# Lecture 1e: Information Theory

## Entropy, KL divergence, and mutual information

Jue Guo

University at Buffalo

May 7, 2026

# Outline

# Where we are

**Chapter 1, last stop.** We've built a probabilistic model (1b), chosen its complexity (1c), and learned how to turn its outputs into decisions (1d).

**What's left.** A small toolbox of information-theoretic quantities that will reappear constantly later in the course:

- Entropy – how much uncertainty is in a distribution.
- KL divergence – how different two distributions are; secretly, what maximum likelihood is minimising.
- Mutual information – how much one variable tells you about another.

## Why bother

Density estimation, model selection (evidence), variational inference, representation learning, and self-supervised pretraining all live in this language.

# Information content of an outcome

Observing a value of a random variable conveys some information. What should it be a function of? **Desiderata.**

- Depends only on $p(x)$, via some monotonic $h(p(x))$.
- Rare events are more informative than common ones.
- Independent events combine additively: $h(x, y) = h(x) + h(y)$ when $p(x, y) = p(x)\,p(y)$.

Only one function does it (up to a base):

$$h(x) \;=\; -\log_2 p(x). \tag{1}$$

Units are bits for $\log_2$, nats for ln. We'll use natural logs from now on – they connect cleanly to ML expressions.

# Entropy: average information

The expected information content under $p$ is the entropy:

$$H[x] \;=\; -\sum_x p(x)\,\ln p(x), \qquad \text{(continuous: } H[x] = -\int p(x)\,\ln p(x)\,dx\text{)}. \tag{2}$$

Convention: $p \ln p \to 0$ as $p \to 0$, so zero-probability outcomes don't break the sum.

**Properties (discrete case).**

- $H[x] \geq 0$, with equality only for a deterministic distribution.
- Bounded: $H[x] \leq \ln M$ for $M$ states, attained at the uniform distribution.
- Concave in $p$ (mixing distributions can only increase entropy).

# Worked example – coding eight states

**Uniform.** Eight equally-likely states. Send the index in 3 bits:

$$H[x] = -8 \cdot \tfrac{1}{8} \log_2 \tfrac{1}{8} = 3 \text{ bits.}$$

**Nonuniform.** States $\{a, b, c, d, e, f, g, h\}$ with probabilities $\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}\right)$:

$$H[x] = -\left(\tfrac{1}{2} \log_2 \tfrac{1}{2} + \tfrac{1}{4} \log_2 \tfrac{1}{4} + \tfrac{1}{8} \log_2 \tfrac{1}{8} + \tfrac{1}{16} \log_2 \tfrac{1}{16} + 4 \cdot \tfrac{1}{64} \log_2 \tfrac{1}{64}\right) = 2 \text{ bits.}$$

A prefix code with strings $\{0, 10, 110, 1110, 111100, 111101, 111110, 111111\}$ gives an average length of exactly 2 bits.

## Shannon's noiseless coding theorem

Entropy is the *lower bound* on the average code length needed to transmit the value of a random variable.

# Entropy as "disorder"

Drop $N$ identical objects into bins of frequency $p_i = n_i/N$. The number of microstates compatible with the macrostate is the multiplicity

$$W = \frac{N!}{\prod_i n_i!}. \tag{3}$$

Stirling's approximation $\ln N! \simeq N \ln N - N$ gives

$$\frac{1}{N} \ln W \xrightarrow{N\to\infty} -\sum_i p_i \ln p_i = H[p]. \tag{4}$$

**Physics.** Boltzmann's $S = k \ln W$ – entropy as the log of the number of microscopic arrangements. Same formula, same intuition: peaked distributions have few arrangements (low entropy), spread-out distributions have many (high entropy).

Compare: a uniform distribution over 30 bins has $H = \ln 30 \approx 3.40$; a sharply peaked one over the same 30 bins might be $H \approx 1.77$. (PRML Fig. 1.30.)

