6.1 Entropy

For example, suppose we observe a sequence of symbols $X_n \sim p$ generated from distribution p. If p has **high entropy**, it will be **hard to predict** the value of each osbervation X_n. Hence we say that the dataset $\mathcal{D} = (X_1, \ldots, X_n)$ has **high information** content.

Entropy for discrete random variables

$$\mathbb{H}(X) = -\sum_{k=1}^{K} p(X=k) \log_2 p(X=k) = -\mathbb{E}_X[\log p(X)]$$

Usually we use log base 2, in which case the units are called **bits**. The discrete distribution with **maximum entropy** is the **uniform distribution**. Hence for a $K-$ary random variable, the entropy is maximized if $p(x=k) = 1/K$;

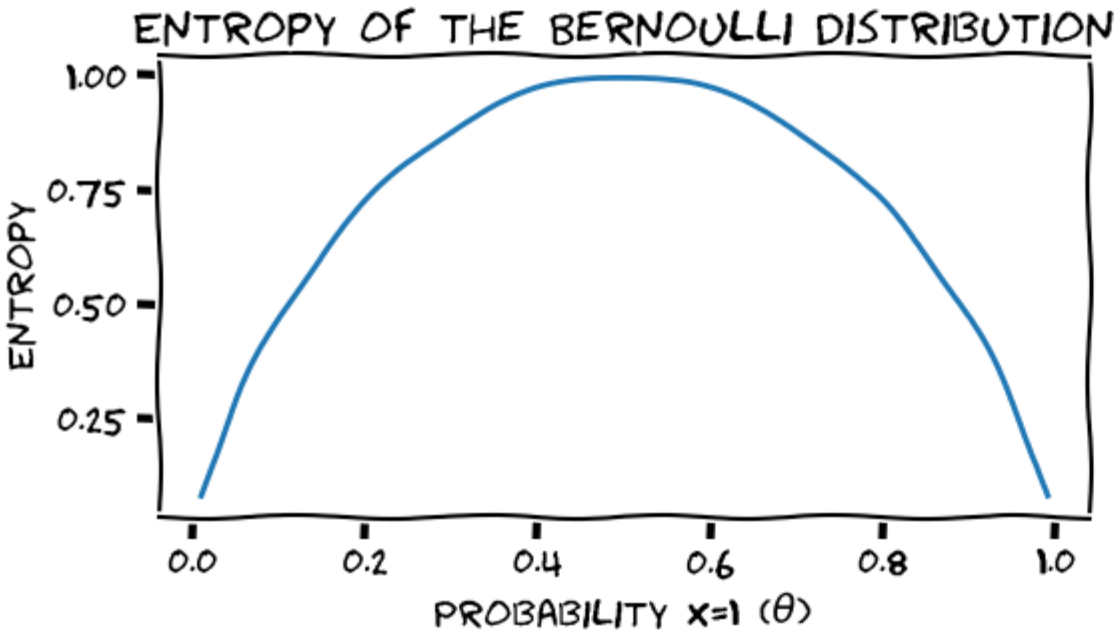
$$\mathbb{H}(X) = -\sum_{k=1}^{K} \frac{1}{K} \log(1/K) = -\log(1/K) = \log(K)$$

Conversely, the distribution with minimum entropy (which is zero) is any delta-function that puts all its mass on one state.

Example: Bernoulli Distribution

- Probability Mass Function (PMF): $p(X=x) = \theta^x (1-\theta)^{(1-x)}$ where $x \in \{0, 1\}$
- The entropy:

$$\begin{aligned}\mathbb{H}(X) &= -[p(X=1) \log_2 p(X=1) + p(X=0) \log_2 p(X=0)] \\ &= -[\theta \log_2 \theta + (1-\theta) \log_2 (1-\theta)]\end{aligned}$$



Cross entropy

$$\mathbb{H}(p, q) \triangleq -\sum_{k=1}^{K} p_k \log q_k$$

The cross entropy is the expected number of bits needed to compress some data samples drawn from distribution p using a code based on distribution q. This can be minimized by setting $q = p$, in which case the expected number of bits of the optimal code is $\mathbb{H}(p, p) = \mathbb{H}(p)$, this is known as **Shannon's source coding theorem**.

For example, if we know that the letter 'e' is more common than the letter 'z', we would assign a shorter codeword to 'e' and a longer codeword to 'z'. This way, we can compress the data more efficiently.

