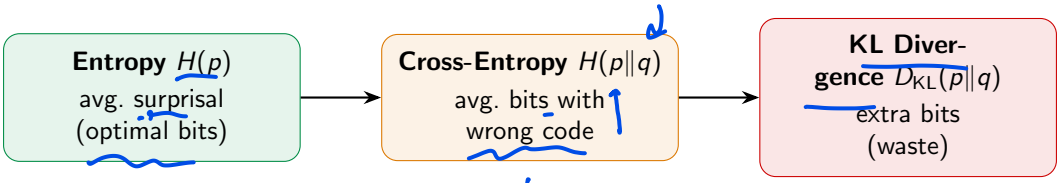
Information Theory II: ML Connections

KL = MLE · Cross-Entropy Loss · Forward/Reverse KL · Mutual Information

Recap: The Three Pillars



$$H(p\|q) = H(p) + D_{\mathrm{KL}}(p\|q)$$

Today: We connect these concepts to machine learning.
KL $\rightarrow$ MLE, cross-entropy $\rightarrow$ log-loss, forward/reverse KL,
maximum entropy principle, and mutual information.

KL = Maximum Likelihood

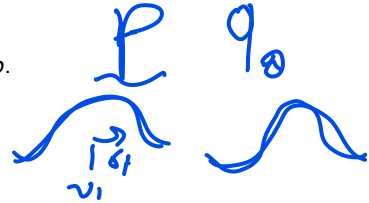
The single most important connection between information theory and machine learning.

Minimizing KL = Maximizing Log-Likelihood

Data from true p, model family q_θ. Find θ making q_θ closest to p.

Step 1: Write out $D_{\mathrm{KL}}(p\|q_\theta) = \sum_x p(x)[\log p(x) - \log q_\theta(x)]$.

Which of those two terms depends on θ?



Minimizing KL = Maximizing Log-Likelihood

Data from true p, model family q_θ. Find θ making q_θ closest to p.

Step 1: Write out $D_{\mathrm{KL}}(p\|q_\theta) = \sum_x p(x)[\log p(x) - \log q_\theta(x)]$.

Which of those two terms depends on θ?

$$D_{\mathrm{KL}}(p\|q_\theta) = \underbrace{\sum_x p(x)\log p(x)}_{= -H(p),\ \text{doesn't depend on } \theta!} - \sum_x p(x)\log q_\theta(x)$$

Minimizing KL = Maximizing Log-Likelihood

Data from true p, model family q_θ. Find θ making q_θ closest to p.

Step 1: Write out $D_{\mathrm{KL}}(p\|q_\theta) = \sum_x p(x)[\log p(x) - \log q_\theta(x)]$.

Which of those two terms depends on θ?

$$D_{\mathrm{KL}}(p\|q_\theta) = \underbrace{\sum_x p(x)\log p(x)}_{= -H(p),\ \text{doesn't depend on } \theta!} - \sum_x p(x)\log q_\theta(x)$$

Step 2: Drop the constant:

$$\arg\min_\theta \ D_{\mathrm{KL}}(p\|q_\theta) = \arg\max_\theta \sum_x p(x)\log q_\theta(x)$$

Minimizing KL = Maximizing Log-Likelihood

Data from true p, model family q_θ. Find θ making q_θ closest to p.

Step 1: Write out $D_{\mathrm{KL}}(p\|q_\theta) = \sum_x p(x)[\log p(x) - \log q_\theta(x)]$.
Which of those two terms depends on θ?

$$D_{\mathrm{KL}}(p\|q_\theta) = \underbrace{\sum_x p(x)\log p(x)}_{= -H(p),\ \text{doesn't depend on } \theta!} - \sum_x p(x)\log q_\theta(x)$$

Step 2: Drop the constant:

$$\arg\min_\theta \; D_{\mathrm{KL}}(p\|q_\theta) = \arg\max_\theta \sum_x p(x)\log q_\theta(x)$$

