PROBABILITY THEORY REVIEW

CSCI-B555

Predrag Radivojac

DEPARTMENT OF COMPUTER SCIENCE AND INFORMATICS
INDIANA UNIVERSITY, BLOOMINGTON

Spring, 2015
Axioms of Probability

\( \Omega = \) sample space, all outcomes of the experiment
\( \mathcal{F} = \) event space, set of subsets of \( \Omega \)

\( \Omega \) and \( \mathcal{F} \) must be non-empty

If the following conditions hold:

1. \( A \in \mathcal{F} \implies A^c \in \mathcal{F} \)

2. \( A_1, A_2, \ldots \in \mathcal{F} \implies \bigcup_{i=1}^{\infty} A_i \in \mathcal{F} \)

\( \mathcal{F} \) is called a sigma field (sigma algebra)

\( (\Omega, \mathcal{F}) \) = a measurable space
AXIOMS OF PROBABILITY

$(\Omega, \mathcal{F}) = \text{a measurable space}$

Any function $P : \mathcal{F} \rightarrow [0, 1]$ such that

1. $P(\Omega) = 1$

2. $\forall A, B \in \mathcal{F}$ and $A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$

is called a probability measure (probability distribution)

$(\Omega, \mathcal{F}, P) = \text{a probability space}$
CONSEQUENCES OF THE AXIOMS OF PROBABILITY

\((\Omega, \mathcal{F}, P) = \text{a probability space}\)

1. \(P(\emptyset) = 0\)

2. \(P(A^c) = 1 - P(A)\)

3. \(P(A) = \sum_{i=1}^{k} P(A \cap B_i), \text{ where } \{B_i\}_{i=1}^{k} \text{ is a partition of } \Omega\)

4. \(P(A \cup B) = P(A) + P(B) - P(A \cap B)\)

... and everything else.
**Sample Spaces**

\[ \Omega \]

- **Discrete (countable)**
  - \( \Omega = \{1, 2, 3, 4, 5, 6\} \)
  - \( \Omega = \mathbb{N} \)

- **Continuous (uncountable)**
  - \( \Omega = [0, 1] \)
  - \( \Omega = \mathbb{R} \)

Typically: \( \mathcal{F} = \mathcal{P}(\Omega) \)

Typically: \( \mathcal{F} = \mathcal{B}(\Omega) \)

\( \Omega = [0, 1] \cup \{2\} = \text{mixed space} \)

- **Power set**
- **Borel field**
(\Omega, \mathcal{F}) = \text{a measurable space}

**Example:** \[ \Omega = \{0, 1\} \]
\[ \mathcal{F} = \{\emptyset, \{0\}, \{1\}, \Omega\} \]

\[ P(A) = \begin{cases} 
1 - \alpha & A = \{0\} \\
\alpha & A = \{1\} \\
0 & A = \emptyset \\
1 & A = \Omega 
\end{cases} \quad \alpha \in [0, 1] \]

**How can we choose** \( P \) **in practice?**

Clearly, we cannot do it arbitrarily.

**How can we satisfy all constraints?**
PROBABILITY MASS FUNCTIONS

\[ \Omega = \text{discrete sample space} \]
\[ \mathcal{F} = \mathcal{P}(\Omega) \]

Probability mass function:

1. \( p : \Omega \rightarrow [0, 1] \)
2. \( \sum_{\omega \in \Omega} p(\omega) = 1 \)

The probability of any event \( A \in \mathcal{F} \) is defined as

\[ P(A) = \sum_{\omega \in A} p(\omega) \]
PMFs YOU HEARD ABOUT

Bernoulli distribution: \[ \Omega = \{S, F\} \quad \alpha \in (0, 1) \]

\[ p(\omega) = \begin{cases} 
\alpha & \omega = S \\
1 - \alpha & \omega = F 
\end{cases} \]

Alternatively, \( \Omega = \{0, 1\} \)

\[ p(k) = \alpha^k \cdot (1 - \alpha)^{1-k} \quad \forall k \in \Omega \]
PMFs You Heard About

**Binomial distribution:**

\[
\Omega = \{0, 1, \ldots, n\} \quad \alpha \in (0, 1)
\]

\[
p(k) = \binom{n}{k} \alpha^k (1 - \alpha)^{n-k} \quad \forall k \in \Omega
\]

\[
n = 30
\]
PMFs you heard about

Poisson distribution:

\[ \Omega = \{0, 1, \ldots\} \quad \lambda \in (0, \infty) \]

\[ p(k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad \forall k \in \Omega \]
PMFs YOU HEARD ABOUT