Joint entropy

The joint entropy of two random variables X and Y is defined as

$$\mathbb{H}(X, Y) = -\sum_{x,y} p(x, y) \log_2 p(x, y)$$

Example: Consider choosing an integer from 1 to 8, $n \in \{1, \ldots, 8\}$. Let $X(n) = 1$ if n is even, and $Y(n) = 1$ if n is prime:

n	1	2	3	4	5	6	7	8
X	0	1	0	1	0	1	0	1
Y	0	1	1	0	1	0	1	0

The joint distribution is

$p(X, Y)$	$Y = 0$	$Y = 1$
$X = 0$	1/8	3/8

$p(X,Y)$	$Y = 0$	$Y = 1$
$X=1$	3/8	1/8

so the joint entropy is given by

$$\mathbb{H}(X,Y) = -\left[\frac{1}{8}\log_2\frac{1}{8} + \frac{3}{8}\log_2\frac{3}{8} + \frac{3}{8}\log_2\frac{3}{8} + \frac{1}{8}\log_2\frac{1}{8}\right] = 1.81 \text{ bits}$$

Clearly the marginal probabilities are uniform:

$p(X=1) = p(X=0) = p(Y=0) = p(Y=1) = 0.5$,

So $\mathbb{H}(X) = \mathbb{H}(Y) = 1$.

Hence $\mathbb{H}(X,Y) = 1.81 \text{ bits } < \mathbb{H}(X) + \mathbb{H}(Y) = 2 \text{ bits}$.

If X and Y are independent, then $\mathbb{H}(X,Y) = \mathbb{H}(X) + \mathbb{H}(Y)$, so the bound is tight.

This makes intuitive sense: when the parts are correlated in some way, it reduces the **"degrees of freedom"** of the system, and hence reduces the overall entropy.

If Y is a deterministic function of X, then $\mathbb{H}(X,Y) = \mathbb{H}(X)$. So

$$\mathbb{H}(X,Y) \geq \max\{\mathbb{H}(X), \mathbb{H}(Y)\} \geq 0$$

Intuitively this says **combining variables** together **does not** make the **entropy go down**: you cannot reduce uncertainty merely by adding **more unknowns** to the problem, *you need to observe some data*.

Conditional Entropy

The conditional entropy of Y given X is the uncertainty we have in Y after seeing X, averaged over possible values for X:

$$\begin{aligned}
\mathbb{H}(Y|X) &\triangleq \mathbb{E}_{p(x)}[\mathbb{H}(p(Y|X))] \\
&= \sum_x p(x)\mathbb{H}(p(Y|X=x)) \\
&= -\sum_x p(x) \sum_y p(y|x) \log p(y|x) \\
&= -\sum_{x,y} p(x,y) \log p(y|x) \\
&= -\sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)} \\
&= -\sum_{x,y} p(x,y) \log p(x,y) - \sum_x p(x) \log \frac{1}{p(x)} \\
&= \mathbb{H}(X,Y) - \mathbb{H}(X)
\end{aligned}$$

- If Y is a deterministic function of X, then knowing X completely determines Y , so $\mathbb{H}(Y|X) = 0$.

- If X and Y are independent, knowing X tells us nothing about Y and $\mathbb{H}(Y|X) = \mathbb{H}(Y)$.
- Since $\mathbb{H}(X,Y) \leq \mathbb{H}(Y) + \mathbb{H}(X)$, we have $\mathbb{H}(Y|X) \leq \mathbb{H}(Y)$
- In general the chain rule for entropy:

$$\mathbb{H}(X_1, X_2) = \mathbb{H}(X_1) + \mathbb{H}(X_2|X_1) \to \mathbb{H}(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} \mathbb{H}(X_i|X_1, \ldots, X$$

6.2 Relative entropy (KL divergence)

In fact, we consider a divergence measure $\mathcal{D}(p,q)$ which quantifies *how far* q is from p, without requiring that $\mathcal{D}$ be a metric.

- More precisely, we say that $\mathcal{D}$ is a divergence if $\mathcal{D}(p,q) \geq 0$ with equality iff $p = q$, whereas a metric also requires that $\mathcal{D}$ be symmetric and satisfy the triangle inequality, $\mathcal{D}(p,r) \leq \mathcal{D}(p,q) + \mathcal{D}(q,r)$.
- We focus on the **Kullback-Leibler divergence** or KL divergence, also known as the **information gain** or **relative entropy**, between two distributions p and q.

For discrete distributions, KL divergence is:

$$D_{\mathrm{KL}}(p||q) \triangleq \sum_{k=1}^{K} p_k \log \frac{p_k}{q_k} = \underbrace{\sum_{k=1}^{K} p_k \log p_k}_{-\mathbb{H}(p)} - \underbrace{\sum_{k=1}^{K} p_k \log q_k}_{\mathbb{H}(p,q)}$$

we can interpret the KL divergence as the *"extra number of bits"* you need to pay when compressing data samples if you use the incorrect distribution q as the basis of your coding scheme compared to the true distribution p.

