Artificial Intelligence

Probabilities

Marc Toussaint, WS 14/15
U Stuttgart
Objective Probability

The double slit experiment:
Objective Probability

The double slit experiment:

![Images of the double slit experiment results.](image-url)
Probability Theory

- Why do we need probabilities?
Probability Theory

- Why do we need probabilities?

  - Obvious: to express inherent (objective) stochasticity of the world
Probability Theory

• Why do we need probabilities?
  – Obvious: to express inherent (objective) stochasticity of the world

• But beyond this: (also in a “deterministic world”):
  – lack of knowledge!
  – hidden (latent) variables
  – expressing uncertainty
  – expressing information (and lack of information)
  – Subjective Probability

• Probability Theory: an information calculus
Outline

• Basic definitions
  – Random variables
  – joint, conditional, marginal distribution
  – Bayes’ theorem

• Probability distributions:
  – Binomial & Beta
  – Multinomial & Dirichlet
  – Conjugate priors
  – Gauss
  – Dirak & Particles

• Utilities, decision theory, entropy, KLD
Probability: Frequentist and Bayesian

• Frequentist probabilities are defined in the limit of an infinite number of trials
  \textit{Example:} “The probability of a particular coin landing heads up is 0.43”

• Bayesian (subjective) probabilities quantify degrees of belief
  \textit{Example:} “The probability of it raining tomorrow is 0.3”
  – Not possible to repeat “tomorrow”
Basic definitions
Probabilities & Sets

• Sample Space/domain $\Omega$, e.g. $\Omega = \{1, 2, 3, 4, 5, 6\}$

• Probability $P : A \subset \Omega \mapsto [0, 1]$
eq, $P(\{1\}) = \frac{1}{6}$, $P(\{4\}) = \frac{1}{6}$, $P(\{2, 5\}) = \frac{1}{3}$,

• Axioms: $\forall A, B \subseteq \Omega$
  – Nonnegativity $P(A) \geq 0$
  – Additivity $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$
  – Normalization $P(\Omega) = 1$

• Implications
  $0 \leq P(A) \leq 1$
  $P(\emptyset) = 0$
  $A \subseteq B \Rightarrow P(A) \leq P(B)$
  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  $P(\Omega \setminus A) = 1 - P(A)$
Probabilities & Random Variables

- For a random variable $X$ with discrete domain $\text{dom}(X) = \Omega$ we write:
  $\forall x \in \Omega : 0 \leq P(X = x) \leq 1$
  $\sum_{x \in \Omega} P(X = x) = 1$

Example: A dice can take values $\Omega = \{1, .., 6\}$. $X$ is the random variable of a dice throw. $P(X = 1) \in [0, 1]$ is the probability that $X$ takes value 1.

- A bit more formally: a random variable is a map from a measurable space to a domain (sample space) and thereby introduces a probability measure on the domain ("assigns a probability to each possible value")
Probabilty Distributions

- \( P(X = 1) \in \mathbb{R} \) denotes a specific probability
- \( P(X) \) denotes the probability distribution (function over \( \Omega \))
Probabilty Distributions

- $P(X = 1) \in \mathbb{R}$ denotes a specific probability
- $P(X)$ denotes the probability distribution (function over $\Omega$)

Example: A dice can take values $\Omega = \{1, 2, 3, 4, 5, 6\}$. By $P(X)$ we describe the full distribution over possible values $\{1, \ldots, 6\}$. These are 6 numbers that sum to one, usually stored in a *table*, e.g.: $[\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}]$

- In implementations we typically represent distributions over discrete random variables as tables (arrays) of numbers

- Notation for summing over a RV:
  In equation we often need to sum over RVs. We then write
  \[ \sum_X P(X) \cdots \]
  as shorthand for the explicit notation $\sum_{x \in \text{dom}(X)} P(X = x) \cdots$
Joint distributions

Assume we have two random variables $X$ and $Y$

- Definitions:
  - Joint: $P(X, Y)$
  - Marginal: $P(X) = \sum_Y P(X, Y)$
  - Conditional: $P(X|Y) = \frac{P(X, Y)}{P(Y)}$

The conditional is normalized: $\forall Y : \sum_X P(X|Y) = 1$

- $X$ is independent of $Y$ iff: $P(X|Y) = P(X)$
  (table thinking: all columns of $P(X|Y)$ are equal)
Joint distributions

\begin{itemize}
  \item joint: \( P(X, Y) \)
  \item marginal: \( P(X) = \sum_Y P(X, Y) \)
  \item conditional: \( P(X | Y) = \frac{P(X, Y)}{P(Y)} \)
\end{itemize}