Geometric distribution:

$$\Omega = \{1, 2, \ldots\} \quad \alpha \in (0, 1)$$

$$p(k) = (1 - \alpha)^{k-1} \alpha$$

$$\forall k \in \Omega$$
\[ \Omega = \text{continuous sample space} \]
\[ \mathcal{F} = \mathcal{B}(\Omega) \]

**Probability density function:**

1. \( p : \Omega \rightarrow [0, \infty) \)
2. \( \int_{\Omega} p(\omega) d\omega = 1 \)

The probability of any event \( A \in \mathcal{F} \) is defined as

\[
P(A) = \int_{A} p(\omega) d\omega.
\]
Uniform distribution: $\Omega = [a, b]$

$$p(\omega) = \frac{1}{b - a} \quad \forall \omega \in [a, b]$$
PDFs You Heard About

Gaussian distribution:

\[ p(\omega) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\omega-\mu)^2} \]

\[ \Omega = \mathbb{R} \quad \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+ \]

\[ \forall \omega \in \mathbb{R} \]
PDFs You Heard About

Exponential distribution: \( \Omega = [0, \infty) \quad \lambda > 0 \)

\[ p(\omega) = \lambda e^{-\lambda \omega} \quad \forall \omega \geq 0 \]
PMFs vs. PDFs

\( \Omega = \text{discrete sample space} \)

Consider a singleton event \( \{\omega\} \in \mathcal{F} \), where \( \omega \in \Omega \)

\[ P(\{\omega\}) = p(\omega) \]

\( \Omega = \text{continuous sample space} \)

Consider an interval event \( A = [x, x + \Delta x] \), where \( \Delta \) is small

\[ P(A) = \int_{x}^{x+\Delta x} p(\omega) \, d\omega \]

\[ \approx p(x) \Delta x \]
**MULTIDIMENSIONAL PMFs**

\[ \Omega = \Omega_1 \times \Omega_2 \times \ldots \times \Omega_k \]
\[ \mathcal{F} = \mathcal{P}(\Omega) \]

**Probability mass function:**

1. \( p : \Omega_1 \times \Omega_2 \times \ldots \times \Omega_k \rightarrow [0, 1] \)
2. \( \sum_{\omega_1 \in \Omega_1} \ldots \sum_{\omega_k \in \Omega_k} p(\omega_1, \omega_2, \ldots, \omega_k) = 1 \)

The probability of any event \( A \in \mathcal{F} \) is defined as

\[ P(A) = \sum_{\omega \in A} p(\omega) \]
\[ \omega = (\omega_1, \omega_2, \ldots, \omega_k) \]
\( \Omega = \mathbb{R}^k \)
\( \mathcal{F} = \mathcal{B}(\mathbb{R})^k \)

**Probability density function:**

1. \( p : \mathbb{R}^k \rightarrow [0, \infty) \)
2. \( \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(\omega_1, \omega_2, \ldots, \omega_k) \, d\omega_1 \cdots d\omega_k = 1 \)

The probability of any event \( A \in \mathcal{F} \) is defined as

\[
P(A) = \int_{\omega \in A} p(\omega) \, d\omega.
\]
\( \omega = (\omega_1, \omega_2, \ldots, \omega_k) \)
**MULTIDIMENSIONAL GAUSSIAN**

\[ \Omega = \mathbb{R}^k \]
\[ \mathcal{F} = \mathcal{B}(\mathbb{R})^k \]

\[ \mu \in \mathbb{R}^k \]
\[ \Sigma = \text{positive definite } k\text{-by-}k \text{ matrix} \]
\[ |\Sigma| = \text{determinant of } \Sigma \]

\[
\begin{align*}
p(\omega) &= \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} \exp \left( -\frac{1}{2} (\omega - \mu)^T \Sigma^{-1} (\omega - \mu) \right) \\
\end{align*}
\]

\[ \mu = (0, 0) \]
\[ \Sigma = \begin{bmatrix} 1 & .75 \\ .75 & 1 \end{bmatrix} \]

\[ k = 2 \]
ELEMENTARY CONDITIONAL PROBABILITIES

\((\Omega, \mathcal{F}, P)\) = a probability space

\(B =\) event that already occurred

The probability that any event \(A \in \mathcal{F}\) has also occurred is

\[
P(A|B) = \frac{P(A \cap B)}{P(B)}
\]

where \(P(B) > 0\).

