Information theory allows us to summarize uncertainty about, and relations between, random variables using real numbers. Often these numbers can serve as objective functions or constraints for algorithms and learning agents. The basic measures are (1) *entropy*, (2) *mutual information*, and (3) *relative entropy* or *KL divergence*. There are a few forms of each and there are important differences between the discrete and continuous cases.

Discrete entropy

History & definition: Info theory was founded by Shannon in his seminal 1948 paper "A mathematical theory of communication" [10]. To quantify the amount of information contained in a "communication", Shannon considered a scenario where there is a finite set of mutually exclusive and collectively exhaustive possibilities x_i, and some pre-communication belief about plausibility $p(x_i)$ of each x_i. That is, we have a **discrete variable** X with pmf p, and a communication provides information about its value. Shannon uses $H(X)$ to denote the total uncertainty in X (i.e., the expected information content of a communication that identifies the true x_i) and asserts the following axioms and theorem:

1. *H is continuous in $p(x_i)$.
2. *If x_i are equally likely, more alternatives means higher $H(X)$.
3. *If $X = (Y, Z)$, $H(X) = H(Y, Z) = H(Z) + \sum_i P(z_i)H(Y \,|\, z_i)$ (this says that the expected remaining uncertainty $H(Y \,|\, z_i)$ after receiving partial information $z_i \sim Z$ and the amount of partial information received $H(Z)$ sum to the total uncertainty).

Theorem 1. *Given the above axioms, H must have the form:*

$$H_b(X) = -\sum_x p(x)\log_b p(x) \tag{1}$$

for some base b. The functional H is known as ***entropy****.*

Shannon had this to say about the name:

> *My greatest concern was what to call it. I thought of calling it 'information,' but the word was overly used, so I decided to call it 'uncertainty.' When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, 'You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, no one knows what entropy really is, so in a debate you will always have the advantage.' [11]*

The key point of all this history is to get an intuition for where entropy came from, and why entropy has its peculiar $-\Sigma p \log p$ form.

***Entropy as *expected* code length:** The base b in Theorem 1 is unspecified. In information theory, we typically use $b = 2$, in which case $H(X)$ represents the expected length, in *bits*, of the shortest "code" that can be used to communicate the value of X. E.g., a constant has 0 entropy because we know what it is without any bits of communication. A Bernoulli variable with $p = 0.5$ requires $-\log_2 0.5 = 1$ bit, to tell us whether it is 1 or 0. A categorical with probabilities $(0.5, 0.25, 0.25)$ requires an average of $-0.5\log_2 0.5 - 0.5\log_2 0.25 = 1.5$ bits using the code $\{0, 10, 11\}$. In ML we typically use $b = e$ for convenience, in which case $H(X)$ is measured in *nats*.

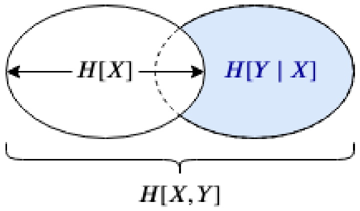
Some properties

- $H(X) = -\mathbb{E}_X\{p(X)\} = \mathbb{E}_X\{1/p(X)\}$
- $H(X) > 0$ (*for discrete entropy only!*)
- $H_b(X) = (\log_b a)H_a(X)$ (*for converting nats ⟷ bits*)
- Uniform distribution has highest H, and constants have 0 H.

Joint and Conditional Entropies

- **Entropy**: $H(X) = -\sum_x p(x)\log p(x)$.
- **Joint Entropy**: $H(X, Y) = -\sum_{x,y} p(x, y)\log p(x, y)$.
- **Cond. Entropy**: $H(Y \,|\, X) = -\sum_{x,y} p(x, y)\log p(y \,|\, x)$
- The **chain rule for entropy** (basically Axiom 3 above) is:

$$H(X, Y) = H(X) + H(Y \,|\, X)$$

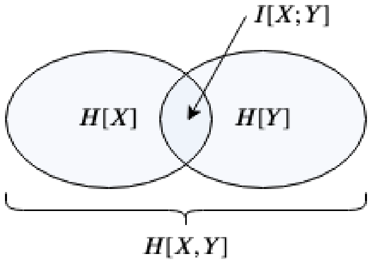


The above diagram is helpful visualizing this property. You can interpret each outline as defining some uncertainty. When the uncertainty that the shape represents is resolved, we remove the whole shape. Thus, the union of $H(X)$ and $H(Y)$ forms $H(X, Y)$. When $H(X)$ is resolved, only part of $H(Y)$ remains: $H(Y|X)$.