# Differential entropy

Bin a continuous $p(x)$ into intervals of width $\Delta$. The discrete entropy of the binned distribution is

$$H_\Delta \;=\; -\sum_i p(x_i)\,\Delta \ln(p(x_i)\,\Delta) \;=\; -\sum_i p(x_i)\,\Delta \ln p(x_i) \;-\; \ln \Delta.$$

Drop the diverging $-\ln \Delta$ and let $\Delta \to 0$:

$$H[x] \;=\; -\int p(x)\,\ln p(x)\,dx. \tag{5}$$

- Multivariate version: $H[\mathbf{x}] = -\int p(\mathbf{x})\,\ln p(\mathbf{x})\,d\mathbf{x}$.
- Unlike its discrete counterpart, the differential entropy can be negative.
- It is also *not invariant* under reparameterisation – a Jacobian shifts it. That's why pure differential entropy isn't a great absolute measure; *differences* (i.e. KL divergences) behave better.

# Maximum entropy distributions

**Discrete, fixed support $M$.** Maximise $H$ subject only to $\sum p_i = 1$. A Lagrange multiplier gives $p_i = 1/M$: the uniform distribution maximises entropy.

**Continuous, with mean $\mu$ and variance $\sigma^2$ fixed.** Maximise $-\int p \ln p \, dx$ under

$$\int p(x)\,dx = 1, \quad \int x\,p(x)\,dx = \mu, \quad \int (x-\mu)^2\,p(x)\,dx = \sigma^2.$$

Calculus of variations yields

$$p^\star(x) \;=\; \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\!\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

Conclusion. Among distributions on $\mathbb{R}$ with given mean and variance, the Gaussian maximises entropy. This is one of the deeper reasons it is the workhorse distribution of the course.

# Differential entropy of a Gaussian

Plug $p^\star$ back into $H$:

$$H[\mathcal{N}(\mu, \sigma^2)] \;=\; \tfrac{1}{2}\big(1 + \ln(2\pi\sigma^2)\big). \tag{6}$$

- Independent of $\mu$ – entropy is a property of *shape*, not location.
- Increases with $\sigma^2$: broader Gaussians are higher-entropy.
- Negative when $\sigma^2 < 1/(2\pi e) \approx 0.0585$: a sharply peaked density can have differential entropy below zero – a reminder it is not the same kind of object as discrete entropy.

Multivariate generalisation: $H[\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})] = \frac{1}{2}\big(D + \ln\big((2\pi)^D|\boldsymbol{\Sigma}|\big)\big)$. The log-determinant of the covariance is the only $\boldsymbol{\Sigma}$-dependent piece.

## Joint and conditional entropy

For a joint distribution $p(x, y)$:

$$H[x, y] \;=\; -\iint p(x, y)\, \ln p(x, y)\, dx\, dy. \tag{7}$$

**Conditional entropy** – the average extra information needed to specify $y$ once $x$ is known:

$$H[y \mid x] \;=\; -\iint p(x, y)\, \ln p(y \mid x)\, dx\, dy \;=\; \mathbb{E}_x\big[H[y \mid X = x]\big]. \tag{8}$$

**Chain rule.** From $\ln p(x, y) = \ln p(y \mid x) + \ln p(x)$:

$$H[x, y] \;=\; H[y \mid x] + H[x]. \tag{9}$$

The total information needed to describe $(x, y)$ is the cost of $x$ plus the additional cost of $y$ given $x$. Conditional entropy is non-negative and $H[y \mid x] \le H[y]$ with equality iff $x$ and $y$ are independent.

# Approximating $p$ by $q$ – the cost of being wrong

Suppose the true distribution is $p(x)$ but we model it with $q(x)$. Coding under $q$ costs $-\ln q(x)$ nats per symbol on average; the optimum (under $p$) is $-\ln p(x)$. The *extra* cost is the Kullback-Leibler divergence:

$$\mathrm{KL}(p \,\|\, q) \;=\; -\int p(x)\,\ln\frac{q(x)}{p(x)}\,dx \;=\; \int p(x)\,\ln\frac{p(x)}{q(x)}\,dx. \tag{10}$$

- $\mathrm{KL}(p \,\|\, q) \geq 0$, with equality iff $p = q$ almost everywhere.
- Asymmetric. $\mathrm{KL}(p \,\|\, q) \neq \mathrm{KL}(q \,\|\, p)$ in general.
- Not a metric – but the right object for many problems anyway.