Step 3: Replace p by empirical $\hat{p}(x) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}[x_i = x]$:

$$\arg\max_\theta \sum_x \hat{p}(x)\log q_\theta(x) = \arg\max_\theta \frac{1}{n}\sum_{i=1}^n \log q_\theta(x_i)$$

$$\arg\min_\theta \; D_{\mathrm{KL}}(p\|q_\theta) \;=\; \arg\max_\theta \sum_{i=1}^n \log q_\theta(x_i) \;=\; \textbf{MLE!}$$

Cross-Entropy = MLE: A Concrete Example

Since $H(p\|q_\theta) = H(p) + D_{\mathrm{KL}}(p\|q_\theta)$ and $H(p)$ is constant in θ: the same optimum solves all three.

Minimize KL $=$ Minimize cross-entropy $=$ Maximize log-likelihood

Numerical example. 3-class classifier, true class is class 1. Loss is $-\log_2 \pi_1$:

Predicted (π_1, π_2, π_3)	**Loss (bits)**	**Interpretation**
$(0.7, 0.2, 0.1)$	0.51	confident, correct
$(0.4, 0.4, 0.2)$	1.32	uncertain
$(0.1, 0.1, 0.8)$	3.32	wrong, but "not too sure"
$(0.05, 0.05, 0.9)$	4.32	**confidently wrong!**
$(1.0, 0.0, 0.0)$	0	confident, correct (saturates)

Cross-entropy is **harsh on confident wrongness** — that's the whole point. Halving the predicted probability *doubles* the loss.

Cross-Entropy Loss in Classification

Multi-class classification with g classes. For observation i:

- True label $y_i \in \{1, \ldots, g\}$, one-hot encoded as $\mathbf{d}^{(i)} = (0, \ldots, 1, \ldots, 0)$
- Model outputs probability vector $\boldsymbol{\pi}(\mathbf{x}_i \mid \theta) = (\pi_1, \ldots, \pi_g)$

The cross-entropy between one-hot $\mathbf{d}^{(i)}$ and model $\boldsymbol{\pi}$:

$$H(\mathbf{d}^{(i)} \| \boldsymbol{\pi}) = -\sum_{k=1}^{g} d_k^{(i)} \log \pi_k(\mathbf{x}_i \mid \theta) = -\log \pi_{y_i}(\mathbf{x}_i \mid \theta)$$

Cross-Entropy Loss in Classification

Multi-class classification with g classes. For observation i:

- True label $y_i \in \{1, \ldots, g\}$, one-hot encoded as $\mathbf{d}^{(i)} = (0, \ldots, 1, \ldots, 0)$
- Model outputs probability vector $\boldsymbol{\pi}(\mathbf{x}_i \mid \theta) = (\pi_1, \ldots, \pi_g)$

The cross-entropy between one-hot $\mathbf{d}^{(i)}$ and model $\boldsymbol{\pi}$:

$$H(\mathbf{d}^{(i)} \| \boldsymbol{\pi}) = -\sum_{k=1}^{g} d_k^{(i)} \log \pi_k(\mathbf{x}_i \mid \theta) = -\log \pi_{y_i}(\mathbf{x}_i \mid \theta)$$

Summing over all observations:

$$\mathcal{R}(\theta) = -\sum_{i=1}^{n} \log \pi_{y_i}(\mathbf{x}_i \mid \theta) \qquad (= \text{negative log-likelihood} = \text{log-loss})$$

Cross-entropy loss IS log-loss IS negative log-likelihood.
Three names for the same thing. Softmax + cross-entropy = the standard recipe.

Binary Cross-Entropy = Bernoulli Log-Loss

Binary classification: $y \in \{0, 1\}$, model outputs $\pi(\mathbf{x}) = P(Y{=}1 \mid \mathbf{x})$.

$$L(y, \pi) = -y \log \pi(\mathbf{x}) - (1 - y) \log(1 - \pi(\mathbf{x}))$$

This is the cross-entropy between two Bernoulli distributions: $p = \text{Bern}(y)$ and $q = \text{Bern}(\pi(\mathbf{x}))$.

