Advanced Probabilistic Machine Learning

Ralf Herbrich, Rainer Schlosser

Probability Theory

HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

Course Setup

- **Goal**: Enabled you to develop machine learning algorithms from scratch
  - We will pick the pace that helps you to get excited; please interrupt and ask questions!
- **Format**: We have one topic per week with a lecture and tutorial
  - **Lecture**: Tuesday, 9:15am – 10:45am (HS2)
  - **Tutorial**: Wednesday: 3:15pm – 4:45pm (HS3)
- **Assignment**: Five assignments to solve them (groups of two). They account for 40% of all points (8 points each)
  - Handed out every other week on Monday (starting 2nd week, October 20)
  - Each assignment has a theory and a practice part
- **Tutorial**: Supporting the material of the lecture and the assignments
  - In the tutorial, Rainer will solve similar exercises to the assignments with you
  - We will answer questions you have with the actual assignments
- **Exam (60 points)**: Counts for 60% of all points; 120 minutes long (**February 24**)

Course Material

- **Books**: All our material and communication will happen over Moodle
  - Bishop, C. Pattern Recognition and Machine Learning. Springer. 2006.
  - MacKay, D. Information Theory, Inference, and Learning Algorithms. CUP. 2003
- **Moodle**: Share our lecture slides, tutorials, solutions
  - **Location**: https://moodle.hpi.de/course/view.php?id=982
  - **Announcements**: https://moodle.hpi.de/mod/forum/view.php?id=33393
- **GitHub Repository**: Supporting material as well as code samples
  - **Location**: https://github.com/HPI-Artificial-Intelligence-Teaching/pml-wise2025
  - If you find mistakes, please submit issues and pull requests
- **GitHub Classrooms**: Used for all our assignments
  - If you do not have a GitHub account, please create one now
  - Find a team member as assignments are solved in groups of two
  - More details tomorrow in the first tutorial

Course Structure

Foundation

1. Probability Theory (Unit 1)
2. Linear Algebra (Unit 2)

Methods

3. Graphical Models (Unit 3)
4. Exact Inference (Unit 4)
5. Approximate Inference: Expectation Propagation (Unit 5)
6. Approximate Inference: Variational Inference (Unit 6)
7. Approximate Inference: Mixture Models (Unit 7)
8. Approximate Inference: Message Approximation (Unit 8)

Applications

8. Text: Topic Models (Unit 9)
9. Images: Conditional Markov Random Fields (Unit 10)
10. Policies: Reinforcement Learning (Unit 11)

Julia

- 2012 developed by Jeff Bezanson, Alan Edelman, Stefan Karpinski and Viral B. Shah at MIT
- Used for numerical and scientific computing with high performance
  - Execution speed is similar to C and FORTRAN
  - Hierarchical and parameterized type system as well as method overloading („multiple dispatching“) as central concepts
  - Native calls from C-(compiled) code possible (without wrappers)
- Unicode is efficiently supported (e.g., UTF-8)
- Alongside C, C++ and FORTRAN, the only programming language that has entered the “PetaFlop Club“


Jeff Bezanson (1981–)


Alan Edelman (1963 –)


Stefan Karpinski (1981–)


Viral Shah

Overview

1. Probability in Machine Learning
2. Probability Theory
3. Exponential Family Distributions

Overview

1. **Probability in Machine Learning**
2. Probability Theory
3. Exponential Family Distributions

What is Probability?

- **Weather forecast**: A meteorologist says

„Tomorrow, it is going to rain in Bangalore with 60%“

- **Two interpretations**:
  1. The meteorologist has analyzed all regions which have similar environmental conditions than Bangalore today. His (**objective**) estimate based on past data is that the procedure which predicts rain tomorrow is correct 60% of the time.
  2. The meteorologist *believes* that it is more likely that it rains tomorrow in Bangalore (than it is to not rain tomorrow). 60% is the quantification of the (**subjective**) belief of the meteorologist.

Frequentist vs. Subjectivist Interpretation

- **Frequentist Interpretation**
  - Probability is a property of the event (“it rains tomorrow in Bangalore”)
  - Is operationalized by repeated experiments
  - Typically used by scientists and engineers

- **Subjective Interpretation**
  - Probability is an expression of belief of the person that makes a statement
  - Is subjective and people-dependent: Two people with identical data can come to different probabilities
  - Typically used by philosophers and economists

1. Probability is not a physical measure but a thought model for randomness!
2. The mathematical rules for probability are **identical** for both interpretations!