Implications of these definitions:

\begin{itemize}
  \item Product rule: \( P(X, Y) = P(X | Y) \ P(Y) = P(Y | X) \ P(X) \)
  \item Bayes’ Theorem: \( P(X | Y) = \frac{P(Y | X) \ P(X)}{P(Y)} \)
\end{itemize}
Bayes’ Theorem

\[ P(X|Y) = \frac{P(Y|X) \cdot P(X)}{P(Y)} \]

posterior = \frac{\text{likelihood} \cdot \text{prior}}{\text{normalization}}
Multiple RVs:

- Analogously for \( n \) random variables \( X_{1:n} \) (stored as a rank \( n \) tensor)
  
  **Joint:** \( P(X_{1:n}) \)
  
  **Marginal:** \( P(X_1) = \sum_{X_{2:n}} P(X_{1:n}) \),
  
  **Conditional:** \( P(X_1|X_{2:n}) = \frac{P(X_{1:n})}{P(X_{2:n})} \)

- \( X \) is *conditionally independent* of \( Y \) given \( Z \) iff:
  
  \[
P(X|Y, Z) = P(X|Z)
  \]

- Product rule and Bayes’ Theorem:
  
  \[
P(X_{1:n}) = \prod_{i=1}^{n} P(X_i|X_{i+1:n})
  \]
  
  \[
P(X_1|X_{2:n}) = \frac{P(X_2|X_{1},X_{3:n}) P(X_1|X_{3:n})}{P(X_2|X_{3:n})}
  \]
  
  \[
P(X,Y|Z) = \frac{P(X,Z|Y) P(Y)}{P(Z)}
  \]
Probability distributions
Reference


http://research.microsoft.com/en-us/um/people/cmbishop/prml/
Bernoulli & Binomial

- We have a binary random variable \( x \in \{0, 1\} \) (i.e. \( \text{dom}(x) = \{0, 1\} \))

  The **Bernoulli** distribution is parameterized by a single scalar \( \mu \),

  \[
  P(x = 1 \mid \mu) = \mu, \quad P(x = 0 \mid \mu) = 1 - \mu
  \]

  \[
  \text{Bern}(x \mid \mu) = \mu^x (1 - \mu)^{1-x}
  \]

- We have a data set of random variables \( D = \{x_1, \ldots, x_n\} \), each \( x_i \in \{0, 1\} \). If each \( x_i \sim \text{Bern}(x_i \mid \mu) \) we have

  \[
  P(D \mid \mu) = \prod_{i=1}^{n} \text{Bern}(x_i \mid \mu) = \prod_{i=1}^{n} \mu^{x_i} (1 - \mu)^{1-x_i}
  \]

  \[
  \argmax_{\mu} \log P(D \mid \mu) = \argmax_{\mu} \sum_{i=1}^{n} x_i \log \mu + (1 - x_i) \log(1 - \mu) = \frac{1}{n} \sum_{i=1}^{n} x_i
  \]

- The **Binomial distribution** is the distribution over the count \( m = \sum_{i=1}^{n} x_i \)

  \[
  \text{Bin}(m \mid n, \mu) = \binom{n}{m} \mu^m (1 - \mu)^{n-m}, \quad \binom{n}{m} = \frac{n!}{(n-m)! \ m!}
  \]
Beta

How to express uncertainty over a Bernoulli parameter $\mu$

- The $Beta$ distribution is over the interval $[0, 1]$, typically the parameter $\mu$ of a Bernoulli:

$$Beta(\mu | a, b) = \frac{1}{B(a, b)} \mu^{a-1}(1 - \mu)^{b-1}$$

with mean $\langle \mu \rangle = \frac{a}{a+b}$ and mode $\mu^* = \frac{a-1}{a+b-2}$ for $a, b > 1$

- The crucial point is:
  - Assume we are in a world with a “Bernoulli source” (e.g., binary bandit), but don’t know its parameter $\mu$
  - Assume we have a prior distribution $P(\mu) = Beta(\mu | a, b)$
  - Assume we collected some data $D = \{x_1, .., x_n\}$, $x_i \in \{0, 1\}$, with counts $a_D = \sum_i x_i$ of $[x_i = 1]$ and $b_D = \sum_i (1 - x_i)$ of $[x_i = 0]$
  - The posterior is

$$P(\mu | D) = \frac{P(D | \mu)}{P(D)} P(\mu) \propto Bin(D | \mu) \, Beta(\mu | a, b) \propto \mu^{a_D}(1 - \mu)^{b_D} \mu^{a-1}(1 - \mu)^{b-1} = \mu^{a-1+a_D} (1 - \mu)^{b-1+b_D} = Beta(\mu | a + a_D, b + b_D)$$
**Beta**