Connection to MLE / logistic regression: Logistic regression maximizes likelihood
$\Leftrightarrow$ minimizes binary cross-entropy $\Leftrightarrow$ minimizes KL to the true conditional.

The cross-entropy loss in logistic regression **is** information-theoretic!

Maximum Entropy Principle

Given partial knowledge, choose the distribution that is **maximally uncertain** about everything else.

The Maximum Entropy Principle (Jaynes, 2003)

We know constraints (e.g., $\mathbb{E}[g_m(X)] = \alpha_m$ for $m = 1, \ldots, M$). Many distributions satisfy them. Which to pick?

MaxEnt Principle: of all distributions satisfying the constraints, choose the one with **maximum entropy**.

Shore & Johnson, 1980: MaxEnt is the unique inference rule satisfying consistency under coarsening, independence, and continuity.

The Maximum Entropy Principle (Jaynes, 2003)

We know constraints (e.g., $\mathbb{E}[g_m(X)] = \alpha_m$ for $m = 1, \ldots, M$). Many distributions satisfy them. Which to pick?

MaxEnt Principle: of all distributions satisfying the constraints, choose the one with **maximum entropy**.

Shore & Johnson, 1980: MaxEnt is the unique inference rule satisfying consistency under coarsening, independence, and continuity.

Solution (Lagrangian):

$$p^*(x) = \frac{1}{Z} \exp\left(\sum_{m=1}^{M} \lambda_m \, g_m(x)\right) \qquad Z = \sum_x \exp\left(\sum_m \lambda_m \, g_m(x)\right)$$

The MaxEnt distribution is always an **exponential family** (Gibbs distribution)!

This is no coincidence — exponential families **are** the MaxEnt distributions.

MaxEnt Recovers Familiar Distributions

Different constraints $\Rightarrow$ different MaxEnt distributions:

Constraint(s)	**MaxEnt distribution**	**Connection**
None (fixed support $\{1, \ldots, g\}$)	Uniform	Max entropy $= \log g$
$\mathbb{E}[X] = \mu$ (on $\{1, \ldots, g\}$)	Exponential/Gibbs	Biased die
$\mathbb{E}[X] = \mu$, $\mathrm{Var}(X) = \sigma^2$	**Gaussian** $N(\mu, \sigma^2)$	Most common in ML!
$\mathbb{E}[X] = 1/\lambda$ (on $[0, \infty)$)	**Exponential**(λ)	Memoryless

Why the Gaussian is everywhere in ML: if you only know the mean and variance, the Gaussian is the **least informative** distribution compatible with that knowledge — any other choice smuggles in extra assumptions.

Proof: Gaussian Maximizes Differential Entropy

Claim: Among all densities p on $\mathbb{R}$ with mean μ and variance σ^2, the Gaussian $g = N(\mu, \sigma^2)$ maximizes differential entropy.

Proof (via Gibbs' inequality applied to differential entropy):

Let p be any density with the same mean and variance as g. Consider $D_{\mathrm{KL}}(p\|g) \geq 0$:

$$0 \leq D_{\mathrm{KL}}(p\|g) = \int p(x) \log_2 \frac{p(x)}{g(x)}\, dx = -h(p) - \int p(x) \log_2 g(x)\, dx.$$

Plug in $\log_2 g(x) = -\frac{1}{2}\log_2(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2 \ln 2}$ and take the integral. Since p has the *same* mean and variance as g, the integral of $-\log_2 g(x)$ under p equals its integral under g:

Proof: Gaussian Maximizes Differential Entropy

Claim: Among all densities p on $\mathbb{R}$ with mean μ and variance σ^2, the Gaussian $g = N(\mu, \sigma^2)$ maximizes differential entropy.