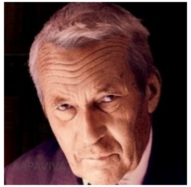
Rules of Probability

- **Mathematical Definition**. *A number* $P(A) \in [0,1]$ *assigned to an* event *or* statement A *that indicates how* likely A *is to occur.*
- **Set Theory**. We model events and statements via set theory and assume
  - A countably infinite total set $\Omega \supseteq A$
  - If $A(x)$ is a 1st order logic statement, then $\mathrm{A} := \{x \mid A(x)\}$ and
    - $\mathrm{A} \subseteq \mathrm{B} \equiv \forall x: A(x) \rightarrow B(x)$ and $\mathrm{A}^c \equiv \forall x: \neg A(x)$
    - $\mathrm{A} \cup \mathrm{B} \equiv \forall x: A(x) \vee B(x)$ and $\mathrm{A} \cap \mathrm{B} \equiv \forall x: A(x) \wedge B(x)$
- **Rules**: For all $A, B \subseteq \Omega$
  - **Monotonicity**: If $A \subseteq B$ then $P(A) \leq P(B)$
  - **Complement Rule**: $P(A^c) = 1 - P(A)$
  - **Sum Rule**: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
  - **Product Rule**: $P(A \cap B) = \underbrace{\frac{P(A \cap B)}{P(B)}}_{P(A|B)} \cdot P(B)$

Frequentist vs. Subjective Probabilities

- **Kolmogorov (1933)**: *The rules of probability for **sets** follow from the following 3 axioms*
  1. $P(A) \geq 0$ *for all* $A \subseteq \Omega$
  2. $P(\Omega) = 1$
  3. $P(\bigcup_i A_i) = \sum_i P(A_i)$ *if for all* $i \neq j: A_i \cap A_j = \emptyset$

- **Cox (1944)**: *The rules of probability for **logic** follow from the following 3 axioms*
  1. $P(A) \in [0,1]$ *for all logical statements* A
  2. $P(A)$ *is independent of how the statement is represented*
  3. *If* $P(A|C') > P(A|C)$ *and* $P(B|A \wedge C') = P(B|A \wedge C)$ *then*
  $$P(A \wedge B|C') \geq P(A \wedge B|C)$$



Andrey Kolmogorov (1903 – 1987)

Richard Threlkeld Cox (1898 – 1991)

The Role of Probability in Machine Learning

- **Theory**: *How likely is it, that the accuracy of a predictor $\mathcal{A}(D)$ learned from training data D is good?*

$$P\big(\text{Accuracy}\big(\mathcal{A}(D)\big) < \varepsilon\big) \le \delta$$

Objective probability over the repeated draw of training sets D

- **Typical Assumptions**
  1. Independent identically distributed data (IID)
  2. Accuracy is an expected performance measure on the next test example
- **Frequentist view on probability**: Over N applications of the learning algorithm and draws of random training data D, for how many is the learned predictor accurate?

- **Practice**: *What can we say about the plausibility of a single predictor f in light of training data D?*

$$P(f\,|D)$$

Subjective belief that f is the right predictor given D

- **Typical Assumptions**
  1. Independent identically distributed data (IID)
  2. Known conditional dependence of data and predictor
- **Subjectivist view on probability**: Given the certain and known training data D, what is the remaining uncertainty over the right predictor for (future) data?

Overview

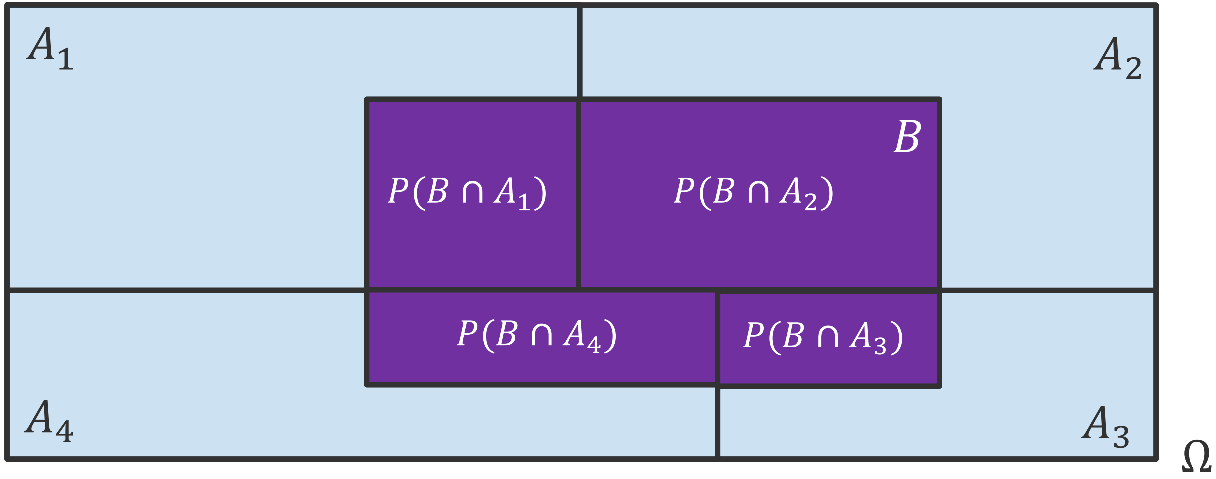
1. Probability in Machine Learning
2. **Probability Theory**
3. Exponential Family Distributions