Just as in probability, we can condition everything on Z, so it is also true that: $H(X, Y \,|\, Z) = H(X \,|\, Z) + H(Y \,|\, X, Z)$.

Mutual Information

The little bit in the middle—the information shared between X and Y—is aptly named the "mutual information" $I(X, Y)$:



From the diagram, we immediately obtain the properties:

- $I(X; Y) = H(X) - H(X \,|\, Y) = H(Y) - H(Y \,|\, X)$.
- $I(X; Y) = H(X) + H(Y) - H(X, Y)$.
- $I(X; X) = H(X)$.

To confirm the properties algebraically, you can use the definition:

$$I(X; Y) = \sum_{x,y} p(x, y)\log\frac{p(x, y)}{p(x)p(y)}.$$

***MI as Expected Info Gain:** Intuitively, we can interpret $I(X; Y)$ as the *expected reduction in uncertainty about Y that results from knowing X* (and vice versa). Thus, mutual information is often used as an "information gain" objective—e.g., in active learning [3] and exploration in RL [7]—where we have a (Bayesian) belief $\boldsymbol{\theta}$ about our model parameters $\theta \sim \boldsymbol{\theta}$, and we expect next action a to produce observation $o \sim \mathcal{O} \,|\, a, \boldsymbol{\theta}$. We seek a that will maximally reduce the uncertainty in $\boldsymbol{\theta}$; i.e., our objective is $\max_a I(\boldsymbol{\theta}; \mathcal{O} \,|\, a, \theta)$.

***Entropy as diversity:** Another interpretation of uncertainty is *diversity*. So, e.g., if we want an RL agent to explore a diverse set of states, or if we wanted to maximize the diversity of hidden activations across a mini-batch, we might add an entropy bonus to our objective function. But given our definitions so far, we can only do this for discrete states / activations. Before we extend entropy to the continuous case, let's diverge a bit...

Relative entropy, also known as KL divergence

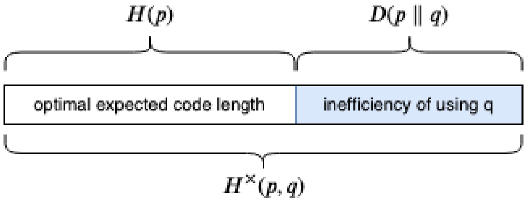
Cross Entropy To design the optimal code for communicating X, we need to know $p(x)$. Suppose we only have an approximation of p; by convention, we denote the approximation as q. How much longer does our code need to be? The total length of the optimal code when p is approximated by q is captured by the *cross entropy*:

$$H^{\times}(p,q) = -\sum_x p(x)\log q(x). \tag{2}$$

NB: Usually $H^{\times}(p,q)$ is written as just $H(p,q)$, which is notationally similar to joint entropy. What $H(\cdot,\cdot)$ refers to will usually be clear from the context and its arguments: joint entropy is a function of two variables (often with different ranges), whereas cross entropy is a function of two distributions on the same domain.

Below, for X with pmf p, we use $H(X)$ and $H(p)$ interchangeably.