*The prior is $\text{Beta}(\mu \mid a, b)$, the posterior is $\text{Beta}(\mu \mid a + a_D, b + b_D)$*

- Conclusions:
  - The semantics of $a$ and $b$ are counts of $[x_i = 1]$ and $[x_i = 0]$, respectively
  - The Beta distribution is conjugate to the Bernoulli (explained later)
  - With the Beta distribution we can represent beliefs (state of knowledge) about uncertain $\mu \in [0, 1]$ and know how to update this belief given data
Beta

\[ a = 0.1 \]
\[ b = 0.1 \]

\[ a = 1 \]
\[ b = 1 \]

\[ a = 2 \]
\[ b = 3 \]

\[ a = 8 \]
\[ b = 4 \]

taken from Bishop
Multinomial

- We have an integer random variable $x \in \{1, \ldots, K\}$
  The probability of a single $x$ can be parameterized by $\mu = (\mu_1, \ldots, \mu_K)$:

\[
P(x = k \mid \mu) = \mu_k
\]

with the constraint $\sum_{k=1}^{K} \mu_k = 1$ (probabilities need to be normalized)

- We have a data set of random variables $D = \{x_1, \ldots, x_n\}$, each $x_i \in \{1, \ldots, K\}$. If each $x_i \sim P(x_i \mid \mu)$ we have

\[
P(D \mid \mu) = \prod_{i=1}^{n} \mu_{x_i} = \prod_{i=1}^{n} \prod_{k=1}^{K} \mu_{k}^{[x_i = k]} = \prod_{k=1}^{K} \mu_{k}^{m_k}
\]

where $m_k = \sum_{i=1}^{n} [x_i = k]$ is the count of $[x_i = k]$. The ML estimator is

\[
\arg\max_{\mu} \log P(D \mid \mu) = \frac{1}{n} (m_1, \ldots, m_K)
\]

- The Multinomial distribution is this distribution over the counts $m_k$

\[
\text{Mult}(m_1, \ldots, m_K \mid n, \mu) \propto \prod_{k=1}^{K} \mu_{k}^{m_k}
\]
How to express uncertainty over a Multinomial parameter $\mu$

- The *Dirichlet* distribution is over the $K$-simplex, that is, over $\mu_1, \ldots, \mu_K \in [0, 1]$ subject to the constraint $\sum_{k=1}^{K} \mu_k = 1$:

$$\text{Dir}(\mu | \alpha) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k - 1}$$

It is parameterized by $\alpha = (\alpha_1, \ldots, \alpha_K)$, has mean $\langle \mu_i \rangle = \frac{\alpha_i}{\sum_j \alpha_j}$ and mode $\mu_i^* = \frac{\alpha_i - 1}{\sum_j \alpha_j - K}$ for $\alpha_i > 1$.

- The crucial point is:
  - Assume we are in a world with a “Multinomial source” (e.g., an integer bandit), but don’t know its parameter $\mu$
  - Assume we have a *prior* distribution $P(\mu) = \text{Dir}(\mu | \alpha)$
  - Assume we collected some data $D = \{x_1, \ldots, x_n\}$, $x_i \in \{1, \ldots, K\}$, with counts $m_k = \sum_i [x_i = k]$
  - The posterior is

$$P(\mu | D) = \frac{P(D | \mu) \ P(\mu)}{P(D)} \propto \text{Mult}(D | \mu) \ \text{Dir}(\mu | a, b)$$

$$\propto \prod_{k=1}^{K} \mu_k^{m_k} \ \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} = \prod_{k=1}^{K} \mu_k^{\alpha_k - 1 + m_k}$$

$$\text{Dir}(\mu | a + \alpha)$$
Dirichlet

The prior is $\text{Dir}(\mu \mid \alpha)$, the posterior is $\text{Dir}(\mu \mid \alpha + m)$

• Conclusions:
  – The semantics of $\alpha$ is the counts of $[x_i = k]$ 
  – The Dirichlet distribution is conjugate to the Multinomial 
  – With the Dirichlet distribution we can represent beliefs (state of knowledge) about uncertain $\mu$ of an integer random variable and know how to update this belief given data
Dirichlet

Illustrations for $\alpha = (0.1, 0.1, 0.1)$, $\alpha = (1, 1, 1)$ and $\alpha = (10, 10, 10)$:

taken from Bishop
Motivation for Beta & Dirichlet distributions

- **Bandits:**
  - If we have binary [integer] bandits, the Beta [Dirichlet] distribution is a way to represent and update beliefs.
  - The belief space becomes discrete: The parameter $\alpha$ of the prior is continuous, but the posterior updates live on a discrete “grid” (adding counts to $\alpha$).
  - We can in principle do belief planning using this.