Probability Theory: Law of Total Probability

- **Total Probability Theorem**. *Let $A_1, A_2, \ldots, A_n \subseteq \Omega$ be disjoint events that form a partition of the sample space Ω and $P(A_i) > 0$ for all A_i. Then, for any event $B \subseteq \Omega$*

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i) = \sum_{i=1}^{n} P(B|A_i) \cdot P(A_i)$$

- **Geometric Proof**



Probability Theory: Bayes Rule

(Rev) Thomas Bayes
(1701 – 1761)

- **Bayes' Theorem**. *Let $A_1, A_2, \dots, A_n$ be disjoint events that form a partition of the sample space S and $P(A_i) > 0$ for all A_i. Then, for any event B with $P(B) > 0$*

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{n} P(B|A_j) \cdot P(A_j)}$$

law of total probability

- **Proof**. Follows from the definition of conditional probability and "multiply-by-1"

$$P(A_i \cap B) \cdot \frac{P(B)}{P(B)} = P(A_i \cap B) \cdot \frac{P(A_i)}{P(A_i)}$$

=1 (by definition $P(A_i) > 0$ and $P(B) > 0$)

$$P(A_i|B) \cdot P(B) = P(B|A_i) \cdot P(A_i)$$

by definition of conditional probability

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

- **Simplified view** when looking at the probabilities as functions of A_i

$$P(A_i|B) \propto P(B|A_i) \cdot P(A_i)$$

posterior **likelihood** **prior**

Bayes Rule: False-Positive Puzzle

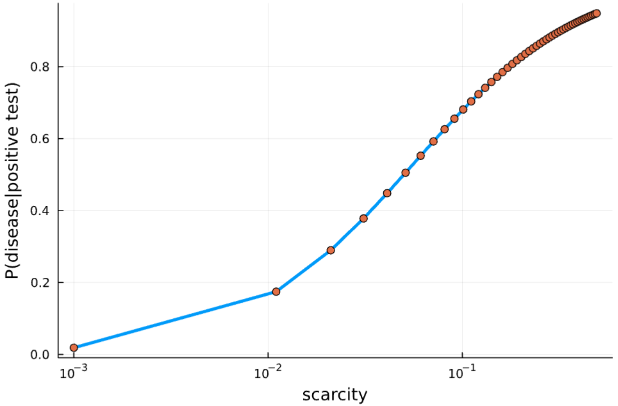
- **Situation**: A test for a rare disease is assumed to be correct 95% of the time (i.e., the probability that the test shows the disease or lack thereof is 95%). It's a rare disease that occurs in 0.1% of the population. If you have a positive test outcome, what is the probability that you have the disease?
- **Solution**:

$$A = \text{"Person has the disease"}$$

$$B = \text{"Test result is positive"}$$

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

$$P(A|B) = \frac{0.95 \cdot 0.001}{0.95 \cdot 0.001 + 0.05 \cdot 0.999} \approx 0.0187$$



- **Counterintuitive**: According to *The Economist* (February 20, 1999), 80% of leading American hospital staff guessed the probability to be 95%!

Probability Theory: Independence

- **Independence**. *We say that the events* $A_1, A_2, \ldots, A_n$ *are independent if*

$$P\left(\bigcap_{i\in I} A_i\right) = \prod_{i\in I} P(A_i), \qquad \text{for all subsets } I \text{ of } \{1, \ldots, n\}$$

- **Intuition**. Knowledge of an event A with $P(A) > 0$ does not provide information about the probability of an independent event B

$= P(B \cap A)$

$$P(B|A) = P(B) \quad \Leftrightarrow \boxed{P(B|A) \cdot P(A)} = P(B) \cdot P(A)$$

- **Important modelling assumption** (often implicitly) used in machine learning when making assumptions about training and test data generation: knowing one training example provides no information about the probability of any other training example (realistic?!)
- **Counterintuitive geometry**: If A and B are disjoint, they are **not** independent!

Probability Theory: Random Variable

- **Random Variable**. *A random variable is a real-valued function of the outcome of the experiment. A function of a random variable defines another random variable.*
  - **Examples**:
    - Tossing a coin N times, the **number** of heads
    - Given an image, the **pixel intensity** of the top-left pixel (in 8-bit)
- **Probability Mass Function**. *The probability mass function $p(x)$ assigns each value x the probability that the random variable takes the value x.*
  - **Example**: Coin toss: If $N = 2$ then

