64$
- error prob $\delta \rightarrow$ lossy compression
- Can we do better? yes!
- send / encode blocks $X^N \equiv (X_1 \dots X_N)$
- say $A_x = \{1, 2, \dots, 26\}$, $P(X = k) = P_k$
- $X_i \overset{iid}{\sim} P(x) \rightarrow H(X_1 \cdots X_N) = N H(X)$ (derive this!)
  $\hookrightarrow$ this additivity is why Shannon defined info as $\log_2 \frac{1}{P(x)}$
- X^N can be encoded in $N H_0(X)$ bits with no loss
- Source coding theorem:

  X^N can be encoded in $N H(X)$ bits with negligible loss as $N \rightarrow \infty$

$P(x)_{typical} = \prod_i P(x_i) \simeq \prod_{k \in A} P_k^{(N p_k)}$

$h(x) = \log_2 \text{''} \quad \text{''} \simeq \sum_k N p_k \log_2 p_k$

$= N H(p)$

- Define: $N_k \triangleq \sum_{i=1} \mathbb{1}(x_i = k)$
- In a typical sequence x^N, there will be

$$N_k \simeq N p_k$$

- $P(x^N) \simeq \prod_{k \in \mathcal{A}_x} p_k^{(N p_k)}$ (typically)
- $h(x^N) \simeq N \sum_k p_k \log_2 \frac{1}{p_k} = \underline{N H(x)}$
- info content of a typical x^N is $\simeq N H(x)$
- x^N is $\underline{\varepsilon\text{-typical}}$ if $\left| \frac{1}{N} \sum_{i=1}^{N} h(x_i) - H(x) \right| < \varepsilon$
- typical set:

$$\mathcal{A}_{\varepsilon, N} \triangleq \left\{ x^N : x^N \text{ is } \varepsilon\text{-typical} \right\}$$

- How likely is $x^N \in \mathcal{A}_{\varepsilon,N}$? $P(x^N \in \mathcal{A}_{\varepsilon,N})$?
- $R_i \triangleq h(x_i) = \log_2 \frac{1}{p(x_i)}$
- $R_i \overset{iid}{\sim} P(R)$, $i = 1 \dots n$
- $\mathbb{E}[R] = \mathbb{E}_{x \sim p(x)}[h(x)] = H(x)$
- $\lim_{N \to \infty} P\left(\left| \frac{1}{N} \sum_{i=1}^{N} R_i - \mathbb{E}[R] \right| < \varepsilon \right) \to 1$

(Law of Large Numbers)

- "x^N is almost surely in the typical set for large N
- How big is $\mathcal{A}_{\varepsilon,N}$?
- x^N is ε-typical if:

$$H(x) - \varepsilon \leq \frac{1}{N} \log_2 \frac{1}{p(x^N)} \leq H(x) + \varepsilon$$

$$\hookrightarrow 2^{-N(H(x)+\varepsilon)} \leq p(x^N) \leq 2^{-N(H(x)-\varepsilon)}$$

- So, if all $x^N \sim P(x^N)$ are in $A_{\epsilon,N}$ and $P(x^N) \simeq 2^{-NH(X)}$

  then $|A_{\epsilon,N}| \simeq 2^{NH(X)}$

- this is the Asymptotic equipartition principle (AEP)

- this is much smaller than $2^{NH_0(X)} = |A_X|^N$

  (raw bit content)

- e.g.: $|A_X| = 2$ (binary), $P(X=1) = 0.4$

- $$\frac{|A_{\epsilon,N}|}{|A_X|^N} \simeq 2^{N(H(X) - H_0(X))} = 2^{N(0.97 - 1.0)} = 2^{-0.03N}$$

- So only 2^{-30} fraction of $N = 3000$ sequences are typical

- conclusion: a fixed-length code for x^N can be

  $\log_2 |A_{\epsilon,N}| = NH(X)$ bits with negligible loss of info

(4) Lossless variable-length symbol codes

- $\mathcal{A}_X = \{a, b, c, d\}$
- symbol code c_i for $i \in \mathcal{A}_X$, e.g. $c_a = 11$

i	c_i	l_i	p_i
a	11	2	1/4
b	10	2	1/4
c	00	2	1/4
d	01	2	1/4

$abacd \rightarrow \underbrace{11}_{a}\underbrace{10}_{b}\underbrace{11}_{a}\underbrace{00}_{c}\underbrace{01}_{d}$

- must be prefix codes to be uniquely decodable
- $\mathbb{E}[l] = \sum_i p_i l_i$ "expected code length"
- for unequal probabilities we can do better, e.g.