Bayes’ rule:

\[
P(A|B) = \frac{P(B|A)P(A)}{P(B)}
\]
(Ω, ℱ, P) = a probability space

**Chain rule**

\[ P(A_1 \cap A_2 \ldots \cap A_k) = P(A_1)P(A_2|A_1)\ldots P(A_k|A_1 \cap A_2 \ldots \cap A_{k-1}) \]

where \( \{A_i\}_{i=1}^k \) is a collection of \( k \) events
SUM RULE, PRODUCT RULE

$(\Omega, \mathcal{F}, P) = \text{a probability space}$

**Sum rule:**

$$P(A) = \sum_{i=1}^{k} P(A \cap B_i)$$

where $\{B_i\}_{i=1}^{k}$ is a partition of $\Omega$

**Product rule:**

$$P(A \cap B) = P(A|B) \cdot P(B)$$

where $P(B) > 0$
INDEPENDENCE OF EVENTS

$(Ω, ℱ, P) = \text{a probability space}$

Events $A$ and $B$ are independent if:

$$P(A \cap B) = P(A) \cdot P(B)$$

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

$$P(A \cap B|C) = P(A|C) \cdot P(B|C)$$

What if we had multiple events?
INDEPENDENCE EXAMPLES

$$(\Omega, \mathcal{F}, P) = \text{a probability space}$$
RANDOM VARIABLES

\[ (\Omega, \mathcal{F}, P) \]

\[ A = \{ \omega \in \Omega : Musician(\omega) = \text{yes} \} \]

\[ \Omega = \text{voltage at any time } t \]

\[ \text{Analog} \rightarrow \text{A/D converter} \rightarrow \text{Digital} \]

- Age: 35
- Height: 1.85m
- Weight: 75kg
- IQ: 104
- Likes sports: Yes
- Smokes: No
- Marital status: Single
- Occupation: Musician

- Age: 26
- Height: 1.75m
- Weight: 79kg
- IQ: 103
- Likes sports: Yes
- Smokes: No
- Marital status: Divorced
- Occupation: Athlete
Example: three consecutive (fair) coin tosses
\[ X = \text{the number of heads in the first toss} \]
\[ Y = \text{the number of heads in all three tosses} \]
Find the probability spaces after the transformations.

Where is the probability space \((\Omega, \mathcal{F}, P)\)?
Where is the randomness?

\[ \Omega = \{\text{HHH, HHT, HTH, HTT, TTH, THT, TTT}\} \]
\[ \mathcal{F} = \mathcal{P}(\Omega) \]
\[ P = ? \]

\[ P(\Omega) = 1 \]
\[ P(\{\text{HHH, TTT}\}) = \frac{2}{8} \]
\[ \vdots \]
$X : \Omega \rightarrow \{0, 1\}$
$Y : \Omega \rightarrow \{0, 1, 2, 3\}$

<table>
<thead>
<tr>
<th>$\omega$</th>
<th>HHH</th>
<th>HHT</th>
<th>HTH</th>
<th>HTT</th>
<th>THH</th>
<th>THT</th>
<th>TTH</th>
<th>TTT</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X(\omega)$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$Y(\omega)$</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

What are the probability spaces $(\Omega_X, \mathcal{F}_X, P_X)$ and $(\Omega_Y, \mathcal{F}_Y, P_Y)$?

Where does the randomness come from?
RANDOM VARIABLE: FORMAL DEFINITION

\((\Omega, \mathcal{F}, P) = \) a probability space

Random variable:

1. \(X : \Omega \rightarrow \Omega_X\)

2. \(\forall A \in \mathcal{B}(\Omega_X)\) it holds that \(\{\omega : X(\omega) \in A\} \in \mathcal{F}\)

It follows that:

\[P_X(A) = P(\{\omega : X(\omega) \in A\})\]
**Discrete Random Variable**

\((\Omega, \mathcal{F}, P) = \) a discrete probability space

Probability mass function (pmf):

\[
p_X(x) = P_X(\{x\})
= P(\{\omega : X(\omega) = x\}) \quad \forall x \in \Omega_X
\]

The probability of an event \(A\):

\[
P_X(A) = \sum_{x \in A} p_X(x) \quad \forall A \subseteq \Omega_X
\]

\[
P(\{\omega : X(\omega) \in A\})
\]
CONTINUOUS RANDOM VARIABLE

Cumulative distribution function (cdf):

\[ F_X(t) = P_X \{x : x \leq t\} \]
\[ = P_X ((-\infty, t]) \]
\[ = P (X \leq t) \]
\[ = P (\{\omega : X(\omega) \leq t\}) \]