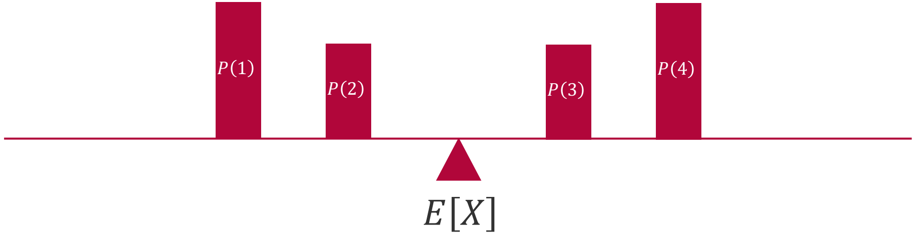
$$
\begin{aligned}
p(0) &= P(\text{tail}, \text{tail}) \\
p(1) &= P(\text{head}, \text{tail}) + P(\text{tail}, \text{head}) \\
p(2) &= P(\text{head}, \text{head})
\end{aligned}
$$

Probability Theory: Expectation and Variance

- **Expected Value**. *The expected value* $E[X]$ *(also called* expectation*) of a random variable* X *is defined by*

$$E[X] := \sum_x x \cdot p(x)$$

- **Intuition**. Center of gravity when placing the weight $p(x)$ at position x on a straight line



- **Variance**. *The variance* $\mathrm{var}[X]$ *of a random variable* X *is defined by*

$$\mathrm{var}[X] := \sum_x (x - E[X])^2 \cdot p(x) = E[(X - E[X])^2]$$

Overview

1. Probability in Machine Learning
2. Probability Theory
3. **Exponential Family Distributions**

Probability Distributions

- Only defined for **random variables**, *not* for events or logic statements!
  - **Discrete random variables**: $p: \mathbb{Z} \mapsto [0,1]$ and $\sum_x p(x) = 1$
  - **Continuous random variables**: $p: \mathbb{R} \mapsto \mathbb{R}^+$ and $\int p(x)\, \mathrm{d}x = 1$
    - Note that, by definition, they are only a **model** for real data!
- In computational statistics some classes of probability distributions have emerged whose distributions can be fully described with a small number of parameters $\boldsymbol{\theta} \in \mathbb{R}^d$
  - **Advantages**:
    1. **Storage Efficiency**: Only d real numbers for whole function!
    2. **Compute Efficiency**: Only $O(d)$ computation for rules of probability!
  - **Disadvantages**:
    1. **Real World**: Too restrictive to represent true phenomena in real data
    2. **Bayes' Rule**: Function classes often not closed under Bayes' rule

Efficient Bayes' Rule

- **Bayes' rule** over random variables X and Y

$$p(x|y) = \frac{p(y|x) \cdot p(x)}{p(y)} = \underbrace{\frac{1}{p(y)} \cdot \left[\int p(y|x)\, \mathrm{d}x\right]}_{\text{Normalization constant independent of } x} \cdot \underbrace{\left[\frac{p(y|x)}{\int p(y|x)\, \mathrm{d}x}\right]}_{\text{Probability distribution } p_y(\cdot) \text{ over } x} \cdot \underbrace{p(x)}_{\text{Probability distribution } p(\cdot) \text{ over } x}$$

- **Idea**: If $p(y|x)$ and $p(x)$ are exponentials of a linear function of transformation of x then (up to normalization), the product becomes a linear operation!

$$p(y|x) \propto \exp\left(\boldsymbol{\theta}_y^{\mathrm{T}} \cdot \boldsymbol{T}(x)\right)$$

$$p(x) \propto \exp\left(\boldsymbol{\theta}^{\mathrm{T}} \cdot \boldsymbol{T}(x)\right)$$

$$\Rightarrow \quad p(x|y) \propto \exp\left(\underbrace{\left[\boldsymbol{\theta}_y + \boldsymbol{\theta}\right]^{\mathrm{T}}}_{\text{Product in distribution becomes addition!}} \cdot \boldsymbol{T}(x)\right)$$

Exponential Family

- **Exponential Family Distribution**. *A distribution* $p(\cdot\,|\boldsymbol{\theta})$ *over a random variable* X *is called an* exponential family distribution *if it can be expressed as follows:*

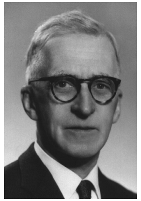
$$p(x|\boldsymbol{\theta}) = \exp\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{T}(x) - A(\boldsymbol{\theta})\right)$$

where $\boldsymbol{\theta}$ *is called the* natural parameters *of the distribution,* $\boldsymbol{T}(\cdot)$ *is called the* sufficient statistics *and* $A(\boldsymbol{\theta})$ *is the* log-normalization *given by*

$$A(\boldsymbol{\theta}) = \log\left(\int \exp\left(\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{T}(x)\right)\,\mathrm{d}x\right)$$

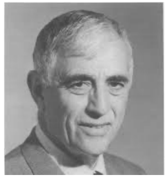
- Every discrete probability distribution over the outcome $x \in \{1, \dots, K\}$ is an exponential family distribution with

$$\boldsymbol{\theta} = \begin{bmatrix} \log(p(1)) \\ \vdots \\ \log(p(x-1)) \\ \log(p(x)) \\ \log(p(x+1)) \\ \vdots \\ \log(p(K)) \end{bmatrix} \qquad \boldsymbol{T}(x) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$