Accepting that $H(p) = H^{\times}(p,p)$ is the optimal code length given the true distribution, and $H^{\times}(p,q)$ is the optimal code length given a suboptimal distribution (it's true, but we haven't proved either), it is intuitive that $H^{\times}(p,q) > H^{\times}(p,p)$. Then the difference $H^{\times}(p,q) - H(p)$ can be used measure the distance "from q to p"; i.e., it a measure of how good an approximation q is to p. We can use this to define **relative entropy** or **KL divergence** $D(p\,\|\,q)$:



KL divergence Using the definitions of $H(p)$, $H^{\times}(p,q)$, we have:

$$D(p\,\|\,q) = H^{\times}(p,q) - H^{\times}(p,p) = \sum_x p(x)\log\frac{p(x)}{q(x)}, \tag{3}$$

where $0\log\frac{0}{q>=0} = 0$ and $(p>0)\log\frac{p>0}{0} = \infty$ by convention.

Non-negativity of KL divergence From the above figure, we have $D(p\,\|\,q) \geq 0$ with equality if and only if $p = q$. Algebraically,

$$-D(p\,\|\,q) = \mathbb{E}_p \log\frac{q(x)}{p(x)} \leq \log\mathbb{E}_p\frac{q(x)}{p(x)} = \log 1 = 0, \tag{4}$$

where we've used Jensen's inequality (which says that for convex f, we have $\mathbb{E}f(X) \geq f(\mathbb{E}X)$, and vice versa for concave f.)

*Convexity and concavity of H, I, and D

- KL divergence is *convex* in both arguments.
- Entropy is *concave* ($H(\lambda p_1 + (1-\lambda)p_2) \geq \lambda H(p_1) + (1-\lambda)H(p_2)$).
- Let $(X,Y) \sim p(x,y) = p(x)p(y\,|\,x)$. Fixing $p(x)$, $I(X;Y)$ is *convex* in $p(y\,|\,x)$. Fixing $p(y\,|\,x)$, $I(X;Y)$ is *concave* in $p(x)$.

***KL as a starting point:** we motivated KL divergence from a code length perspective. But it may actually be a better *analytical* starting point than entropy, insofar as (1) it is better behaved in the continuous case, and (2) we can define both H and I in terms of D:

- $H(X) = \log n - D(p\,\|\,\mathcal{U}(n))$ for a n-valued variable $X \sim p$ (prove by putting $q = \mathcal{U}(n)$ in (3))
- $I(X;Y) = D(p(x,y)\,\|\,p(x)p(y))$ (easily verified)

Corollaries

- $I(X;Y) > 0$.
- $H(X|Y) \leq H(X)$ (information can't hurt).

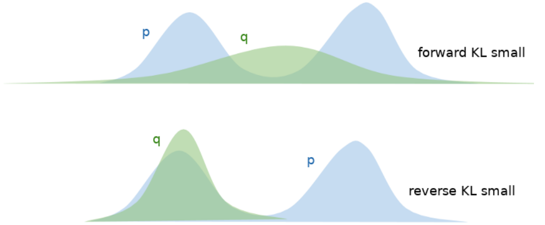
Minimizing KL and log likelihood. If our approximation q_θ of p is parameterized by θ (e.g., it is a neural network), notice that:

$$\arg\min_\theta D(p\,\|\,q_\theta) = \arg\min_\theta H^{\times}(p,q_\theta) = \arg\max_\theta \mathbb{E}_p \log q_\theta(x). \tag{5}$$

Thus, minimizing KL divergence is the same as minimizing cross entropy, which is the same as maximizing log likelihood.

Forward and reverse KL Occasionally we have a choice between minimizing $D(p\,\|\,q)$ and $D(q\,\|\,p)$ to make p and q more similar. KL is asymmetric, and there is an important differences between the two objectives. In particular, when p and q are in the usual (forward) alphabetical order, $D(p\,\|\,q)$ is known as *forward KL*. In reverse alphabetical order, $D(q\,\|\,p)$, corresponds to *reverse KL*.

forward KL	**reverse KL**
$D(p\,\|\,q)$	$D(q\,\|\,p)$
small when q "covers" p	small when q "covered by" p
blows up if $q = 0$ anywhere on the support of p	blows up if q has support anywhere that p does not
q is mode covering	q is mode seeking



In the top half of the figure, the green single hump (q) "covers" the two blue humps (p), so that forward KL is small. In the bottom half, the single hump (q) seeks out the mode of the two humps (p) and forward KL is very large (or ∞), while reverse KL is small.