Probability density function (pdf), if it exists:

\[ p_X(x) = \frac{dF_X(t)}{dt} \bigg|_{t=x} \]
CONTINUOUS RANDOM VARIABLE

If the probability density function (pdf) exists:

\[ F_X(t) = \int_{-\infty}^{t} p_X(x) \, dx \]

The probability of an event \( A = (a, b) \):

\[
P_X((a, b]) = \int_{a}^{b} p_X(x) \, dx
= F_X(b) - F_X(a)
\]

\( P(a < X \leq b) \)
(\Omega, \mathcal{F}, P) = \text{a discrete probability space}

**Joint probability distribution:**

\[
p_{XY}(x, y) = P(X = x, Y = y) = P(\{\omega : X(\omega) = x\} \cap \{\omega : Y(\omega) = y\})
\]

Extend to \(k\)-D vector \(X = (X_1, X_2, \ldots, X_k)\)

**Marginal probability distribution:**

\[
p_{X_i}(x_i) = \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_k} p_X(x_1, \ldots, x_k)
\]
JOINT AND MARGINAL DISTRIBUTIONS

\((\Omega, \mathcal{F}, P) = (\mathbb{R}^k, \mathcal{B}(\mathbb{R})^k, P_X)\) = a continuous probability space

Joint probability distribution:

\[
F_X(t) = P_X \left( \{x : x_i \leq t_i, i = 1 \ldots k\} \right) \\
= P (X_1 \leq t_1, X_2 \leq t_2 \ldots)
\]

\[
p_X(x) = \frac{\partial^k}{\partial t_1 \cdots \partial t_k} F_X(t_1, \ldots, t_k) \bigg|_{t=x}
\]

(if it exists)

Marginal probability distribution:

\[
p_{X_i}(x_i) = \int_{x_1} x_{i-1} \int_{x_{i+1}} x_k p_X(x) \, dx_1 \cdots dx_{i-1} \, dx_{i+1} \cdots dx_k
\]
CONDITIONAL DISTRIBUTIONS

Conditional probability distribution:

\[ p_{Y|X}(y|x) = \frac{p_{XY}(x,y)}{p_X(x)} \]

The probability of an event \( A \), given that \( X = x \), is:

\[ P_{Y|X}(Y \in A|X = x) = \begin{cases} \sum_{y \in A} p_{Y|X}(y|x) & Y: \text{discrete} \\ \int_{y \in A} p_{Y|X}(y|x) \, dy & Y: \text{continuous} \end{cases} \]
**Chain Rule**

Conditional probability distribution:

\[
p(x_k | x_1, \ldots, x_{k-1}) = \frac{p(x_1, \ldots, x_k)}{p(x_1, \ldots, x_{k-1})}
\]

This leads to:

\[
p(x_1, \ldots, x_k) = p(x_1) \prod_{l=2}^{k} p(x_l | x_1, \ldots, x_{l-1})
\]
INDEPENDENCE OF RANDOM VARIABLES

$X$ and $Y$ are **independent** if:

$$p_{XY}(x, y) = p_X(x) \cdot p_Y(y)$$

$X$ and $Y$ are **conditionally independent** given $Z$ if:

$$p_{XY|Z}(x, y|z) = p_{X|Z}(x|z) \cdot p_{Y|Z}(y|z)$$

What if we had $k$ random variables?
(\Omega_X, \mathcal{B}(\Omega_X), P_X) = \text{ a probability space}

Consider a function \( f : \Omega_X \to \mathbb{C} \)

\[
E_x [f(x)] = \begin{cases} 
\sum_{x \in \Omega_X} f(x)p_X(x) & X : \text{discrete} \\
\int_{\Omega_X} f(x)p_X(x)dx & X : \text{continuous}
\end{cases}
\]
# Expectations You Know About