Edwin J. G. Pitman
(1897 – 1993)

Georges Darmois
(1888 - 1960)



Bernhard O. Koopman
(1900 – 1981)

Exponential Family Distributions: Categorical Distribution

Jacob Bernoulli
(1655 – 1705)

- **Categorical Distribution**. *Given* $\boldsymbol{\eta} \in \mathbb{R}^K$ *or a probability vector* $\boldsymbol{\pi} \in [0,1]^K$ *such that* $\sum_{k=1}^{K} \pi_k = 1$, *a discrete random variable* $\boldsymbol{X}$ *over the first* K *unit vectors is said to have a categorical distribution if the density is given by*

$$\mathcal{C}(\boldsymbol{x};\boldsymbol{\eta}) = \frac{\exp\left(\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{x}\right)}{\sum_{k=1}^{K}\exp(\eta_k)} \qquad \mathrm{Cat}(\boldsymbol{x};\boldsymbol{\pi}) = \boldsymbol{\pi}^{\mathrm{T}}\boldsymbol{x}$$

- **Properties**:

$$E[X_i] = \frac{\exp(\eta_i)}{\sum_{k=1}^{K}\exp(\eta_k)} = \pi_i \qquad \boldsymbol{T}(\boldsymbol{x}) = \begin{bmatrix} x_1 \\ \vdots \\ x_K \end{bmatrix} \qquad \boldsymbol{\theta} = \begin{bmatrix} \log(\pi_1) \\ \vdots \\ \log(\pi_K) \end{bmatrix}$$

$$\mathrm{var}[X_i] = \pi_i \cdot (1 - \pi_i)$$

- **Conversions**:

$$\mathcal{C}(\boldsymbol{x};\boldsymbol{\eta}) = \mathrm{Cat}\left(\boldsymbol{x}; \frac{1}{\sum_{k=1}^{K}\exp(\eta_k)} \cdot \begin{bmatrix} \exp(\eta_1) \\ \vdots \\ \exp(\eta_K) \end{bmatrix}\right) \qquad \mathrm{Cat}(\boldsymbol{x};\boldsymbol{\pi}) = \mathcal{C}\left(\boldsymbol{x}; \begin{bmatrix} \log(\pi_1) \\ \vdots \\ \log(\pi_K) \end{bmatrix}\right)$$

- **Importance**. The categorical distribution plays a key role in selection processes and classification and is also known as the *generalized Bernoulli distribution*.

Categorical Distribution: Efficient Products & Divisions

- **Theorem (Multiplication).** *Given two categorical distributions* $\mathcal{C}(\boldsymbol{x};\boldsymbol{\eta}_1)$ *and* $\mathcal{C}(\boldsymbol{x};\boldsymbol{\eta}_2)$

Correction factor

$$\mathcal{C}(\boldsymbol{x};\boldsymbol{\eta}_1)\cdot\mathcal{C}(\boldsymbol{x};\boldsymbol{\eta}_2)=\mathcal{C}(\boldsymbol{x};\boxed{\boldsymbol{\eta}_1+\boldsymbol{\eta}_2})\cdot\boxed{\frac{\sum_{k=1}^{K}\exp\left(\eta_{1,k}+\eta_{2,k}\right)}{\left[\sum_{k=1}^{K}\exp\left(\eta_{1,k}\right)\right]\cdot\left[\sum_{k=1}^{K}\exp\left(\eta_{2,k}\right)\right]}}$$

Additive updates!

- **Theorem (Division).** *Given two categorical distributions* $\mathcal{C}(\boldsymbol{x};\boldsymbol{\eta}_1)$ *and* $\mathcal{C}(\boldsymbol{x};\boldsymbol{\eta}_2)$

Correction factor

$$\frac{\mathcal{C}(\boldsymbol{x};\boldsymbol{\eta}_1)}{\mathcal{C}(\boldsymbol{x};\boldsymbol{\eta}_2)}=\mathcal{C}(\boldsymbol{x};\boxed{\boldsymbol{\eta}_1-\boldsymbol{\eta}_2})\cdot\boxed{\frac{\left[\sum_{k=1}^{K}\exp\left(\eta_{1,k}-\eta_{2,k}\right)\right]\cdot\left[\sum_{k=1}^{K}\exp\left(\eta_{2,k}\right)\right]}{\left[\sum_{k=1}^{K}\exp\left(\eta_{1,k}\right)\right]}}$$

Subtractive updates!