For continuous distributions, KL divergence is:

$$D_{\mathrm{KL}}(p||q) \triangleq \int dx\ p(x) \log \frac{p(x)}{q(x)}$$

Example: KL divergence between two Gaussians

$$D_{\mathrm{KL}}(\mathcal{N}(x|\mu_1, \Sigma_1)||\mathcal{N}(x|\mu_2, \Sigma_2)) = \frac{1}{2}\Big[\mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)^T \Sigma_2^{-1}(\mu_2 - \mu_1) - D +$$

In the scalar case, this becomes

$$D_{\mathrm{KL}}(\mathcal{N}(x|\mu_1, \sigma_1^2)||\mathcal{N}(x|\mu_2, \sigma_2^2)) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}$$

We use **Jensen's inequality**. This states that, for any **convex** function f, we have that

$$f(\sum_{i=1}^{n} \lambda_i x_i) \leq \sum_{i=1}^{n} \lambda_i f(x_i)$$

where $\lambda_i \geq 0$ and $\sum_{i=1}^{n} \lambda_i = 1$. In words, this result says that f of the average is less than the average of the f's.

To prove for general n, we can use induction. For example, if $f(x) = \log(x)$, which is a concave function, we have

$$\log(\mathbb{E}_x g(x)) \geq \mathbb{E}_x \log(g(x))$$

Theorem(Information inequality) $D_{KL}(p||q) \geq 0$ with equality iff $p = q$.

Proof. Let $A = \{x : p(x) > 0\}$ be the support of $p(x)$. Using the concavity of the log function and Jensen's inequality, we have that

$$\begin{aligned} -D_{KL}(p||q) = -\sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} &= \sum_{x \in A} p(x) \log \frac{q(x)}{p(x)} \\ \leq \log \sum_{x \in A} p(x) \frac{q(x)}{p(x)} &= \log \sum_{x \in A} q(x) \end{aligned}$$

Since $\log(x)$ is a strictly concave function ($-\log(x)$ is convex), in the above equation, we have equality iff $p(x) = c\, q(x)$ for some c that tracks the fraction of the whole space $\mathcal{X}$ contained in A.

$$-D_{KL}(p||q) \leq \log \sum_{x \in \mathcal{X}} q(x) = \log 1 = 0$$

We have equality iff $\sum_{x \in A} q(x) = \sum_{x \in \mathcal{X}} q(x) = 1$, which implies $c = 1$. Hence $D_{KL}(p||q) = 0$ iff $p(x) = q(x)$ for all x.

Corollary (Uniform distribution maximizes the entropy) $\mathbb{H}(X) \leq \log |X|$, where $|X|$ is the number of states for X, with equality iff $p(x)$ is uniform.

Proof. Let $u(x) = 1/|X|$. Then

$$0 \leq D_{KL}(p||u) = \sum_{x} p(x) \log \frac{p(x)}{u(x)} = \log |X| - \mathbb{H}(X)$$

KL divergence and MLE

Suppose we want to find the distribution q that is *as close as* possible to p, as measured by KL divergence:

$$\begin{aligned} q^* &= \underset{q}{\operatorname{argmin}}\, D_{KL}(p||q) && (1) \\ &= \underset{q}{\operatorname{argmin}} \left\{ \int p(x) \log p(x) dx - \int p(x) \log q(x) dx \right\} && (2) \end{aligned}$$

Now suppose p is the empirical distribution:

$$p_D(x) = \frac{1}{N} \sum_{n=1}^{N} \delta(x - x_n)$$

Then

$$\begin{aligned} D_{KL}(p_{\mathcal{D}}||q) &= -\int p_{\mathcal{D}}(x) \log q(x) dx + C && (3) \\ &= -\int \left[\frac{1}{N} \sum_n \delta(x - x_n) \right] \log q(x) dx + C && (4) \\ &= -\frac{1}{N} \sum_n \log q(x_n) + C && (5) \end{aligned}$$

where $C = \int p(x) \log\, p(x)\, dx$ is a constant independent of q.