<table>
<thead>
<tr>
<th>$f(x)$</th>
<th>Symbol</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
<td>$E[X]$</td>
<td>Mean</td>
</tr>
<tr>
<td>$(x - E[X])^2$</td>
<td>$V[X]$</td>
<td>Variance</td>
</tr>
<tr>
<td>$x^k$</td>
<td>$E[X^k]$</td>
<td>$k$-th moment; $k \in \mathbb{N}$</td>
</tr>
<tr>
<td>$(x - E[X])^k$</td>
<td>$E[(x - E[X])^k]$</td>
<td>$k$-th central moment; $k \in \mathbb{N}$</td>
</tr>
<tr>
<td>$e^{tx}$</td>
<td>$M_X(t)$</td>
<td>Moment generating function</td>
</tr>
<tr>
<td>$e^{itx}$</td>
<td>$\varphi_X(t)$</td>
<td>Characteristic function</td>
</tr>
<tr>
<td>$\log \frac{1}{p_X(x)}$</td>
<td>$H(X)$</td>
<td>(Differential) entropy</td>
</tr>
<tr>
<td>$\log \frac{p_X(x)}{q(x)}$</td>
<td>$D(p_X</td>
<td></td>
</tr>
<tr>
<td>$\left( \frac{\partial}{\partial \theta} \log p_X(x</td>
<td>\theta) \right)^2$</td>
<td>$\mathcal{I}(\theta)$</td>
</tr>
</tbody>
</table>
Consider a function $f : \Omega_Y \rightarrow \mathbb{C}$

$$E_y [f(y)|x] = \begin{cases} \sum_{y \in \Omega_Y} f(y)p_{Y|X}(y|x) & Y : \text{discrete} \\ \int_{\Omega_Y} f(y)p_{Y|X}(y|x)dy & Y : \text{continuous} \end{cases}$$

$$E [Y|x] = \sum yp_{Y|X}(y|x)$$

$$E [Y|x] = \int yp_{Y|X}(y|x)dy$$

$\rightarrow$ Regression function!
EXPECTATIONS FOR TWO VARIABLES

Consider a function $f : \mathbb{R}^2 \rightarrow \mathbb{C}$

$$E_{x,y} [f(x, y)] = \begin{cases} 
\sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} f(x, y)p_{XY}(x, y) & X, Y : \text{discrete} \\
\int_{\Omega_X} \int_{\Omega_Y} f(x, y)p_{XY}(x, y) \, dx \, dy & X, Y : \text{continuous}
\end{cases}$$
## Expectations You Know About

<table>
<thead>
<tr>
<th>$f(x, y)$</th>
<th>Symbol</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>$(x - E[X])(y - E[Y])$</td>
<td>$\text{cov}(X, Y)$</td>
<td>Covariance</td>
</tr>
<tr>
<td>$\frac{(x-E[X])(y-E[Y])}{\sqrt{V[X]V[Y]}}$</td>
<td>$\text{corr}(X, Y)$</td>
<td>Correlation</td>
</tr>
<tr>
<td>$\log \frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}$</td>
<td>$I(X;Y)$</td>
<td>Mutual information</td>
</tr>
<tr>
<td>$\log \frac{1}{p_{XY}(x,y)}$</td>
<td>$H(X,Y)$</td>
<td>Joint entropy</td>
</tr>
<tr>
<td>$\log \frac{1}{p_{X</td>
<td>Y}(x</td>
<td>y)}$</td>
</tr>
</tbody>
</table>
MIXTURES OF DISTRIBUTIONS

Mixture model:

A set of $m$ probability distributions, $\{p_i(x)\}_{i=1}^{m}$

$$p(x) = \sum_{i=1}^{m} w_i p_i(x)$$

where $\mathbf{w} = (w_1, w_2, \ldots, w_m)$ and non-negative and

$$\sum_{i=1}^{m} w_i = 1$$
MIXTURES OF GAUSSIANS

Mixture of $m = 2$ Gaussian distributions:

$$w_1 = 0.75, \quad w_2 = 0.25$$

$$p(x) = \sum_{i=1}^{m} w_i p_i(x)$$

![Graph showing the mixture of two Gaussian distributions with $\mu = 1, \sigma = 1$ and $\mu = -2, \sigma = 0.75$, along with their mixture.](image-url)
**Graphical Representations**

**Bayesian Network:**

\[ p(x) = \prod_{i=1}^{k} p(x_i | x_{\text{Parents}(X_i)}) \]

**Factorization:**

\[ p(x, y, z) = p(x)p(y|x)p(z|x, y) \]
**Graphical Representations**

Bayesian Network: \( p(\mathbf{x}) = \prod_{i=1}^{k} p(x_i | \mathbf{x}_{\text{Parents}(X_i)}) \)

---

Factorization:

\[ p(x, y, z) = p(x)p(y|x)p(z|y) \]
Markov Network: \[ p(x_i | x_{-i}) = p(x_i | x_{N(X_i)}) \]

Factorization:

\[ p(x) = \frac{1}{Z} \prod_{C \in \mathcal{C}} \psi_C(x_C) \]