Exponential Family Distributions: Dirichlet Distribution

HPI Hasso Plattner Institut
Digital Engineering · Universität Potsdam

- **Dirichlet Distribution**. *Given $\boldsymbol{\alpha} \in \mathbb{R}^K$ with $\alpha_k > 0$, a continuous random variable $\boldsymbol{X} \in [0,1]^K$ over the $K-1$ dimensional simplex (that is, $\sum_{k=1}^K X_k = 1$) is said to have a Dirichlet distribution if the density is given by*

$$\mathrm{Dir}(\boldsymbol{x}; \boldsymbol{\alpha}) = \frac{1}{\mathcal{B}(\boldsymbol{\alpha})} \cdot \prod_{k=1}^{K} x_k^{\alpha_k - 1} \qquad \mathcal{B}(\boldsymbol{\alpha}) = \frac{\prod_{k=1}^K \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^K \alpha_k\right)}$$

- **Properties**:

$$E[X_i] = \frac{\alpha_i}{\sum_{k=1}^K \alpha_k}$$

$$E[\log(X_i)] = \psi(\alpha_i) - \psi\left(\sum_{k=1}^{K} \alpha_k\right) \qquad \psi(z) = \frac{\mathrm{d}}{\mathrm{d}z} \log\big(\Gamma(z)\big)$$

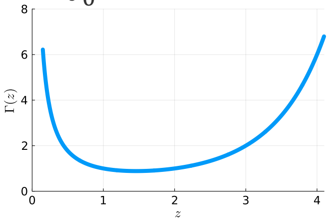
- **Natural Parameters and Sufficient Statistics**:

$$\boldsymbol{T}(\boldsymbol{x}) = \begin{bmatrix} \log(x_1) \\ \vdots \\ \log(x_K) \end{bmatrix} \qquad \boldsymbol{\theta} = \begin{bmatrix} \alpha_1 - 1 \\ \vdots \\ \alpha_K - 1 \end{bmatrix}$$

- **Importance**. The product of a Dirichlet $\mathrm{Dir}(\boldsymbol{\pi}; \boldsymbol{\alpha})$ and a categorical $\mathrm{Cat}(\boldsymbol{x}; \boldsymbol{\pi})$ is again a Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\pi}; \boldsymbol{\alpha} + \boldsymbol{x})$!

Peter Gustav Dirichlet
(1805 – 1859)

$$\Gamma(z) = \int_0^\infty t^{z-1} \cdot \exp(-t)\, \mathrm{d}t$$



$$\forall z > 0: \quad \Gamma(z+1) = z \cdot \Gamma(z)$$

Dirichlet Distribution: Efficient Products & Divisions

- **Theorem (Multiplication).** *Given two Dirichlet distributions* $\mathrm{Dir}(\boldsymbol{x};\boldsymbol{\alpha}_1)$ *and* $\mathrm{Dir}(\boldsymbol{x};\boldsymbol{\alpha}_2)$

$$\mathrm{Dir}(\boldsymbol{x};\boldsymbol{\alpha}_1)\cdot\mathrm{Dir}(\boldsymbol{x};\boldsymbol{\alpha}_2)=\mathrm{Dir}(\boldsymbol{x};\boldsymbol{\alpha}_1+\boldsymbol{\alpha}_2-\mathbf{1})\cdot\frac{\mathcal{B}(\boldsymbol{\alpha}_1+\boldsymbol{\alpha}_2-\mathbf{1})}{\mathcal{B}(\boldsymbol{\alpha}_1)\cdot\mathcal{B}(\boldsymbol{\alpha}_2)}$$

Correction factor

Additive updates!

- **Theorem (Division).** *Given two Dirichlet distributions* $\mathrm{Dir}(\boldsymbol{x};\boldsymbol{\alpha}_1)$ *and* $\mathrm{Dir}(\boldsymbol{x};\boldsymbol{\alpha}_2)$

$$\frac{\mathrm{Dir}(\boldsymbol{x};\boldsymbol{\alpha}_1)}{\mathrm{Dir}(\boldsymbol{x};\boldsymbol{\alpha}_2)}=\mathrm{Dir}(\boldsymbol{x};\boldsymbol{\alpha}_1-\boldsymbol{\alpha}_2+\mathbf{1})\cdot\frac{\mathcal{B}(\boldsymbol{\alpha}_1-\boldsymbol{\alpha}_2+\mathbf{1})\cdot\mathcal{B}(\boldsymbol{\alpha}_2)}{\mathcal{B}(\boldsymbol{\alpha}_1)}$$

Correction factor

Subtractive updates!

Exponential Family Distributions: 1D Gaussian

- **1D Gaussian Distribution**. *A continuous random variable X is said to have a one-dimensional Gaussian distribution if the density is given by*

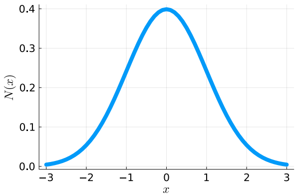
$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- **Properties**:

$$E[X] = \mu$$
$$\mathrm{var}[X] = \sigma^2$$

- **Importance**. The 1D Gaussian distribution plays a fundamental role in ML!
  - **Data Modelling**: The limit distribution for the sum of a large number of independent and identically distributed random variables.
  - **Machine Learning**: The most common belief distribution for the parameters of prediction functions!
  - **Information Theory**: The distribution function with the most uncertainty ("entropy") when fixing mean and variance of the random variable.