Proof (via Gibbs' inequality applied to differential entropy):

Let p be any density with the same mean and variance as g. Consider $D_{\mathrm{KL}}(p\|g) \geq 0$:

$$0 \leq D_{\mathrm{KL}}(p\|g) = \int p(x) \log_2 \frac{p(x)}{g(x)}\, dx = -h(p) - \int p(x) \log_2 g(x)\, dx.$$

Plug in $\log_2 g(x) = -\frac{1}{2}\log_2(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2 \ln 2}$ and take the integral. Since p has the *same* mean and variance as g, the integral of $-\log_2 g(x)$ under p equals its integral under g:

$$-\int p(x) \log_2 g(x)\, dx \;=\; -\int g(x) \log_2 g(x)\, dx \;=\; h(g) \;=\; \tfrac{1}{2}\log_2(2\pi e\sigma^2).$$

Substituting back: $0 \leq -h(p) + h(g) \;\Rightarrow\; h(p) \leq h(g)$. ■

Same trick works for every MaxEnt distribution in the table above: pick the candidate distribution, write $D_{\mathrm{KL}}(p\|\text{candidate}) \geq 0$, exploit matched moments. The exponential family is essentially the MaxEnt family.

The CLT, Information-Theoretically

Classical CLT (recall, Module 20). Let $X_1, X_2, \ldots$ be i.i.d. with mean 0 and variance σ^2. The normalized sum

$$Z_n := \frac{X_1 + X_2 + \cdots + X_n}{\sqrt{n}} \xrightarrow{d} N(0, \sigma^2).$$

The *distribution* converges to a Gaussian. What does the *entropy* do?

Use the differential-entropy scaling rule $h(aX) = h(X) + \log_2 |a|$. With $S_n = X_1 + \cdots + X_n$ (so $\mathrm{Var}(S_n) = n\sigma^2$) and $a = 1/\sqrt{n}$:

$$h(Z_n) = h\left(\frac{S_n}{\sqrt{n}}\right) = h(S_n) - \tfrac{1}{2}\log_2 n.$$

The CLT, Information-Theoretically

Classical CLT (recall, Module 20). Let $X_1, X_2, \ldots$ be i.i.d. with mean 0 and variance σ^2. The normalized sum

$$Z_n := \frac{X_1 + X_2 + \cdots + X_n}{\sqrt{n}} \xrightarrow{d} N(0, \sigma^2).$$

The *distribution* converges to a Gaussian. What does the *entropy* do?

Use the differential-entropy scaling rule $h(aX) = h(X) + \log_2 |a|$.
With $S_n = X_1 + \cdots + X_n$ (so $\mathrm{Var}(S_n) = n\sigma^2$) and $a = 1/\sqrt{n}$:

$$h(Z_n) = h\left(\tfrac{S_n}{\sqrt{n}}\right) = h(S_n) - \tfrac{1}{2}\log_2 n.$$

Entropic CLT (Barron, 1986): $h(Z_n)$ is *monotonically increasing in* n and converges to the Gaussian's entropy,

$$h(Z_n) \nearrow h\big(N(0,\sigma^2)\big) = \tfrac{1}{2}\log_2(2\pi e\,\sigma^2).$$

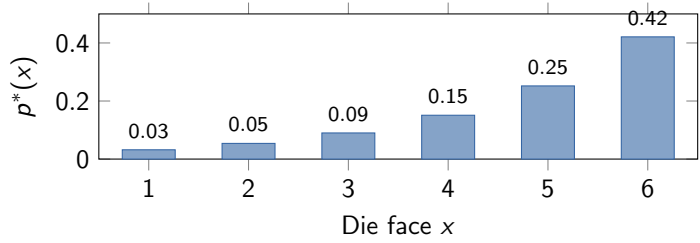
Among distributions with variance σ^2, the Gaussian has **maximum** differential entropy (we proved this). Summation **climbs that hill**: convolution is entropy-increasing, and the Gaussian is its **fixed point** — the entropy attractor.