i	c_i	l_i	p_i
a	0	1	1/2
b	100	3	1/6
c	101	3	1/6
d	111	3	1/6

$\mathbb{E}[l] = \frac{1}{2} + \frac{1}{6} \cdot 3 \cdot 3 = \frac{1}{2} + \frac{9}{6} = \frac{1}{2}\frac{3}{2} = \frac{3}{4}$ bits

(vs. 2 bits)

- Theorem: $\mathbb{E}[l] \geq H(X)$

$\mathbb{E}[l] = H(X)$ only if $l_i = \log_2 1/p_i$

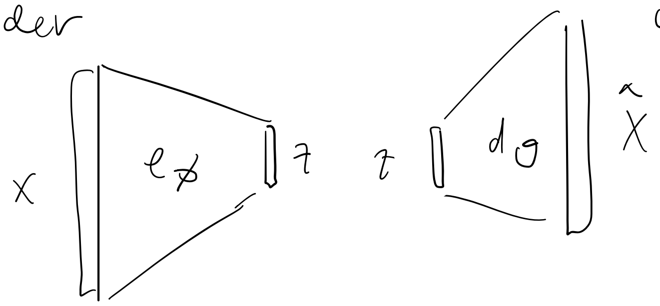
- so again: $NH(X)$ bits (on average) to encode X^N

5 Learning / modeling as compression

- How do we learn an optimal coding scheme?
- Source coding theorem only says one exists
- Example: autoencoders

$e_\phi(x) = z$ encoder

$d_\theta(z) = \hat{x}$ decoder



- $$\hat{\phi}, \hat{\theta} \leftarrow \underset{\phi, \theta}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, d_\theta(e_\phi(x_i)))$$

  for some loss $\ell(\cdots)$

- "Bits back" and VAEs (save for later)

⑥ Info, entropy, etc. with dependent variables

- Is the sub in this search cell?

$$z=1 \quad \text{yes} \quad \pi(z=1) = \pi$$
$$z=0 \quad \text{no} \quad \pi(z=0) = 1-\pi$$

- We can take magnetometer measurements X
- should we? Do we expect to gain info about z?
- Assume we know

$$P(x \mid z=1) \quad \text{vs.} \quad P(x \mid z=0)$$

- e.g.

- $$\pi^* \equiv \pi^*(z=1 \mid X=x) = \frac{\pi\, P(x \mid z=1)}{\pi P(x \mid z=1) + (1-\pi) P(x \mid z=0)}$$
- Did x reduce our uncertainty?

$$H_2(\pi) \equiv H(Z) \quad \text{vs.} \quad H_2(\pi^*) \equiv H(Z \mid X=x)$$

- Not necessarily... e.g., if $\pi = 0.9$, $\pi^* = 0.8$
- uncertainty about z might increase for a given x

- what about in expectation?

the random variable X, not a value $X=x$

$$\mathbb{E}_{x \sim p(x)} \left[H(Z \mid X = x) \right] \equiv H(Z \mid X)$$

"conditional entropy"

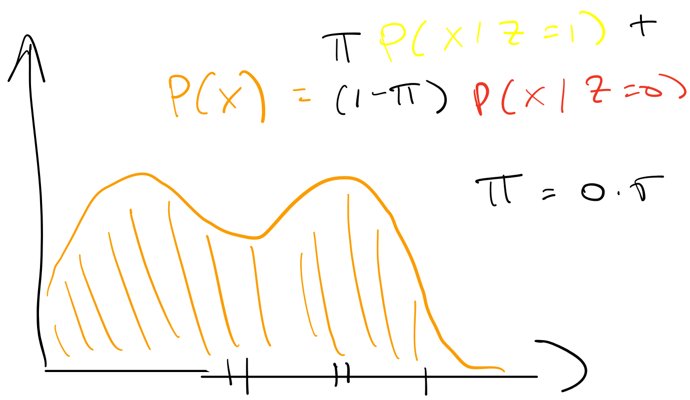
- Fact: $H(Z \mid X) \leq H(Z)$
- You can show this using Jensen's inequality