Carl Friedrich Gauss
(1777 - 1855)



1D Gaussian Distribution: Representations

- **Scale-Location Parameters**

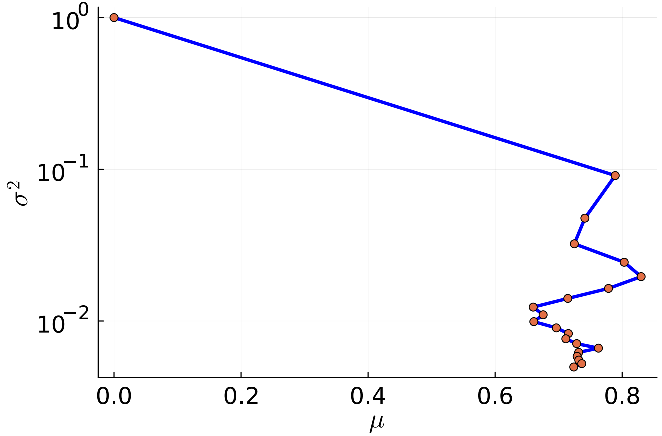
$$\mathcal{N}(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\cdot\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- **Conversions**

$$\mathcal{N}(x;\mu,\sigma^2) = \mathcal{G}\left(x;\frac{\mu}{\sigma^2},\frac{1}{\sigma^2}\right)$$

Two divisions only!

- **Posterior Inference**



- **Natural Parameters**

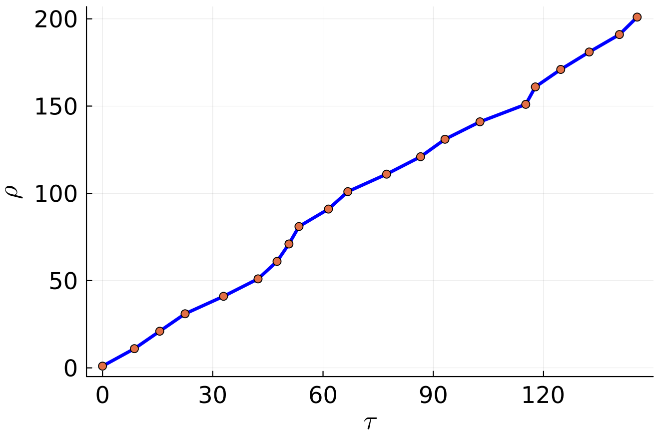
$$\mathcal{G}(x;\tau,\rho) = \sqrt{\frac{\rho}{2\pi}}\cdot\exp\left(-\frac{\tau^2}{2\rho}\right)\cdot\exp\left(\tau\cdot x-\rho\cdot\frac{x^2}{2}\right)$$

- **Conversions**

$$\mathcal{G}(x;\tau,\rho) = \mathcal{N}\left(x;\frac{\tau}{\rho},\frac{1}{\rho}\right)$$

$$\boldsymbol{T}(x) = \begin{bmatrix} x \\ -\frac{x^2}{2}\end{bmatrix} \qquad \boldsymbol{\theta} = \begin{bmatrix}\tau \\ \rho\end{bmatrix}$$

- **Posterior Inference**



1D Gaussian Distribution: Efficient Products & Divisions

- **Theorem (Multiplication)**. *Given two one-dimensional Gaussian distributions $\mathcal{G}(x;\tau_1,\rho_1)$ and $\mathcal{G}(x;\tau_2,\rho_2)$ we have*

Gaussian density

$$\mathcal{G}(x;\tau_1,\rho_1)\cdot\mathcal{G}(x;\tau_2,\rho_2)=\mathcal{G}(x;\tau_1+\tau_2,\rho_1+\rho_2)\cdot\mathcal{N}(\mu_1;\mu_2,\sigma_1^2+\sigma_2^2)$$

Additive updates!

- **Theorem (Division)**. *Given two one-dimensional Gaussian distributions $\mathcal{G}(x;\tau_1,\rho_1)$ and $\mathcal{G}(x;\tau_2,\rho_2)$ where $\rho_1>\rho_2$ we have*

Correction factor

$$\frac{\mathcal{G}(x;\tau_1,\rho_1)}{\mathcal{G}(x;\tau_2,\rho_2)}=\frac{\mathcal{G}(x;\tau_1-\tau_2,\rho_1-\rho_2)}{\mathcal{N}(\mu_1;\mu_2,\sigma_2^2-\sigma_1^2)}\cdot\frac{\sigma_2^2}{\sigma_2^2-\sigma_1^2}$$

Subtractive updates!