- **Reinforcement Learning:**
  - Assume we know that the world is a finite-state MDP, but do not know its transition probability $P(s' \mid s, a)$. For each $(s, a)$, $P(s' \mid s, a)$ is a distribution over the integer $s'$.
  - Having a separate Dirichlet distribution for each $(s, a)$ is a way to represent our belief about the world, that is, our belief about $P(s' \mid s, a)$.
  - We can in principle do belief planning using this $\rightarrow$ *Bayesian Reinforcement Learning*.

- **Dirichlet distributions** are also used to model texts (word distributions in text), images, or mixture distributions in general.
Conjugate priors

- Assume you have data $D = \{x_1, .., x_n\}$ with likelihood

$$P(D \mid \theta)$$

that depends on an uncertain parameter $\theta$
Assume you have a prior $P(\theta)$

- The prior $P(\theta)$ is **conjugate** to the likelihood $P(D \mid \theta)$ iff the posterior

$$P(\theta \mid D) \propto P(D \mid \theta) \, P(\theta)$$

is in the *same distribution class* as the prior $P(\theta)$

- Having a conjugate prior is very convenient, because then you know how to update the belief given data
## Conjugate priors

<table>
<thead>
<tr>
<th>likelihood</th>
<th>conjugate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binomial $\text{Bin}(D \mid \mu)$</td>
<td>Beta $\text{Beta}(\mu \mid a, b)$</td>
</tr>
<tr>
<td>Multinomial $\text{Mult}(D \mid \mu)$</td>
<td>Dirichlet $\text{Dir}(\mu \mid \alpha)$</td>
</tr>
<tr>
<td>Gauss $\mathcal{N}(x \mid \mu, \Sigma)$</td>
<td>Gauss $\mathcal{N}(\mu \mid \mu_0, A)$</td>
</tr>
<tr>
<td>1D Gauss $\mathcal{N}(x \mid \mu, \lambda^{-1})$</td>
<td>Gamma $\text{Gam}(\lambda \mid a, b)$</td>
</tr>
<tr>
<td>$n$D Gauss $\mathcal{N}(x \mid \mu, \Lambda^{-1})$</td>
<td>Wishart $\text{Wish}(\Lambda \mid W, \nu)$</td>
</tr>
<tr>
<td>$n$D Gauss $\mathcal{N}(x \mid \mu, \Lambda^{-1})$</td>
<td>Gauss-Wishart $\mathcal{N}(\mu \mid \mu_0, (\beta \Lambda)^{-1}) \text{ Wish}(\Lambda \mid W, \nu)$</td>
</tr>
</tbody>
</table>
Distributions over continuous domain

- Let $x$ be a continuous RV. The **probability density function (pdf)** $p(x) \in [0, \infty)$ defines the probability

\[
P(a \leq x \leq b) = \int_a^b p(x) \, dx \in [0, 1]
\]

The (cumulative) probability distribution $F(y) = P(x \leq y) = \int_{-\infty}^y dx \, p(x) \in [0, 1]$ is the cumulative integral with $\lim_{y \to \infty} F(y) = 1$

(In discrete domain: **probability distribution** and **probability mass function** $P(x) \in [0, 1]$ are used synonymously.)

- Two basic examples:
  - **Gaussian:** $\mathcal{N}(x \mid \mu, \Sigma) = \frac{1}{\sqrt{2\pi\Sigma}} \, e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1} (x-\mu)}$
  - **Dirac or $\delta$ (“point particle”)** $\delta(x) = 0$ except at $x = 0$, $\int \delta(x) \, dx = 1$
    $\delta(x) = \frac{\partial}{\partial x} H(x)$ where $H(x) = [x \geq 0] = \text{Heavyside step function}$
Gaussian distribution

• 1-dim: \( \mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(x-\mu)^2 / \sigma^2} \)

• \(n\)-dim Gaussian in normal form:

\[
\mathcal{N}(x \mid \mu, \Sigma) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left\{ -\frac{1}{2} (x - \mu)^\top \Sigma^{-1} (x - \mu) \right\}
\]

with mean \( \mu \) and covariance matrix \( \Sigma \). In canonical form:

\[
\mathcal{N}[x \mid a, A] = \frac{\exp\left\{ -\frac{1}{2} a^\top A^{-1} a \right\}}{\sqrt{2\pi|A^{-1}|}} \exp\left\{ -\frac{1}{2} x^\top A x + x^\top a \right\}
\]

with precision matrix \( A = \Sigma^{-1} \) and coefficient \( a = \Sigma^{-1} \mu \) (and mean \( \mu = A^{-1} a \)).