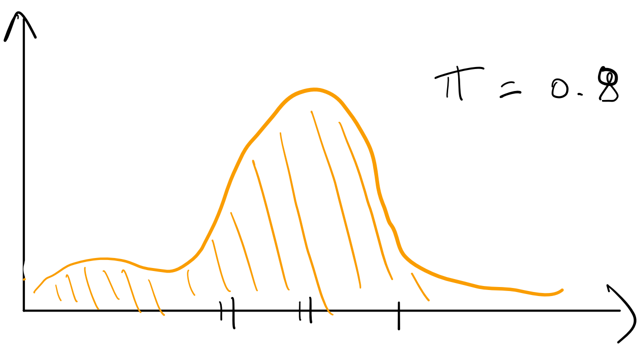
$$\mathbb{E}_X \left[H(p(Z \mid X)) \right] \leq H\left(\mathbb{E}_X \left[p(Z \mid X) \right] \right)$$

- So conditioning always reduces uncertainty
- But by how much?
- $H(Z) - H(Z \mid X)$

$= I(Z; X)$ "mutual information"

$= H(X) - H(X \mid Z)$ (symmetric)





$$= H(X) - \left[\pi H(X \mid Z=1) + (1-\pi) H(X \mid Z=0) \right]$$

- Mutual info is high when
  - $H(Z)$ or $H(X)$ is high
  - $H(X \mid Z)$ or $H(Z \mid X)$ is small at likely values of Z or X
- Another view:

$$I(X;Z) = \sum_x \sum_z p(x,z) \log_2 \frac{p(x,z)}{p(x)\,p(z)}$$

$$= KL\left(p(x,z) \,\|\, p(x)\,p(z) \right)$$

$$= 0 \quad \text{iff} \quad X \perp\!\!\!\perp Z$$

- "if $X \perp\!\!\!\perp Z$, then X does not reduce uncertainty about Z"
- Yet another view:

$$I(X;Z) = H(X) + H(Z) - H(X,Y)$$

- compare to covariance:

$$\mathrm{cov}(X,Z) = \mathbb{E}[XY] - \mathbb{E}[X] - \mathbb{E}[Y]$$

(7) Which cell to search?

- Sub $z \in \{1 \dots k\}$, $\pi(z=k) = \pi_k$
- Search cell k and find it : $Y_k = 1$

$$p(Y_k = 1 \mid z = k) = 1$$
$$p(Y_k = 1 \mid z \neq k) = 0$$

- $\arg\max_k \; H(z) - H(z \mid Y_k)$

$$= \arg\max_k \; H(Y_k) - \underbrace{H(Y_k \mid z)}_{=0}$$

- $p(Y_k = 1) = \sum_{j=1}^{k} \pi_j \underbrace{p(Y_k = 1 \mid z = j)}_{\substack{=0 \text{ if } j \neq k \\ 1 \text{ if } j = k}} = \pi_k$

$$= \arg\max_k \; H_2(\pi_k)$$
$$= \arg\max_k \; \pi_k .$$

(8) SEPs

- $P(Y_k = 1 \mid z = k) = q_k \leq 1$

$$P(Y_k = 1 \mid z \neq k) = 1$$

- $p(Y_k = 1) = \pi_k q_k$

- $H(Y_k \mid Z) = \pi_k H_2(q_k)$

- $\underset{k}{\arg\max}\ H_2(\pi_k q_k) - \pi_k H_2(q_k)$

- max <u>possible</u> $I(Z; Y_k)$ is $\pi_k = 0.5$, $q_k = 1.0$

- e.g. $\pi_k = \frac{1}{8}$, $q_k = 1 \rightarrow H_2(\frac{1}{8}) - 0 \approx \underline{0.5436}$

  vs. $\pi_{k'} = \frac{1}{2}$, $q_{k'} = \frac{1}{2} \rightarrow H_2(\frac{1}{4}) - \frac{1}{2} H_2(\frac{1}{2}) \approx \underline{0.311}$

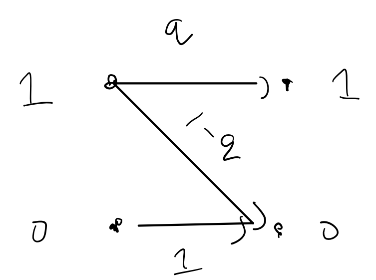
↳ more expected information from $\pi_k = \frac{1}{8}$.

- <u>Puzzle</u>: $q_k = q_{k'} = 0.86$ (equal!)