Jensen-Shannon divergence We may dislike that KL divergence is asymmetric, and blows up when q does not cover p. Jensen-Shannon divergence (JSD) is alternative distance between p and q that is symmetric, and never blows up. It is defined as:

$$\mathrm{JSD}(p\,\|\,q) = \mathrm{JSD}(q\,\|\,p) = \frac{1}{2}D(p\,\|\,m_{pq}) + \frac{1}{2}D(q\,\|\,m_{pq}), \tag{6}$$

where m_{pq} is a 50/50 mixture of p and q (i.e., $m_{pq} = 0.5p + 0.5q$).

***Mutual information characterization of JSD and GANs** Suppose variable X is drawn from mixture m_{pq} between p and q (e.g., p is real data and q is data generated by a GAN generator), and binary variable $Z \sim$ Bernoulli(0.5) identifies the active mixture component (e.g., Z is the label the GAN discriminator is trying to guess). It turns out that $I(X;Z) = \mathrm{JSD}(p;q)$. Thus, if we view the traditional objective for the GAN generator as optimizing JSD (assuming optimal discriminator; see proof of Theorem 1 in [6]), it can also be understood as minimizing the mutual information between the mixture data and its source. Proof:

$$\begin{aligned}
I(X;Z) &= H(X) - H(X\,|\,Z) \\
&= -\sum m_{pq}\log m_{pq} + \frac{1}{2}\Big[\sum p\log p + \sum q\log q\Big] \\
&= -\frac{1}{2}\Big[\sum p\log m_{pq} + \sum q\log m_{pq}\Big] + \frac{1}{2}\Big[\sum p\log p + \sum q\log q\Big] \\
&= \frac{1}{2}\Big[\sum p(\log p - \log m_{pq}) + \sum q(\log q - \log m_{pq})\Big] \\
&= \mathrm{JSD}(p\,\|\,q).
\end{aligned}$$

Since the $I(X;Z) \leq \min(H(X),H(Z))$ and we have $H(Z) = 1$ when using bits (base 2), this also proves that $0 \leq \mathrm{JSD}(p\,\|\,q) \leq 1$, $\forall p,q$.

Variational Approximations

When we have a function f or distribution p that is unknown or intractable, we can sometimes approximate it by solving an optimization problem. This is known as a "variational approach".

***Variational approach to $\log x$ [8]** To understand the usage of the term, consider a variational approach to computing $\log x$. As you can verify by differentiating, $\log x = \min_\theta(\theta x - \log\theta - 1)$. Thus we can compute $\log x$ by introducing *variational parameter* θ and minimizing the *variational upper bound* $\theta x - \log\theta - 1$.

Variational inference If we use a variational approach to Bayesian inference, we are doing *variational inference*. Typically our model is a joint distribution $p(x,z) = p(z)p(x\,|\,z)$ (e.g., z is the latent cause of observation x) and we seek a variational approximation $q_\phi(z)$ to the model posterior $p(z\,|\,x)$. For arbitrary $q(z)$ we have:

$$\begin{aligned}\log p(x) &= \log\mathbb{E}_{z\sim p(z)}p(z)p(x\,|\,z)\\ &= \log\mathbb{E}_{z\sim q(z)}\frac{p(z)}{q(z)}p(x\,|\,z)\\ &\geq \mathbb{E}_{z\sim q(z)}\log\frac{p(z)}{q(z)}p(x\,|\,z)\\ &= \mathbb{E}_{z\sim q(z)}\big[p(x\,|\,z)\big] - D(q(z)\,\|\,p(z)),\end{aligned} \tag{7}$$

where the second line uses importance sampling and the third uses Jensen's inequality. The final line is the *variational* or *evidence lower bound* (ELBO) on $\log p(x)$. While this "I.S.-Jensen" derivation is simple, the following "KL-Bayes" one is more illuminating:

