MEAM/EE/CBE/MSE 4600: AI for Science and Engineering

A Primer on Probability and Information Theory

Instructor: Paris Perdikaris

Lecture 1: Foundations of Probability Theory

1.1 What is Probability?

Learning Objectives:

- Understand the Bayesian vs. frequentist interpretations of probability
- Master the fundamental rules of probability
- Recognize probability as the language of uncertainty in machine learning

Two Interpretations of Probability

> "Probability theory is nothing but common sense reduced to calculation." — Pierre Laplace, 1812

Frequentist Interpretation: Probabilities represent long-run frequencies of repeatable events. For example, saying a fair coin has $P(\text{heads}) = 0.5$ means that if we flip the coin many times, approximately half the outcomes will be heads.

Bayesian Interpretation: Probability quantifies our *uncertainty* or *degree of belief* about events. This interpretation is fundamentally tied to **information** rather than repeated trials.

Why Bayesian for ML? The Bayesian interpretation allows us to:

- Model one-off events (e.g., "probability that a neural network will generalize well")
- Update beliefs as new data arrives
- Quantify uncertainty in predictions (crucial for scientific applications!)

Types of Uncertainty

Understanding uncertainty is critical for building reliable AI systems:

Type	Also Called	Source	Example
Epistemic	Model uncertainty	Ignorance of underlying mechanisms	Limited training data
Aleatoric	Data uncertainty	Intrinsic randomness	Measurement noise

Why This Matters for ML:

- Epistemic uncertainty *can* be reduced by collecting more data
- Aleatoric uncertainty *cannot* be reduced
- Active learning strategies should target high epistemic uncertainty regions

---

1.2 Basic Rules of Probability

Events and Probabilities

An **event** A is a proposition that either holds or doesn't. We denote:

- $\Pr(A)$: probability that A is true, where $0 \leq \Pr(A) \leq 1$
- $\Pr(\bar{A}) = 1 - \Pr(A)$: probability that A does not hold

The Three Fundamental Rules

1. Product Rule (Conjunction):

$$\Pr(A, B) = \Pr(A \cap B) = \Pr(A) \cdot \Pr(B|A)$$

If A and B are **independent**:

$$\Pr(A, B) = \Pr(A) \cdot \Pr(B)$$

2. Sum Rule (Disjunction):

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$

If A and B are **mutually exclusive**:

$$\Pr(A \cup B) = \Pr(A) + \Pr(B)$$

3. Conditional Probability:

$$\Pr(B|A) = \frac{\Pr(A, B)}{\Pr(A)}, \quad \text{provided } \Pr(A) > 0$$

Conditional Independence

Events A and B are **conditionally independent** given C if:

$$\Pr(A, B|C) = \Pr(A|C) \cdot \Pr(B|C)$$

Notation: $A \perp B \mid C$

> **ML Connection:** This is the foundation of graphical models and Bayesian networks. Many ML models assume conditional independence to make computation tractable.

1.3 Random Variables and Distributions

Discrete Random Variables

A **discrete random variable** X takes values from a finite or countable set $\mathcal{X}$.

The **probability mass function (pmf)** is:

$$p(x) \triangleq \Pr(X = x)$$

Properties:

- $0 \leq p(x) \leq 1$ for all x
- $\sum_{x \in \mathcal{X}} p(x) = 1$

Continuous Random Variables

For a **continuous random variable** $X \in \mathbb{R}$:

Cumulative Distribution Function (CDF):

$$F(x) \triangleq \Pr(X \leq x)$$

Probability Density Function (PDF):

$$p(x) = \frac{dF(x)}{dx}$$

Properties:

- $p(x) \geq 0$ (but can be > 1!)

- $\int_{-\infty}^{\infty} p(x)\,dx = 1$
- $\Pr(a < X \leq b) = \int_a^b p(x)\,dx = F(b) - F(a)$

For an infinitesimal interval:

$$\Pr(x < X \leq x + dx) \approx p(x) \cdot dx$$

1.4 Moments of Distributions

Expected Value (Mean)

For continuous X:

$$\mathbb{E}[X] \triangleq \int_{\mathcal{X}} x \cdot p(x)\,dx = \mu$$

For discrete X:

$$\mathbb{E}[X] \triangleq \sum_{x \in \mathcal{X}} x \cdot p(x)$$

Linearity of Expectation:

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$$

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i]$$

Variance and Standard Deviation

$$\mathbb{V}[X] \triangleq \mathbb{E}[(X - \mu)^2] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \sigma^2$$

Useful identity:

$$\mathbb{E}[X^2] = \sigma^2 + \mu^2$$

Scaling property:

$$\mathbb{V}[aX + b] = a^2\mathbb{V}[X]$$

For **independent** random variables:

$$\mathbb{V}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{V}[X_i]$$

Mode

The **mode** is the value with highest probability:

$$x^* = \arg\max_x p(x)$$

> **Warning:** For multimodal distributions, the mode may not be representative!

1.5 Bayes' Rule — The Heart of Probabilistic ML

The Formula

$$p(H|Y) = \frac{p(H) \cdot p(Y|H)}{p(Y)}$$

Or in words:

$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

Components:

Term	Symbol	Meaning
Prior	$p(H)$	Belief before seeing data
Likelihood	$p(Y	H)$
Evidence/Marginal	$p(Y) = \sum_h p(H=h)p(Y	H=h)$
Posterior	$p(H	Y)$

Example: COVID-19 Testing

Suppose:

- Test sensitivity (true positive rate): $\Pr(Y = 1|H = 1) = 0.875$
- Test specificity (true negative rate): $\Pr(Y = 0|H = 0) = 0.975$
- Disease prevalence (prior): $\Pr(H = 1) = 0.10$

If you test positive, what's the probability you have COVID?

$$\Pr(H = 1|Y = 1) = \frac{0.875 \times 0.1}{0.875 \times 0.1 + 0.025 \times 0.9} = \frac{0.0875}{0.1100} \approx 0.795$$

Only 79.5% chance of infection despite a positive test!

If prevalence drops to 1%:

$$\Pr(H = 1|Y = 1) = \frac{0.875 \times 0.01}{0.875 \times 0.01 + 0.025 \times 0.99} \approx 0.26$$

Now only 26% chance! This counterintuitive result shows the importance of **base rates**.

1.6 Key Distributions for ML

Bernoulli and Binomial

Bernoulli (single binary trial):

$$\text{Ber}(y|\theta) = \theta^y(1-\theta)^{1-y}, \quad y \in \{0, 1\}$$

Binomial (N trials, counting successes):

$$\text{Bin}(k|N, \theta) = \binom{N}{k}\theta^k(1-\theta)^{N-k}$$

The Sigmoid Function

To predict binary outcomes from unconstrained inputs, we use the **sigmoid** (logistic) function:

$$\sigma(a) = \frac{1}{1+e^{-a}} = \frac{e^a}{1+e^a}$$

Key Properties:

- Maps $\mathbb{R} \to (0, 1)$
- $\sigma(-a) = 1 - \sigma(a)$
- Derivative: $\frac{d\sigma}{da} = \sigma(a)(1 - \sigma(a))$
- Inverse (logit): $\sigma^{-1}(p) = \log \frac{p}{1-p}$

> **ML Connection:** This is the activation function in logistic regression and the output layer of binary classifiers!

Categorical and Multinomial

Categorical (generalization of Bernoulli to C classes):

$$\text{Cat}(y|\boldsymbol{\theta}) = \prod_{c=1}^{C} \theta_c^{\mathbb{I}(y=c)}$$

where $\sum_{c=1}^{C} \theta_c = 1$

The Softmax Function

Generalizes sigmoid to multiple classes:

$$\text{softmax}(\mathbf{a})_c = \frac{e^{a_c}}{\sum_{c'=1}^{C} e^{a_{c'}}}$$

Properties:

- Maps $\mathbb{R}^C \to [0, 1]^C$ with $\sum_c \text{softmax}(\mathbf{a})_c = 1$
- Temperature scaling: As $T \to 0$, becomes argmax (winner-take-all)
- $\text{softmax}(\mathbf{a} + c) = \text{softmax}(\mathbf{a})$ for any constant c

> **Numerical Stability (Log-Sum-Exp Trick):** $$\log \sum_c e^{a_c} = m + \log \sum_c e^{a_c - m}, \quad m = \max_c a_c$$

Gaussian (Normal) Distribution

The **univariate Gaussian**:

$$\mathcal{N}(y|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

Why is the Gaussian so important?

1. Characterized by just mean and variance
2. Central Limit Theorem: sums of i.i.d. variables → Gaussian
3. Maximum entropy distribution for given mean and variance
4. Mathematically convenient (conjugate priors, closed-form updates)

95% Confidence Interval: $\mu \pm 1.96\sigma \approx \mu \pm 2\sigma$

Lecture 2: Multivariate Distributions and Linear Gaussian Models

2.1 Joint Distributions and Marginalization

For two random variables X and Y:

Joint distribution: $p(x, y)$

Marginal distribution (sum rule):

$$p(x) = \sum_y p(x, y) \quad \text{or} \quad p(x) = \int p(x, y)\, dy$$

Conditional distribution:

$$p(y|x) = \frac{p(x, y)}{p(x)}$$

Product rule:

$$p(x, y) = p(x) \cdot p(y|x)$$

Chain rule (for D variables): $$p(x_1, x_2, \ldots, x_D) = p(x_1) \cdot p(x_2|x_1) \cdot p(x_3|x_1, x_2) \cdots p(x_D|x_{1:D-1})$$

> **ML Connection:** This factorization is the basis of autoregressive models like GPT!

2.2 Covariance and Correlation

Covariance

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

Covariance Matrix for vector $\mathbf{x} \in \mathbb{R}^D$:

$$\boldsymbol{\Sigma} = \mathrm{Cov}[\mathbf{x}] = \mathbb{E}[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top]$$

Important identity:

$$\mathbb{E}[\mathbf{x}\mathbf{x}^\top] = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top$$

Correlation

Pearson correlation coefficient:

$$\rho_{XY} = \mathrm{corr}[X, Y] = \frac{\mathrm{Cov}[X, Y]}{\sqrt{\mathbb{V}[X] \cdot \mathbb{V}[Y]}}$$

Properties:

- $-1 \leq \rho \leq 1$
- $\rho = \pm 1$ iff $Y = aX + b$ (perfect linear relationship)
- $\rho = 0$ means uncorrelated (but **not necessarily independent**!)

> **Warning:** Correlation measures *linear* dependence only. Two variables can be perfectly dependent yet have $\rho = 0$.

Independence vs. Uncorrelated

Condition	Meaning
Independent	$p(X, Y) = p(X)p(Y)$ — strongest
Uncorrelated	$\mathrm{Cov}[X, Y] = 0$ — weaker

Important: Independent $\Rightarrow$ Uncorrelated, but Uncorrelated $\not\Rightarrow$ Independent!

Example: Let $X \sim \text{Uniform}(-1, 1)$ and $Y = X^2$. Then Y is completely determined by X, yet $\text{Cov}[X, Y] = 0$.

2.3 The Multivariate Gaussian (MVN)

Definition

$$\mathcal{N}(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right)$$

where:

- $\boldsymbol{\mu} \in \mathbb{R}^D$ is the mean vector
- $\boldsymbol{\Sigma} \in \mathbb{R}^{D \times D}$ is the (symmetric, positive definite) covariance matrix
- $|\boldsymbol{\Sigma}|$ is the determinant of $\boldsymbol{\Sigma}$

Mahalanobis Distance

The quadratic form in the exponent defines the **Mahalanobis distance**:

$$d_M(\mathbf{y}, \boldsymbol{\mu}) = \sqrt{(\mathbf{y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})}$$

This is Euclidean distance in a transformed coordinate system that accounts for correlations.

> **ML Connection:** Contours of constant probability are ellipsoids. Understanding the geometry helps with PCA and Gaussian mixture models.

Types of Covariance Structures

Type	Form	Parameters	Shape
Full	$\boldsymbol{\Sigma}$ arbitrary	$D(D+1)/2$	Any ellipse orientation
Diagonal	$\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_D^2)$	D	Axis-aligned ellipse
Spherical/Isotropic	$\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$	1	Circle/sphere

---

2.4 Gaussian Conditioning and Marginalization

The Key Results (Memorize These!)

Let $\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix}$ be jointly Gaussian with:

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

Marginals:

$$p(\mathbf{y}_1) = \mathcal{N}(\mathbf{y}_1 | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$$

Conditionals:

$$\boxed{p(\mathbf{y}_1 | \mathbf{y}_2) = \mathcal{N}(\mathbf{y}_1 | \boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2})}$$

where:

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_2)$$

$$\boldsymbol{\Sigma}_{1|2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

Key Insights:

1. Conditional mean is a **linear function** of the observed variable
2. Conditional covariance is **independent** of the observed value
3. The matrix $\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}$ acts like a "regression coefficient"

> **ML Connection:** This is the foundation of Gaussian processes, Kalman filters, and optimal linear estimation!

2.5 Linear Gaussian Systems (Bayes Rule for Gaussians)

Setup

Let:

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z) \quad \text{(prior)}$$

$$p(\mathbf{y}|\mathbf{z}) = \mathcal{N}(\mathbf{y}|\mathbf{W}\mathbf{z} + \mathbf{b}, \boldsymbol{\Sigma}_y) \quad \text{(likelihood)}$$

Posterior

$$\boxed{p(\mathbf{z}|\mathbf{y}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_{z|y}, \boldsymbol{\Sigma}_{z|y})}$$

where:

$$\boldsymbol{\Sigma}_{z|y}^{-1} = \boldsymbol{\Sigma}_z^{-1} + \mathbf{W}^\top \boldsymbol{\Sigma}_y^{-1} \mathbf{W}$$

$$\boldsymbol{\mu}_{z|y} = \boldsymbol{\Sigma}_{z|y} \left[\mathbf{W}^\top \boldsymbol{\Sigma}_y^{-1} (\mathbf{y} - \mathbf{b}) + \boldsymbol{\Sigma}_z^{-1} \boldsymbol{\mu}_z \right]$$

Marginal Likelihood

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{W}\boldsymbol{\mu}_z + \mathbf{b}, \boldsymbol{\Sigma}_y + \mathbf{W}\boldsymbol{\Sigma}_z\mathbf{W}^\top)$$

Example: Inferring a Hidden Scalar

Given noisy observations $y_1, \ldots, y_N$ of unknown z with:

- Prior: $p(z) = \mathcal{N}(z|\mu_0, \tau_0^{-1})$ (precision τ_0)
- Likelihood: $p(y_n|z) = \mathcal{N}(y_n|z, \tau_y^{-1})$ (precision τ_y)

Posterior after N observations:

$$p(z|\mathbf{y}) = \mathcal{N}(z|\mu_N, \tau_N^{-1})$$

where:

$$\tau_N = \tau_0 + N\tau_y \quad \text{(precision adds!)}$$

$$\mu_N = \frac{\tau_0\mu_0 + N\tau_y\bar{y}}{\tau_0 + N\tau_y} = \frac{\tau_0}{\tau_0 + N\tau_y}\mu_0 + \frac{N\tau_y}{\tau_0 + N\tau_y}\bar{y}$$

Interpretation: Posterior mean is a **weighted average** of prior mean and sample mean, weighted by their respective precisions.

2.6 Transformations of Random Variables

Change of Variables (Scalar)

If $y = f(x)$ is monotonic with inverse $x = g(y)$:

$$p_Y(y) = p_X(g(y)) \left| \frac{dg}{dy} \right|$$

Multivariate Change of Variables

For invertible $\mathbf{y} = f(\mathbf{x})$:

$$p_Y(\mathbf{y}) = p_X(g(\mathbf{y})) \left| \det \left(\frac{\partial g}{\partial \mathbf{y}} \right) \right|$$

where the Jacobian $\frac{\partial g}{\partial \mathbf{y}}$ measures how volumes change under the transformation.

> **ML Connection:** This is the foundation of **normalizing flows** — a powerful class of generative models!

Linear Transformations

For $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$:

$$\mathbb{E}[\mathbf{y}] = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}$$

$$\text{Cov}[\mathbf{y}] = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top$$

Central Limit Theorem

If $X_1, X_2, \ldots, X_N$ are i.i.d. with mean μ and variance σ^2, then:

$$\frac{\bar{X}_N - \mu}{\sigma / \sqrt{N}} \xrightarrow{d} \mathcal{N}(0, 1) \quad \text{as } N \to \infty$$

Implication: Sample means become approximately Gaussian regardless of the original distribution!

Lecture 3: Information Theory for Machine Learning

3.1 Motivation: Information as the Foundation of ML

Why Information Theory?

- Provides a principled way to measure uncertainty
- Defines optimal compression and communication limits
- Gives rise to fundamental ML concepts: cross-entropy loss, KL divergence
- Connects probability to coding theory and data compression

> "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point." — Claude Shannon, 1948

3.2 Entropy: Measuring Uncertainty

Definition

The **entropy** of a discrete random variable X with pmf p is:

$$H(X) \triangleq -\sum_{x \in \mathcal{X}} p(x) \log p(x) = \mathbb{E}_p[-\log p(X)]$$

Units:

- Base 2 logarithm: bits
- Natural logarithm: nats

Intuition: Entropy as "Average Surprise"

Define **surprise** (or self-information) of event x as:

$$I(x) \triangleq -\log p(x)$$

- Rare events have high surprise: $p(x) \to 0 \Rightarrow I(x) \to \infty$
- Certain events have zero surprise: $p(x) = 1 \Rightarrow I(x) = 0$

Entropy is the expected surprise:

$$H(X) = \mathbb{E}[I(X)] = \mathbb{E}[-\log p(X)]$$

Properties of Entropy

1. **Non-negativity:** $H(X) \geq 0$, with equality iff X is deterministic
2. **Maximum entropy:** For discrete X with K outcomes: $$H(X) \leq \log K with equality iff $p(x) = 1/K$ (uniform distribution)
3. **Additivity for independent variables:** $$H(X, Y) = H(X) + H(Y) \quad \text{if } X \perp Y

Examples

Fair coin ($K = 2$, uniform):

$$H(X) = -\left(\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}\right) = \log 2 = 1 \text{ bit}$$

Biased coin ($p(\text{heads}) = 0.9$):

$$H(X) = -(0.9\log 0.9 + 0.1\log 0.1) \approx 0.47 \text{ bits}$$

The more predictable, the lower the entropy!

Differential Entropy (Continuous Variables)

For continuous X with pdf $p(x)$:

$$h(X) \triangleq -\int p(x)\log p(x)\,dx$$

Warning: Differential entropy can be negative!

Gaussian: $h(X) = \frac{1}{2}\log(2\pi e\sigma^2)$ nats

> **ML Connection:** Among all distributions with fixed variance, the Gaussian has maximum entropy. This is why Gaussian assumptions are often "least informative."

3.3 Conditional Entropy

$$H(Y|X) = \mathbb{E}_{p(x)}\left[H(Y|X=x)\right] = -\sum_{x,y} p(x,y)\log p(y|x)$$

Chain rule for entropy:

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

Interpretation: Total uncertainty = uncertainty in X + remaining uncertainty in Y after observing X

Property: $H(Y|X) \leq H(Y)$ — conditioning reduces entropy (on average)

3.4 Kullback-Leibler (KL) Divergence

Definition

The **KL divergence** (relative entropy) from distribution p to q is:

$$D_{\mathrm{KL}}(p\|q) \triangleq \sum_x p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_p \left[\log \frac{p(X)}{q(X)} \right]$$

For continuous distributions:

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

Properties

1. **Non-negativity:** $D_{\mathrm{KL}}(p\|q) \geq 0$, with equality iff $p = q$
2. **Asymmetry:** $D_{\mathrm{KL}}(p\|q) \neq D_{\mathrm{KL}}(q\|p)$ in general
3. **Not a true distance:** Violates triangle inequality

Interpretations

Information-theoretic: Extra bits needed to encode samples from p using a code optimized for q

Statistical: Amount of information lost when q is used to approximate p

Bayesian: KL divergence measures how much our beliefs change after seeing data

Forward vs. Reverse KL

Name	Form	Behavior
Forward KL	$D_{\mathrm{KL}}(p\|q)$	Mass-covering: q tries to cover all of p
Reverse KL	$D_{\mathrm{KL}}(q\|p)$	Mode-seeking: q focuses on high-probability regions of p

ML Connection:

- Maximum likelihood ≈ minimizing forward KL from data to model
- Variational inference ≈ minimizing reverse KL from model to true posterior

---

3.5 Cross-Entropy

Definition

$$H(p, q) \triangleq -\sum_{x} p(x) \log q(x) = \mathbb{E}_p[-\log q(X)]$$

Relationship to KL Divergence

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

Since $H(p)$ is constant w.r.t. q:

$$\arg\min_q D_{\mathrm{KL}}(p\|q) = \arg\min_q H(p, q)$$

Cross-Entropy Loss in ML

For classification with true labels y and predicted probabilities $\hat{p}(y|x)$:

Binary cross-entropy:

$$\mathcal{L} = -[y \log \hat{p} + (1 - y) \log(1 - \hat{p})]$$

Categorical cross-entropy:

$$\mathcal{L} = -\sum_{c=1}^{C} y_c \log \hat{p}_c$$

where $\mathbf{y}$ is one-hot encoded.

Key Insight: Minimizing cross-entropy loss = maximum likelihood estimation = minimizing KL divergence to the empirical distribution!

3.6 Mutual Information

Definition

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

Equivalent Formulations

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

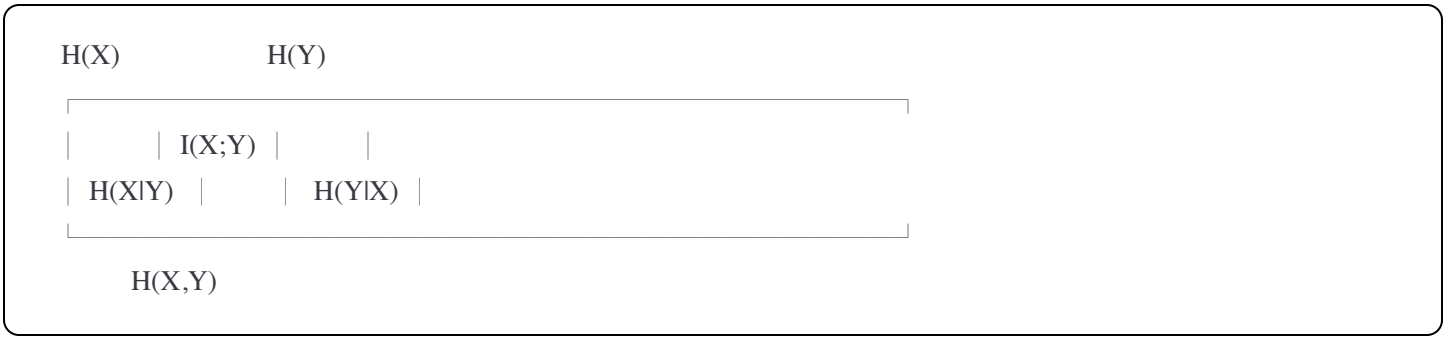
Properties

1. **Symmetry:** $I(X;Y) = I(Y;X)$
2. **Non-negativity:** $I(X;Y) \geq 0$, with equality iff $X \perp Y$
3. **Bounded:** $I(X;Y) \leq \min(H(X), H(Y))$

Interpretation

- Mutual information measures **shared information** between X and Y
- It quantifies **how much knowing one variable reduces uncertainty about the other**
- Unlike correlation, MI captures **all** statistical dependencies (linear and nonlinear)

Venn Diagram View



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

> **ML Connection:**
>
> - Feature selection: choose features with high MI with labels
> - InfoGAN: maximize MI between latent codes and generated samples
> - Contrastive learning: maximize MI between different views of same data

3.7 Connecting Information Theory to ML Loss Functions

Maximum Likelihood and Cross-Entropy

Given data $\{x_1, \ldots, x_N\}$ from true distribution p_{data}:

$$\arg\max_\theta \frac{1}{N} \sum_{i=1}^{N} \log p_\theta(x_i) = \arg\min_\theta H(p_{\text{data}}, p_\theta)$$

Why Log-Likelihood?