- This is called the **cross entropy** objective, and is equal to the **average negative log likelihood** of q on the training set.
- Thus we see that **minimizing KL divergence** to the *empirical distribution* is equivalent to **maximizing likelihood**.
- This perspective points out the **flaw with likelihood-based training**, namely that it puts too much weight on the training set.
- We do not really believe that the **empirical distribution** is a *good representation* of the *true distribution*, since it just puts *"spikes"* on a finite set of points, and zero density everywhere else.
- We could *smooth the empirical distribution* using *kernel density estimation*, but that would require a similar kernel on the space of images.
- An alternative, algorithmic approach is to use **data augmentation**, which is a way of **perturbing the observed data samples** in way that we believe reflects plausible *"natural variation"*.

Forward vs reverse KL

Forwards KL or the *inclusive KL*, where minimizing this wrt q is known as an **moment projection**.:

$$D_{\mathrm{KL}}(p||q) = \int p(x) \log \frac{p(x)}{q(x)} dx$$

Reverse KL or the *exclusive KL*, where minimizing this wrt q is known as an **information projection**.:

$$D_{\mathrm{KL}}(q||p) = \int q(x) \log \frac{q(x)}{p(x)} dx$$

Out[ ]:

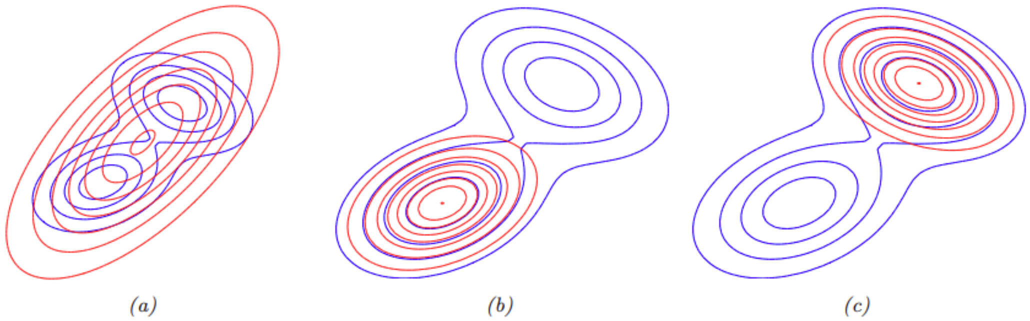


Figure 6.3: Illustrating forwards vs reverse KL on a bimodal distribution. The blue curves are the contours of the true distribution p. The red curves are the contours of the unimodal approximation q. (a) Minimizing forwards KL, $D_{\mathbb{KL}}(p\|q)$, wrt q causes q to "cover" p. (b-c) Minimizing reverse KL, $D_{\mathbb{KL}}(q\|p)$ wrt q causes q to "lock onto" one of the two modes of p. Adapted from Figure 10.3 of [Bis06]. Generated by code at figures.probml.ai/book1/6.3.

6.3 Mutual information

The KL divergence gave us a way to measure how similar two distributions were. How should we measure **the dependence between two random variables**? The answer lies in the similarity of their joint and marginal distributions: **Mutual information (MI)** between two random variables.

$$\mathbb{I}(X;Y) \triangleq D_{\mathbb{KL}}(p(x,y)||p(x)p(y)) = \sum_{y\in\mathcal{Y}}\sum_{x\in\mathcal{X}} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

- Writing $\mathbb{I}(X;Y)$ instead of $\mathbb{I}(X,Y)$ says in case X and/or Y represent sets of variables; for example, we can write $\mathbb{I}(X;Y,Z)$ to represent the MI between X and (Y,Z).
- For continuous random variables, we just replace sums with integrals.
- MI is always non-negative, even for continuous random variables.

$$\mathbb{I}(X;Y) \triangleq D_{\mathbb{KL}}(p(x,y)||p(x)p(y)) \geq 0$$

We achieve the bound of 0 iff $p(x,y) = p(x)p(y)$.

$$\mathbb{I}(X;Y) = \mathbb{H}(X) - \mathbb{H}(X|Y) = \mathbb{H}(Y) - \mathbb{H}(Y|X)$$

The MI between X and Y as the reduction in uncertainty about X after observing Y, or, by symmetry, the reduction in uncertainty about Y after observing X.

- Incidentally, this result gives an alternative proof that conditioning, on average, reduces entropy. In particular, we have $0 \leq \mathbb{I}(X;Y) = \mathbb{H}(X) - \mathbb{H}(X|Y)$, and hence $\mathbb{H}(X|Y) \leq \mathbb{H}(X)$.