MaxEnt Example: The Biased Die

A 6-sided die with $\mathbb{E}[X] = 4.8$ (heavier than the usual 3.5). What's the MaxEnt distribution?

$$p^*(x) = \frac{e^{\lambda x}}{Z(\lambda)}, \qquad Z(\lambda) = \sum_{x=1}^{6} e^{\lambda x}, \qquad \sum_{x=1}^{6} x\, p^*(x) = 4.8$$

Solving numerically: $\lambda \approx 0.514$.



Exponential tilt toward higher faces — the "least opinionated" way to have $\mathbb{E}[X] = 4.8$.

Sign of λ encodes the deviation: $\lambda = 0 \to$ uniform; $\lambda > 0 \to$ higher faces; $\lambda < 0 \to$ lower.

Mutual Information

How much does knowing X tell us about Y?

A universal measure of dependence.

Joint and Conditional Entropy

Joint entropy: Total uncertainty in (X, Y) together:

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

Conditional entropy: Remaining uncertainty in Y after observing X:

$$H(Y \mid X) = -\sum_{x,y} p(x, y) \log p(y \mid x) = \mathbb{E}_X[H(Y \mid X{=}x)]$$

Two natural questions:

- How does $H(X, Y)$ decompose? (Answer: the *chain rule*, next slide.)
- How much does Y *reduce* our uncertainty about X? (Answer: *mutual information*, in 2 slides.)

Joint and conditional entropy will give us the two key identities of the rest of this lecture.

Chain Rule: The Workhorse Identity

The chain rule **factors** joint uncertainty into stages. Both directions are valid:

$$H(X,Y) \;=\; H(X) + H(Y \mid X) \;=\; H(Y) + H(X \mid Y).$$

Symmetry consequence: $H(X) - H(X|Y) = H(Y) - H(Y|X)$. (This will be *mutual information* in 2 slides.)

Worked example: two coin flips $X, Y \in \{0,1\}$, both fair, with $Y = X$ XOR Z where $Z \sim \text{Bern}(0.1)$ independent of X.

- $H(X) = 1$ (fair coin).
- $H(Y|X) = H(Z) = h(0.1) \approx 0.469$ (binary entropy of 0.1).
- $H(X,Y) = H(X) + H(Y|X) \approx 1.469$ bits.

Sanity check the other direction: Y marginally is also fair ($\Rightarrow H(Y) = 1$), and by symmetry $H(X|Y) \approx 0.469$. Same total. ✓

n-variable generalization: $H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i \mid X_1, \ldots, X_{i-1})$.
This is exactly how autoregressive language models factor the joint probability of a sequence.

Mutual Information: Three Equivalent Definitions

Question: How much does knowing X **reduce** our uncertainty about Y?

Form 1 — uncertainty reduction:

$$I(X;Y) = H(X) - H(X \mid Y).$$

Knowing Y shrinks our uncertainty about X by this much.

Mutual Information: Three Equivalent Definitions

Question: How much does knowing X **reduce** our uncertainty about Y?

Form 1 — uncertainty reduction:

$$I(X;Y) \;=\; H(X) - H(X \mid Y).$$

Knowing Y shrinks our uncertainty about X by this much.

Form 2 — by symmetry of the chain rule:

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

MI is symmetric in X and Y — unlike KL.

Mutual Information: Three Equivalent Definitions

Question: How much does knowing X **reduce** our uncertainty about Y?

Form 1 — uncertainty reduction:

$$I(X;Y) \;=\; H(X) - H(X \mid Y).$$

Knowing Y shrinks our uncertainty about X by this much.

Form 2 — by symmetry of the chain rule:

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

MI is symmetric in X and Y — unlike KL.