  $\pi_k = 0.455 < \pi_{k'} = 0.5$

  $I(Z; Y_k) \approx 0.685 \overset{!!}{>} I(Z; Y_{k'}) \approx 0.678$

- <u>Hint</u>: Mackay example 9.11, "Z-channels"



9 "Bits back"

- optimal encoding of X is in $\log_2 \frac{1}{p(X)}$ bits
- what if $P(X)$ is intractable?
- However $p(X \mid Z)$ is $\quad P(X) = \sum_Z P(X, Z)$
- Transmit X and Z
- code X according to $P(X \mid Z)$, Z according to $p(Z)$
- Which Z? $\underset{Z}{\operatorname{argmax}}\, p(Z \mid X)$?
- That would minimize the code length of x but not Z
- Instead: sample $Z \sim P(Z \mid X)$

naive cost: $\log_2 \frac{1}{p(X \mid Z)} + \log_2 \frac{1}{p(Z)}$

- However! The receiver can decode the random bits and gain them back:

$$\log_2 \frac{1}{p(X \mid Z)} + \log_2 \frac{1}{p(Z)} - \underbrace{\log \frac{1}{p(Z \mid X)}}_{\text{"bits back"}}$$

$$= \log_2 \frac{1}{p(X)} \leftarrow \text{optimal rate!}$$

- This is the "bits back" argument

(10) Minimum Description Length (MDL)

- "Bits back" motivates latent variable models
- But which model?
- $P(X ; \mathcal{M}_1)$ vs. $P(X ; \mathcal{M}_2)$
- MDL principle: minimize the joint length of the data and the model

$$\min_{\mathcal{M}} \underbrace{L(X ; \mathcal{M})}_{= h(x; \mathcal{M})} + L(\mathcal{M})$$

- This can be understood as defining a prior over models:

$$p(m) = 2^{-L(m}$$

and doing MAP:

$$\hookrightarrow \max_{m} \log P(x ; M) + p(M)$$

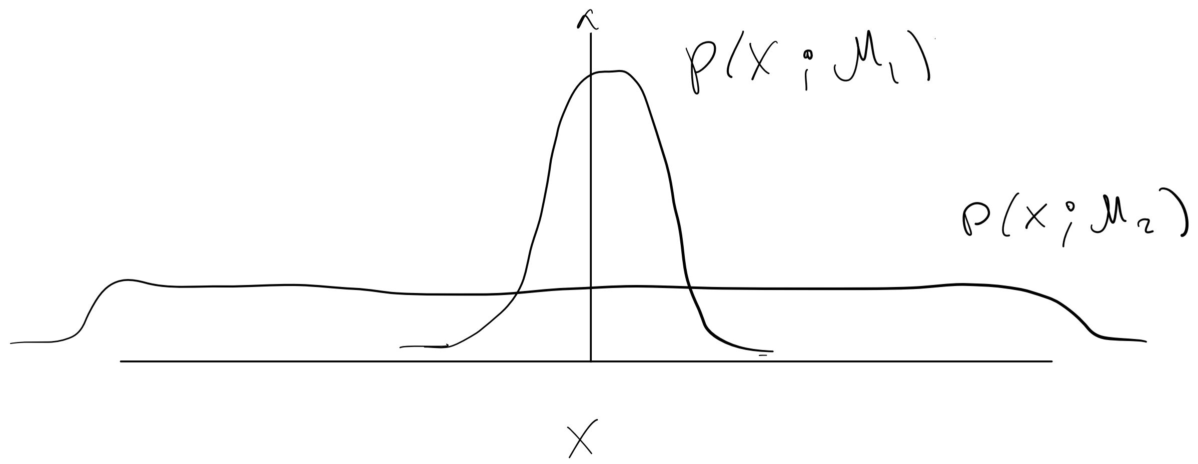
- for two models, the Bayes factor:

$$\frac{P(\mathcal{M}_1 \mid X)}{P(\mathcal{M}_2 \mid X)} = \frac{P(x ; \mathcal{M}_1)}{P(x ; \mathcal{M}_2)} \frac{P(\mathcal{M}_1)}{P(\mathcal{M}_2)}$$

evidence

(11) Occam's Razor and model evidence

- MacKay argues that model evidence alone captures the principle of Occam's Razor.
- Say M_1 is simpler than M_2
- M_2 can fit to more data but therefore spreads probability across many possible X:



(12) Recent updates to this perspective

- Fong and Holmes (2020):

  "Leave p-out" CV score:

$$S_{cv}(X_{1:n}; p) = \frac{1}{\binom{n}{p}} \sum_{t=1}^{\binom{n}{p}} \frac{1}{p} \sum_{j=1}^{p} s(\tilde{x}_j^t \mid x_{1:n-p}^t)$$

Theorem:

$$\log P(X_{1:n}) = \sum_{p=1}^{n} S_{cv}(X_{1:n}; p)$$

$$s(\tilde{x}_j^t \mid x_{1:n-p}^t) = \log P(\tilde{x}_j \mid x_{1:n-p}; M)$$