# Why KL $\geq 0$: convexity and Jensen

**Convex** $f$: every chord lies on or above the curve.

$$f(\lambda a + (1-\lambda)b) \;\leq\; \lambda f(a) + (1-\lambda)f(b), \quad 0 \leq \lambda \leq 1. \tag{11}$$

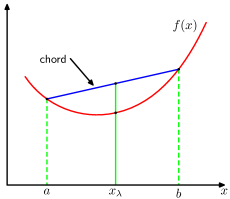


Figure 1.31

**Jensen's inequality.** For convex $f$ and a probability distribution: $f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)]$. Apply with $f(u) = -\ln u$ (strictly convex) to KL:

$$\mathrm{KL}(p \,\|\, q) \;=\; \mathbb{E}_p\left[-\ln \frac{q}{p}\right] \;\geq\; -\ln \mathbb{E}_p\left[\frac{q}{p}\right] \;=\; -\ln \int q(x)\,dx \;=\; 0.$$

Equality only when $q = p$ almost everywhere. (PRML Fig. 1.31.)

# Minimising KL = maximum likelihood

Suppose true data come from $p(\mathbf{x})$; we model them with $q(\mathbf{x} \mid \boldsymbol{\theta})$ and want to minimise $\mathrm{KL}(p \,\|\, q_{\boldsymbol{\theta}})$. Write

$$\mathrm{KL}(p \,\|\, q_{\boldsymbol{\theta}}) \;=\; \mathbb{E}_p[\ln p(\mathbf{x})] \;-\; \mathbb{E}_p[\ln q(\mathbf{x} \mid \boldsymbol{\theta})].$$

The first term is a $\boldsymbol{\theta}$-independent constant. With i.i.d. samples $\{\mathbf{x}_n\}$ standing in for the unknown $p$,

$$\mathbb{E}_p[\ln q(\mathbf{x} \mid \boldsymbol{\theta})] \;\approx\; \frac{1}{N} \sum_{n=1}^{N} \ln q(\mathbf{x}_n \mid \boldsymbol{\theta}).$$

Maximising this is exactly maximum likelihood:

$$\arg\min_{\boldsymbol{\theta}} \mathrm{KL}(p \,\|\, q_{\boldsymbol{\theta}}) \;=\; \arg\max_{\boldsymbol{\theta}} \prod_{n} q(\mathbf{x}_n \mid \boldsymbol{\theta}).$$

Take-home. MLE is what you would do if you had access to the true distribution and wanted to be as close to it (in KL) as possible. The link with the sum-of-squares loss in 1a, via Gaussian noise, is one consequence.

# Concrete example – KL between two Gaussians

For $p = \mathcal{N}(\mu_1, \sigma_1^2)$ and $q = \mathcal{N}(\mu_2, \sigma_2^2)$, a direct integral gives the closed form

$$\mathrm{KL}(p \,\|\, q) \;=\; \ln \frac{\sigma_2}{\sigma_1} \;+\; \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\,\sigma_2^2} \;-\; \tfrac{1}{2}. \tag{12}$$

**Numbers.** Take $p = \mathcal{N}(0, 1)$ and $q = \mathcal{N}(1, 1)$:

$$\mathrm{KL}(p \,\|\, q) \;=\; \ln 1 + \tfrac{1+1}{2} - \tfrac{1}{2} \;=\; \tfrac{1}{2} \text{ nats} \approx 0.72 \text{ bits}.$$

Now reverse it (same shapes, swapped roles): $\mathrm{KL}(q \,\|\, p) = \frac{1}{2}$ too – by symmetry of these particular shapes. *Asymmetry shows up when the variances differ.* Take $p = \mathcal{N}(0, 1)$ and $q = \mathcal{N}(0, 4)$:

$$\mathrm{KL}(p \,\|\, q) = \ln 2 + \tfrac{1}{8} - \tfrac{1}{2} \approx 0.32, \qquad \mathrm{KL}(q \,\|\, p) = -\ln 2 + \tfrac{4}{2} - \tfrac{1}{2} \approx 0.81.$$

Reading. $\mathrm{KL}(p \,\|\, q)$ penalizes $q$ where $p$ has mass; $\mathrm{KL}(q \,\|\, p)$ penalizes $p$ where $q$ has mass. They answer different questions – "where is $q$ too small?" vs. "where is $q$ too large?". Variational methods exploit this distinction.