$$\begin{aligned}D(q(z)\,\|\,p(z\,|\,x)) &= \mathbb{E}_{z\sim q(z)}\big[\log q(z) - \log p(z\,|\,x)\big]\\ &= \mathbb{E}_{z\sim q(z)}\big[\log q(z) - \log p(x\,|\,z) - \log p(z) + \log p(x)\big]\\ &= \mathbb{E}_{z\sim q(z)}\big[\log p(x\,|\,z)\big] + D(q(z)\,\|\,p(z)) + \log p(x),\end{aligned}$$

where the second line uses Bayes theorem. Rearranging we get:

$$\log p(x) - D(q(z)\,\|\,p(z\,|\,x)) = \mathbb{E}_{z\sim q(z)}\big[p(x\,|\,z)\big] - D(q(z)\,\|\,p(z)). \tag{8}$$

This derivation precisely quantifies the difference between $\log p(x)$ and the ELBO (RHS) as $D(q(z)\,\|\,p(z\,|\,x))$. Now $q(z)$ was arbitrary, so if we parameterize $q_\phi(z\,|\,x)$ to make this difference small, and parameterize our data model $p_\theta(x,z) = p_\theta(z)p_\theta(z\,|\,x)$, we recover the cost function for the *variational autoencoder* [9]:

$$J(\theta,\phi,x_i) = -\mathbb{E}_{z\sim q_\phi(z\,|\,x_i)}\big[p_\theta(x_i\,|\,z)\big] + D(q_\phi(z\,|\,x_i)\,\|\,p_\theta(z)). \tag{9}$$

***Variational approximations to I, D, $\mathbb{E}(X)$** As you can infer, finding variational approximations requires some inventiveness. So it's instructive to see a few more examples.
The following upper and lower bounds on $I(X;Y)$, due to [1], are similar to the above in that they replace some p with variational approximation q to obtain a bound with tightness in terms of $D(q\,\|\,p)$ or $D(p\,\|\,q)$. First the upper bound on $I(X;Y)$:

$$\begin{aligned}I(X;Y) &= \mathbb{E}_{x,y\sim p(x,y)}\log\frac{p(y\,|\,x)}{p(y)}\left(\frac{q(y)}{q(y)}\right)\\ &= \mathbb{E}_{x,y\sim p(x,y)}\left[\log\frac{p(y\,|\,x)}{q(y)}\right] + \mathbb{E}_{y\sim p(y)}\left[\log\frac{q(y)}{p(y)}\right]\\ &= \mathbb{E}_{x\sim p(x)}\mathbb{E}_{y\sim p(y\,|\,x)}\left[\log\frac{p(y\,|\,x)}{q(y)}\right] - D(p(y)\,\|\,q(y)\\ &\leq \mathbb{E}_{x\sim p(x)}D(p(y\,|\,x)\,\|\,q(y)).\end{aligned} \tag{10}$$

Similarly, we can obtain a variational lower bound on $I(X;Y)$:

$$\begin{aligned}I(X;Y) &= \mathbb{E}_{x,y\sim p(x,y)}\log\frac{p(x\,|\,y)}{p(x)}\left(\frac{q(x\,|\,y)}{q(x\,|\,y)}\right)\\ &= \mathbb{E}_{x,y\sim p(x,y)}\left[\log\frac{q(x\,|\,y)}{p(x)}\right] + \mathbb{E}_{y\sim p(y)}\mathbb{E}_{x\sim p(x\,|\,y)}\left[\log\frac{p(x\,|\,y)}{q(x\,|\,y)}\right]\\ &= \mathbb{E}_{x,y\sim p(x,y)}\big[\log q(x\,|\,y)\big] + H(X) + \mathbb{E}_{x\sim p(x)}D(p(y\,|\,x)\,\|\,q(y))\\ &\geq \mathbb{E}_{x,y\sim p(x,y)}\big[\log q(x\,|\,y)\big] + H(X).\end{aligned}$$

We also state the Donsker-Varadhan (DV) formula for $D(p\,\|\,q)$:

$$D(p\,\|\,q) = \sup_{f:X\to\mathbb{R}}\mathbb{E}_p\big[f\big] - \log\mathbb{E}_q\big[\exp(f)\big]. \tag{11}$$

If f in the right hand side is parameterized by a neural network, we obtain a variational lower bound on $D(p\,\|\,q)$ [5, 2].