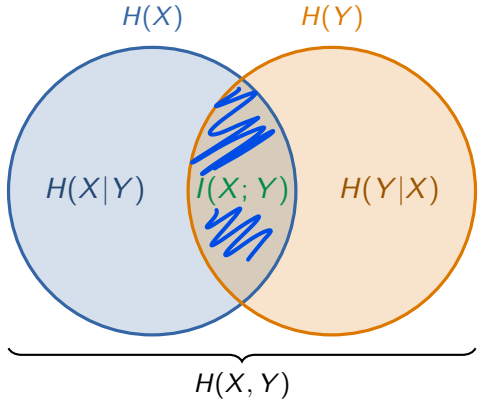
Form 3 — KL from independence:

$$I(X;Y) \;=\; D_{\mathrm{KL}}\big(p(x,y) \,\|\, p(x)\,p(y)\big) \;=\; \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}.$$

MI = how far the joint is from the product of marginals.

$I(X;Y) = 0$ **iff** $X \perp\!\!\!\perp Y$. The larger I, the more X and Y "know" about each other.

The Information Diagram



$I(X;Y) = H(X) + H(Y) - H(X,Y)$ — the overlap = shared information.

Properties of Mutual Information

1. Non-negative: $I(X;Y) \geq 0$ (since $D_{\mathrm{KL}} \geq 0$)

2. Zero iff independent: $I(X;Y) = 0 \Leftrightarrow X \perp\!\!\!\perp Y$

3. Symmetric: $I(X;Y) = I(Y;X)$ (unlike KL!)

4. Self-information: $I(X;X) = H(X)$

5. Bounded: $I(X;Y) \leq \min\{H(X), H(Y)\}$ (for discrete RVs)

6. Invariant: under invertible smooth transformations of X or Y

Conditioning reduces entropy ("information can't hurt"):

$H(X \mid Y) \leq H(X)$, with equality iff $X \perp\!\!\!\perp Y$.

Proof: $0 \leq I(X;Y) = H(X) - H(X \mid Y)$.

Worked Example: Computing MI

Joint distribution $p(x, y)$:

$X \backslash Y$	y_1	y_2	y_3	y_4
x_1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
x_2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
x_3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
x_4	$\frac{1}{4}$	0	0	0

Marginals: $p_X = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, $p_Y = (\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$.

Group the joint by value (read off the table):

- $1 \times \frac{1}{4}$, $2 \times \frac{1}{8}$, $6 \times \frac{1}{16}$, $4 \times \frac{1}{32}$

$H(X) = \log_2 4 = 2$ (uniform).

$H(Y) = \frac{1}{2}(1) + \frac{1}{4}(2) + 2 \cdot \frac{1}{8}(3) = \frac{7}{4}$.

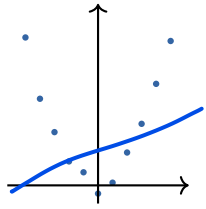
$H(X, Y) = \frac{1}{4}(2) + 2 \cdot \frac{1}{8}(3) + 6 \cdot \frac{1}{16}(4) + 4 \cdot \frac{1}{32}(5)$
$= \frac{4}{8} + \frac{6}{8} + \frac{12}{8} + \frac{5}{8} = \frac{27}{8}$.

Mutual information:

$$I(X; Y) = 2 + \frac{7}{4} - \frac{27}{8} = \boxed{\frac{3}{8} \text{ bits}}$$

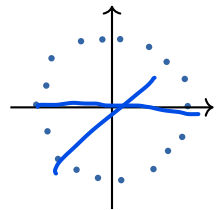
MI vs Correlation: Why MI Is More General

Pearson correlation only measures **linear** dependence. MI captures **any** dependence.



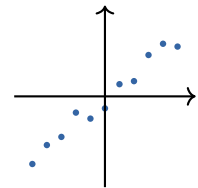
$Y = X^2 + \varepsilon$

$\rho \approx 0, \; I > 0$



(X, Y) **uniform on circle**

$\rho \approx 0, \; I > 0$



Linear

$\rho > 0, \; I > 0$

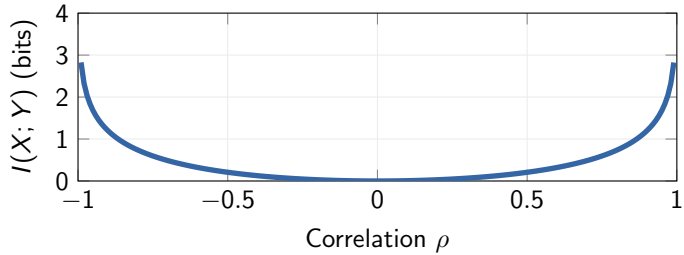
Correlation = 0 does NOT mean independence. MI detects **any** dependence — linear, nonlinear, or otherwise.