# Mutual information

For two variables $x, y$ with joint $p(x, y)$ and marginals $p(x), p(y)$,

$$I[x, y] \;\equiv\; \mathrm{KL}\big(p(x, y) \,\|\, p(x)p(y)\big) \;=\; \iint p(x, y)\, \ln \frac{p(x, y)}{p(x)\, p(y)}\, dx\, dy. \tag{13}$$

**Properties.**

- $I[x, y] \geq 0,\ = 0$ iff $x \perp y$.
- Symmetric: $I[x, y] = I[y, x]$.
- Connects entropy and conditional entropy:

$$I[x, y] \;=\; H[x] - H[x \mid y] \;=\; H[y] - H[y \mid x].$$

"How much knowing $y$ reduces uncertainty about $x$."

**Bayesian reading.** If $p(x)$ is the prior and $p(x \mid y)$ the posterior after seeing $y$, then $I[x, y]$ is the *average* reduction in uncertainty $p(x) \to p(x \mid y)$. Information gain.

# Concrete example – mutual information of a 2×2 joint

Toy joint distribution between $x \in \{0,1\}$ and $y \in \{0,1\}$:

| | $y=0$ | $y=1$ | $p(x)$ |
|---|---|---|---|
| $x=0$ | 0.40 | 0.10 | 0.50 |
| $x=1$ | 0.10 | 0.40 | 0.50 |
| $p(y)$ | 0.50 | 0.50 | |

**If $x, y$ were independent**, every cell would equal $p(x)p(y) = 0.25$. They aren't – the diagonal is heavier.

**Mutual information.** Sum the four contributions $p(x,y)\,\ln\big[p(x,y)/(p(x)p(y))\big]$:

$$I[x,y] \;=\; 2 \cdot 0.40 \ln \tfrac{0.40}{0.25} \;+\; 2 \cdot 0.10 \ln \tfrac{0.10}{0.25} \;=\; 0.376 - 0.183 \;=\; \mathbf{0.193}\ \text{nats} \;\approx\; 0.278\ \text{bits}.$$

**Cross-check.** $H[x] = \ln 2 \approx 0.693$. The conditional entropy $H[x \mid y] = \ln 2 - I[x,y] \approx 0.500$ – knowing $y$ reduces our uncertainty about $x$ from $\ln 2$ nats to about 0.5 nats; mutual information is the size of that drop.

# Where these tools reappear

The same handful of quantities will keep showing up:

- Cross-entropy loss for classification: $-\sum_n \ln q(t_n \mid \mathbf{x}_n)$. Equal (up to a constant) to KL of true labels vs. model.
- ELBO / variational inference: maximising $\ln p(\mathcal{D})$ decomposes into a likelihood term minus $\mathrm{KL}(q(\boldsymbol{\theta}) \,\|\, p(\boldsymbol{\theta} \mid \mathcal{D}))$.
- Information bottleneck and representation learning: trade off $I[\mathbf{x}, \mathbf{z}]$ (compression) against $I[\mathbf{z}, t]$ (predictiveness).
- Mutual-information maximisation (InfoNCE etc.) underlies modern contrastive self-supervised methods.
- Decision trees / feature selection: split or rank features by information gain about the target.

Same vocabulary, different applications.

# Takeaways

- Information of an outcome is $-\ln p(x)$; entropy is its expectation. Concave, non-negative (discrete), maximised at uniform.
- For continuous variables, the differential entropy can be negative and is reparameterisation-dependent – use it via differences (KL).
- Among distributions with given mean and variance, the Gaussian maximises entropy.
- Chain rule: $H[x, y] = H[x] + H[y \mid x]$.
- KL divergence is the extra-bits cost of using $q$ instead of $p$. Always $\geq 0$, zero iff equal. Asymmetric.
- Minimising $\mathrm{KL}(p_{\text{data}} \,\|\, q_{\boldsymbol{\theta}})$ is maximum likelihood.
- Mutual information $I[x, y]$ measures how much one variable tells you about the other; = KL between joint and product of marginals.

## Reading

Bishop, PRML (2006), §1.6. See also Cover & Thomas, MacKay for full treatments.