Finally, we state a variational formula for $\mathbb{E}_p(x)$:

$$\log\mathbb{E}_{x\sim p(x)}(x) = \sup_{q\in Q}\big[\mathbb{E}_{x\sim q(x)}(\log x) - D(q(x)\,\|\,p(x))\big] \tag{12}$$

where Q is the set of distributions ($q(x) \geq 0$, $\sum_x q(x) = 1$). You can prove this extremizing the Lagrangian of the RHS ([4], (8.93)).

*Differential Entropy (briefly)

When X is continuous with density p, we define "differential" entropy $h(X)$ (or $h(f)$) as the continuous analog to the discrete case:

$$h(X) = -\int_{\mathrm{supp}(X)} p(x)\log p(x)\,dx. \tag{13}$$

E.g., $X \sim \mathcal{U}(0,a)$ has $h(X) = -\int_0^a \frac{1}{a}\log\frac{1}{a}\,dx = \log a$. Unlike the discrete case, $h(X)$ can be negative! We see $h(X)$ scales with the size of X's support. So unlike the discrete case, where $H(f(X)) \leq H(X)$, applying f can increase differential entropy.

To understand why differential entropy behaves differently, consider the discrete entropy of an n-bit quantization of continuous variable X with pdf f and support $[0,1]$. Letting $\Delta = 1/2^n$ represent the width of each of the 2^n bins. The i-th bin has probability $\Delta f(i\Delta)$, so that discretized $X^{(n)}$ has entropy:

$$\begin{aligned}H(X^{(n)}) &= -\Sigma_{i=0}^{2^n-1}\Delta f(i\Delta)\log\Delta f(i\Delta)\\ &= -\Sigma_{i=0}^{2^n-1}\Delta f(i\Delta)\log f(i\Delta) - \Sigma_{i=0}^{2^n-1}\Delta f(i\Delta)\log\Delta\\ &= -\Sigma_{i=0}^{2^n-1}\Delta f(i\Delta)\log f(i\Delta) - \log\Delta.\end{aligned} \tag{14}$$

As $n \to \infty$, the second term blows up, while the first term approaches $h(X)$ if $f(x)\log f(x)$ is Riemann integrable.

Fortunately, $I(X;Y) = H(X) - H(X\,|\,Y)$ and $D(p\,\|\,q) = H^\times(p,q) - H(p)$ are both differences—when we quantize each term as above, the $\log\Delta$ cancels out, and the remainder is (often) finite (so long as their integral exists). Unlike entropy, I and D retain their properties in continuous case—i.e., we still have $D(p\,\|\,q) \geq 0$.

References

[1] D. Barber and F. V. Agakov. The im algorithm: a variational approach to information maximization. In *Advances in neural information processing systems*, page None, 2003.

[2] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm. Mutual information neural estimation. In *International Conference on Machine Learning*, pages 531–540, 2018.

[3] W. H. Beluch, T. Genewein, A. Nürnberger, and J. M. Köhler. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9368–9377, 2018.

[4] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[5] M. D. Donsker and S. S. Varadhan. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on Pure and Applied Mathematics*, 36(2):183–212, 1983.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[7] R. Houthooft, X. Chen, Y. Duan, J. Schulman, F. De Turck, and P. Abbeel. Vime: Variational information maximizing exploration. In *Advances in Neural Information Processing Systems*, pages 1109–1117, 2016.

[8] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[9] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[10] C. E. Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

[11] M. Tribus and E. C. McIrvine. Energy and information. *Scientific American*, 225(3):179–190, 1971.