A different interpretation

$$\begin{aligned} \mathbb{I}(X;Y) &= \mathbb{H}(X,Y) - \mathbb{H}(X|Y) - \mathbb{H}(Y|X) && (6) \\ &= \mathbb{H}(X) + \mathbb{H}(Y) - \mathbb{H}(X,Y) && (7) \end{aligned}$$

Out[ ]:

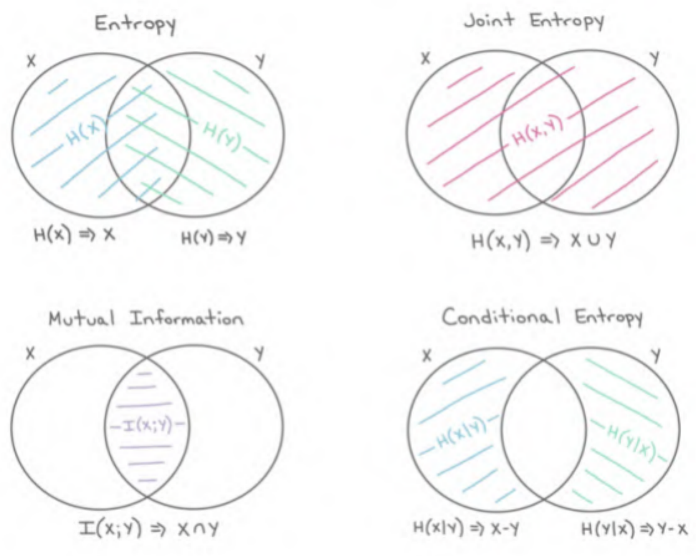


Figure 6.4: The marginal entropy, joint entropy, conditional entropy and mutual information represented as information diagrams. Used with kind permission of Katie Everett.

Example: Consider choosing an integer from 1 to 8, $n \in \{1, \ldots, 8\}$. Let $X(n) = 1$ if n is even, and $Y(n) = 1$ if n is prime:

n	1	2	3	4	5	6	7	8
X	0	1	0	1	0	1	0	1
Y	0	1	1	0	1	0	1	0

Recall that $\mathbb{H}(X) = \mathbb{H}(Y) = 1$. The conditional distribution $p(Y|X)$ is given by normalizing each row:

| $p(Y|X)$ | $Y = 0$ | $Y = 1$ | |----------|---------|---------| | $X = 0$ | 1/4 | 3/4 | | $X = 1$ | 3/4 | 1/4 |

$$\mathbb{H}(Y|X) = -\left[\frac{1}{8}\log_2\frac{1}{4} + \frac{3}{8}\log_2\frac{3}{4} + \frac{3}{8}\log_2\frac{3}{4} + \frac{1}{8}\log_2\frac{1}{4}\right] = 0.81 \text{ bits}$$

and the mutual information is

$$\mathbb{I}(X;Y) = \mathbb{H}(Y) - \mathbb{H}(Y|X) = (1 - 0.81) \text{ bits} = 0.19 \text{ bits}$$

One can easily verify that

$$\begin{aligned} \mathbb{H}(X,Y) &= \mathbb{H}(X|Y) + \mathbb{I}(X;Y) + \mathbb{H}(Y|X) && (8) \\ &= (0.81 + 0.19 + 0.81) \text{ bits} && (9) \\ &= 1.81 \text{ bits} && (10) \end{aligned}$$

Conditional mutual information

$$\begin{aligned}
\mathbb{I}(X;Y|Z) &\triangleq \mathbb{E}_{p(z)}[\mathbb{I}(X;Y|Z)] && (11)\\
&= \mathbb{E}_{p(x,y,z)}\left[\log \frac{p(x,y|z)}{p(x|z)p(y|z)}\right] && (12)\\
&= \mathbb{H}(X|Z) + \mathbb{H}(Y|Z) - \mathbb{H}(X,Y|Z) && (13)\\
&= \mathbb{H}(X|Z) - \mathbb{H}(X|Y,Z) = \mathbb{H}(Y|Z) - \mathbb{H}(Y|X,Z) && (14)\\
&= \mathbb{H}(X,Z) + \mathbb{H}(Y,Z) - \mathbb{H}(Z) - \mathbb{H}(X,Y,Z) && (15)\\
&= \mathbb{I}(Y;X,Z) - \mathbb{I}(Y;Z) && (16)
\end{aligned}$$

$$\mathbb{I}(Z,Y;X) = \mathbb{I}(Z;X) + \mathbb{I}(Y;X|Z)$$

Generalizing to N variables, we get the chain rule for mutual information:

$$\mathbb{I}(Z_1,\ldots,Z_N;X) = \sum_{n=1}^{N} \mathbb{I}(Z_n;X|Z_1,\ldots,Z_{n-1})$$

MI as a "generalized correlation coefficient"

In [ ]: