Day 1: Probability Axioms — Building the Foundation

The Mathematical Bedrock of Uncertainty Reasoning in AI

1. The Fundamental Question

How do we mathematically formalize uncertainty to build reliable AI systems?

Probability theory provides the essential language for reasoning about uncertainty in AI models. Kolmogorov's three axioms serve as the unshakable foundation for all probabilistic reasoning.

2. Kolmogorov's Axioms: The Mathematical Foundation

For any probability space $(\Omega, \mathcal{F}, P)$ where:

- Ω: Sample space (all possible outcomes)
- $\mathcal{F}$: Event space (set of all measurable events)
- P: Probability measure

The Three Axioms of Probability

1. **Non-negativity:** For any event $A \subseteq \Omega$,
$$P(A) \geq 0$$
2. **Normalization:** The probability of the entire sample space is 1,
$$P(\Omega) = 1$$
3. **Countable Additivity:** For any countable sequence of disjoint events $A_1, A_2, \ldots,$
$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

3. Derived Rules and Key Consequences

From these simple axioms, we derive all probability rules used in AI:

3.1. Complement Rule

$$P(A^c) = 1 - P(A)$$

Proof: Since $A \cup A^c = \Omega$ and $A \cap A^c = \emptyset$, by axioms 2 and 3:

$$1 = P(\Omega) = P(A \cup A^c) = P(A) + P(A^c)$$

3.2. Union Rule for Non-Disjoint Events

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

3.3. Monotonicity

If $A \subseteq B$, then $P(A) \leq P(B)$

3.4. Bounds on Probability

$$0 \leq P(A) \leq 1 \quad \text{for any event } A$$

Numerical Example: Fair Six-Sided Die

Let $\Omega = \{1, 2, 3, 4, 5, 6\}$ with $P(\{i\}) = \frac{1}{6}$ for $i = 1, \ldots, 6$

- **Axiom 1:** $P(\{2\}) = \frac{1}{6} \geq 0$
- **Axiom 2:** $P(\Omega) = 6 \times \frac{1}{6} = 1$
- **Axiom 3:** $P(\{1, 2, 3\}) = P(\{1\}) + P(\{2\}) + P(\{3\}) = \frac{3}{6}$

4. Why This Matters in AI Systems

Critical Applications in AI

- **Bayesian Networks:** All probabilistic graphical models rely on these axioms
- **Classification:** Ensures output probabilities sum to 1 (softmax normalization)
- **Generative Models:** VAEs and GANs require valid probability distributions
- **Reinforcement Learning:** Policy and value functions must satisfy probability constraints
- **Uncertainty Quantification:** Proper calibration depends on valid probability assignments

5. Practical Implications for Model Design

5.1. Output Layer Design

In classification networks, the softmax function ensures normalization:

$$P(y = i|x) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

This guarantees $\sum_{i=1}^{K} P(y = i|x) = 1$ and $P(y = i|x) \geq 0$ for all i.

5.2. Bayesian Inference

Posterior probabilities must satisfy:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \quad \text{where } P(D) = \int P(D|\theta)P(\theta)d\theta$$

5.3. Loss Function Design

Cross-entropy loss assumes valid probability distributions:

$$L = -\sum_{i=1}^{K} y_i \log(\hat{y}_i)$$

where $\sum y_i = 1$, $\sum \hat{y}_i = 1$, and $y_i, \hat{y}_i \geq 0$.

6. Common Pitfalls and How to Avoid Them

Modeling Mistakes to Avoid

- **Negative Probabilities:** Using linear output without softmax
- **Non-normalized Outputs:** Forgetting to ensure probabilities sum to 1
- **Invalid Confidence Scores:** Outputting values outside [0,1] as confidence
- **Inconsistent Updates:** Violating additivity when updating beliefs

7. Advanced Topics Building on Axioms

7.1. Conditional Probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{when } P(B) > 0$$

7.2. Bayes' Theorem

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

7.3. Law of Total Probability

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i) \quad \text{for partition } \{B_1, \ldots, B_n\}$$

8. Practical Exercises for Mastery

Hands-On Practice Problems

1. Prove that $P(\emptyset) = 0$ using the axioms
2. If $P(A) = 0.3$ and $P(B) = 0.4$ with $P(A \cap B) = 0.1$, find $P(A \cup B)$
3. A neural network outputs logits [2.0, 1.0, 0.5]. Convert to probabilities and verify they satisfy the axioms
4. Design a function that takes any real-valued scores and converts them to valid probabilities
5. Explain why a classifier outputting [0.7, 0.6] for two classes violates the axioms

9. Real-World AI Examples

Axioms in Production Systems

- **Spam Detection:** $P(\text{spam}) + P(\text{not spam}) = 1$
- **Medical Diagnosis:** Disease probabilities must be non-negative and normalized
- **Autonomous Vehicles:** Action probabilities must form valid distribution
- **Recommendation Systems:** Item preference scores converted to probabilities

10. Key Insight: Mathematical Rigor Enables Reliability

The probability axioms are not just abstract mathematics—they are the engineering constraints that ensure our AI systems produce coherent, interpretable, and reliable results. Every successful probabilistic AI model implicitly respects these rules.

- **Validation:** Always check that your model outputs satisfy $0 \leq P \leq 1$ and $\sum P = 1$
- **Debugging:** Probability violations often indicate model architecture issues
- **Interpretability:** Valid probabilities enable meaningful confidence estimates
- **Comparability:** Standardized probability scales allow model performance comparison

Next: Conditional Probability and Bayes' Theorem

Tomorrow we'll explore how to update beliefs in light of new evidence—the cornerstone of Bayesian reasoning in AI.