MI for Correlated Gaussians

For any $(X, Y) \sim N\left(\boldsymbol{\mu}, \begin{pmatrix} \sigma_X^2 & \rho\,\sigma_X\sigma_Y \\ \rho\,\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}\right)$:

$$I(X;Y) = -\frac{1}{2}\log_2(1-\rho^2).$$

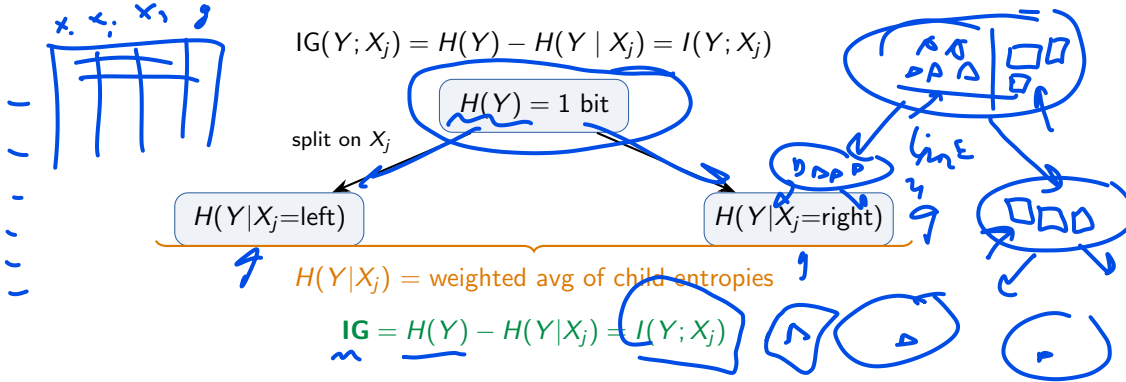
Why: $h(X) + h(Y) - h(X,Y) = \frac{1}{2}\log(2\pi e\sigma_X^2 \cdot 2\pi e\sigma_Y^2) - \frac{1}{2}\log\big((2\pi e)^2\sigma_X^2\sigma_Y^2(1-\rho^2)\big)$. All but ρ cancels.



$\rho = 0 \Rightarrow I = 0$ (Gaussian: uncorrelated = independent). $|\rho| \to 1 \Rightarrow I \to \infty$.

MI in ML: Information Gain in Decision Trees

At each node of a decision tree, we pick the feature that **maximizes information gain**:



Split on the feature with highest MI with the target.
Exactly the criterion in ID3 and C4.5. CART uses Gini by default but can use entropy. Feature selection: rank features by $I(X_j; Y)$, pick top ones.

The Information Theory Toolbox for ML

Concept	Formula	ML Role
Entropy $H(p)$	$-\sum p \log p$	Baseline risk, impurity
Cross-entropy $H(p\|\|q)$	$-\sum p \log q$	**Loss function** (log-loss)
KL divergence D_{KL}	$\sum p \log \frac{p}{q}$	MLE, variational inference
Mutual info $I(X;Y)$	$D_{\mathrm{KL}}(p_{xy} \| p_x p_y)$	Feature selection, info gain
Forward KL	$D_{\mathrm{KL}}(p \| q_\theta)$	Supervised learning
Reverse KL	$D_{\mathrm{KL}}(q_\phi \| p)$	Variational inference
MaxEnt	$\arg\max H$ s.t. constraints	Gaussian, exp family

Key takeaway: Information theory is not just "theory" — it directly gives us the loss functions, optimization objectives, and model selection criteria we use every day in ML.

Questions?