1. **Asymptotic optimality:** MLE is consistent and efficient
2. **Information-theoretic:** Minimizes KL divergence to true distribution
3. **Computational:** Log converts products to sums (numerical stability)

Negative Log-Likelihood as a Loss

For regression with Gaussian likelihood:

$$-\log p(y|x, \theta) = \frac{(y - f_\theta(x))^2}{2\sigma^2} + \text{const}$$

→ **Mean squared error** is negative log-likelihood under Gaussian assumption!

For classification with categorical likelihood:

$$-\log p(y|x, \theta) = -\sum_c y_c \log \text{softmax}(f_\theta(x))_c$$

→ **Cross-entropy loss** is negative log-likelihood under categorical assumption!

3.8 Summary: The Information Theory Toolkit for ML

Concept	Formula	ML Application
Entropy	$H(X) = -\mathbb{E}[\log p(X)]$	Uncertainty quantification
Cross-entropy	$H(p, q) = -\mathbb{E}_p[\log q]$	Classification loss

Concept	Formula	ML Application
KL divergence	$D_{\mathrm{KL}}(p\|\|q) = H(p,q) - H(p)$	Variational inference
Mutual information	$I(X;Y) = H(X) - H(X \| Y)$	

Practice Problems

Lecture 1

1. **Bayes' Rule Application:** A spam filter has 99% sensitivity and 98% specificity. If 10% of emails are spam, what's the probability an email flagged as spam is actually spam?
2. **Independence:** Show that if $X \sim \mathrm{Uniform}(-1, 1)$ and $Y = X^2$, then $\mathrm{Cov}[X, Y] = 0$.
3. **Softmax:** Compute $\mathrm{softmax}([1, 2, 3])$ and verify the sum equals 1.

Lecture 2

4. **Gaussian Conditioning:** For the bivariate Gaussian with $\mu = [0, 0]^\top$ and $\Sigma = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$, find $p(Y_1 | Y_2 = 1)$.
5. **Posterior Mean:** Given prior $p(z) = \mathcal{N}(0, 1)$ and observation $y = 2$ with likelihood $p(y|z) = \mathcal{N}(z, 0.5)$, compute the posterior mean and variance.

Lecture 3

6. **Entropy:** Compute the entropy of a die with probabilities $(1/2, 1/4, 1/8, 1/16, 1/32, 1/32)$.
7. **KL Divergence:** Compute $D_{\mathrm{KL}}(\mathrm{Ber}(0.9) \| \mathrm{Ber}(0.5))$.
8. **Mutual Information:** For jointly Gaussian (X, Y) with correlation ρ, show that $I(X;Y) = -\frac{1}{2}\log(1 - \rho^2)$.

Key Takeaways

1. **Probability is the language of uncertainty** — essential for making decisions under incomplete information.
2. **Bayes' rule is the principled way to update beliefs** — combines prior knowledge with new evidence.
3. **Gaussians are mathematically convenient** — closed under conditioning, marginalization, and linear transformations.

4. **Information theory provides principled loss functions** — cross-entropy loss is not arbitrary but optimal under certain assumptions.
5. **KL divergence measures information loss** — fundamental to variational methods and generative modeling.
6. **Mutual information captures all dependencies** — more powerful than correlation for measuring relationships.

These notes are designed to give you the mathematical foundations needed for the machine learning techniques we'll cover in the rest of the course. The probability concepts here will appear repeatedly when we discuss neural networks, Bayesian inference, and generative models.