• Gaussian identities: see

http://ipvs.informatik.uni-stuttgart.de/mlr/marc/notes/gaussians.pdf
Motivation for Gaussian distributions

- Gaussian Bandits
- Control theory, Stochastic Optimal Control
- State estimation, sensor processing, Gaussian filtering (Kalman filtering)
- Machine Learning
- etc
Particle Approximation of a Distribution

• We approximate a distribution $p(x)$ over a continuous domain $\mathbb{R}^n$

• A particle distribution $q(x)$ is a weighed set $S = \{(x^i, w^i)\}_{i=1}^N$ of $N$ particles
  – each particle has a “location” $x^i \in \mathbb{R}^n$ and a weight $w^i \in \mathbb{R}$
  – weights are normalized, $\sum_i w^i = 1$

$$q(x) := \sum_{i=1}^N w^i \delta(x - x^i)$$

where $\delta(x - x^i)$ is the $\delta$-distribution.
Particle Approximation of a Distribution

Histogram of a particle representation:

- $i=100$
- $i=500$
- $i=1000$
- $i=5000$
Motivation for particle distributions

- Numeric representation of “difficult” distributions
  - Very general and versatile
  - But often needs many samples
- Distributions over games (action sequences), sample based planning, MCTS
- State estimation, particle filters
- etc
Utilities & Decision Theory

- Given a space of events $\Omega$ (e.g., outcomes of a trial, a game, etc) the utility is a function

$$U : \Omega \rightarrow \mathbb{R}$$

- The utility represents preferences as a single scalar – which is not always obvious (cf. multi-objective optimization)

- Decision Theory making decisions (that determine $p(x)$) that maximize expected utility

$$E\{U\}_p = \int x U(x) p(x)$$

- Concave utility functions imply risk aversion (and convex, risk-taking)
Entropy

- The neg-log ($- \log p(x)$) of a distribution reflects something like “error”:
  - neg-log of a Gaussian $\leftrightarrow$ squared error
  - neg-log likelihood $\leftrightarrow$ prediction error

- The $(- \log p(x))$ is the “optimal” coding length you should assign to a symbol $x$. This will minimize the expected length of an encoding

$$H(p) = \int_{x} p(x) [- \log p(x)]$$

- The entropy $H(p) = \mathbb{E}_{p(x)} \{- \log p(x)\}$ of a distribution $p$ is a measure of uncertainty, or lack-of-information, we have about $x$
Kullback-Leibler divergence*

- Assume you use a "wrong" distribution \( q(x) \) to decide on the coding length of symbols drawn from \( p(x) \). The expected length of a encoding is

\[
\int_x p(x) \left[ - \log q(x) \right] \geq H(p)
\]

- The difference

\[
D(p \mid\mid q) = \int_x p(x) \log \frac{p(x)}{q(x)} \geq 0
\]

is called Kullback-Leibler divergence

Proof of inequality, using the Jenson inequality:

\[
- \int_x p(x) \log \frac{q(x)}{p(x)} \geq - \log \int_x p(x) \frac{q(x)}{p(x)} = 0
\]
Some more continuous distributions*

Gaussian

\[ \mathcal{N}(x \mid a, A) = \frac{1}{|2\pi A|^{1/2}} e^{-\frac{1}{2} (x-a)^\top A^{-1} (x-a)} \]

Dirac or \( \delta \)

\[ \delta(x) = \frac{\partial}{\partial x} H(x) \]

Student’s t

(=Gaussian for \( \nu \to \infty \), otherwise heavy tails)

\[ p(x; \nu) \propto [1 + \frac{x^2}{\nu}]^{-\frac{\nu+1}{2}} \]

Exponential

(distribution over single event time)

\[ p(x; \lambda) = [x \geq 0] \lambda e^{-\lambda x} \]

Laplace

(“double exponential”)

\[ p(x; \mu, b) = \frac{1}{2b} e^{-|x-\mu|/b} \]

Chi-squared

\[ p(x; k) \propto [x \geq 0] x^{k/2-1} e^{-x/2} \]

Gamma

\[ p(x; k, \theta) \propto [x \geq 0] x^{k-1} e^{-x/\theta} \]