Gaussian density

Limit 1D Gaussian Distributions: Dirac Delta and Uniform

- **Dirac Delta**. The Dirac delta function $\delta(\cdot)$ is defined as the limit $\sigma^2 \to 0$

$$\delta(x) = \lim_{\sigma^2 \to 0} \mathcal{N}(x; 0, \sigma^2)$$

- **Gaussian Uniform**. The Gaussian uniform $\mathcal{U}(\cdot)$ is defined as the limit $\sigma^2 \to \infty$

$$\mathcal{U}(x) = \lim_{\sigma^2 \to +\infty} \mathcal{N}(x; 0, \sigma^2)$$

- **Theorem (Convolution of Normal with Dirac)**. *For any $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}^+$*

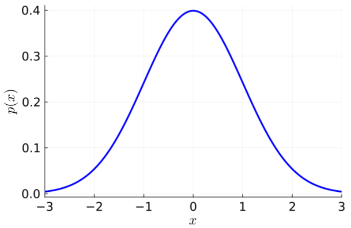
$$\int_{-\infty}^{+\infty} \delta(x) \cdot \mathcal{N}(x; \mu, \sigma^2)\,\mathrm{d}x = \boxed{\mathcal{N}(0; \mu, \sigma^2)}$$

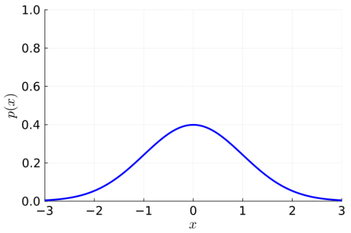
Gaussian density at $x = 0$

- **Theorem (Product of Normal with Uniform)**. *For any $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}^+$*

$$\frac{\mathcal{U}(x) \cdot \mathcal{N}(x; \mu, \sigma^2)}{\int_{-\infty}^{+\infty} \mathcal{U}(\tilde{x}) \cdot \mathcal{N}(\tilde{x}; \mu, \sigma^2)\,\mathrm{d}\tilde{x}} = \boxed{\mathcal{N}(x; \mu, \sigma^2)}$$

Equivalent to multiplying with 1





Exponential Family Distributions: Gamma Distribution

- **Gamma Distribution**. *Given $\beta > -1$ and $\lambda > 0$, a positive continuous random variable X is said to have a Dirichlet distribution if the density is given by*

$$\mathrm{Gam}(x;\beta,\lambda) = \frac{\lambda^{\beta+1}}{\Gamma(\beta+1)} \cdot x^{\beta} \cdot \exp(-\lambda x) \cdot \mathbb{I}(x > 0)$$

- **Properties**:

$$E[X] = \frac{\beta+1}{\lambda}$$

$$E[\log(X)] = \psi(\beta+1) - \log(\lambda) \qquad \psi(z) = \frac{\mathrm{d}}{\mathrm{d}z}\log\big(\Gamma(z)\big)$$

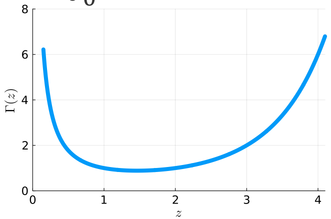
- **Natural Parameters and Sufficient Statistics**:

$$\boldsymbol{T}(x) = \begin{bmatrix}\log(x) \\ -x\end{bmatrix} \qquad \boldsymbol{\theta} = \begin{bmatrix}\beta \\ \lambda\end{bmatrix}$$

- **Importance**. The product of a Gamma distribution $\mathrm{Gam}(\rho;\beta,\lambda)$ and a Normal distribution $\mathcal{N}(x;\mu,\rho^{-1})$ is again a Gamma distribution $\mathrm{Gam}\left(\rho;\beta+\frac{1}{2},\lambda+\frac{(x-\mu)^2}{2}\right)$!

Karl Pearson
(1857 – 1936)

$$\Gamma(z) = \int_0^\infty t^{z-1} \cdot \exp(-t)\,\mathrm{d}t$$



$$\forall z > 0: \quad \Gamma(z+1) = z \cdot \Gamma(z)$$

Gamma Distribution: Efficient Products & Divisions

- **Theorem (Multiplication).** *Given two Gamma distributions* $\mathrm{Gam}(x;\beta_1,\lambda_1)$ *and* $\mathrm{Gam}(x;\beta_2,\lambda_2)$

Correction factor

$$\mathrm{Gam}(x;\beta_1,\lambda_1)\cdot\mathrm{Gam}(x;\beta_2,\lambda_2)=\mathrm{Gam}(x;\boxed{\beta_1+\beta_2,\lambda_1+\lambda_2})\cdot\boxed{\frac{\mathrm{Gam}(1;\beta_1,\lambda_1)\cdot\mathrm{Gam}(1;\beta_2,\lambda_2)}{\mathrm{Gam}(1;\beta_1+\beta_2,\lambda_1+\lambda_2)}}$$

Additive updates!

- **Theorem (Division).** *Given two Gamma distributions* $\mathrm{Gam}(x;\beta_1,\lambda_1)$ *and* $\mathrm{Gam}(x;\beta_2,\lambda_2)$

Correction factor

$$\frac{\mathrm{Gam}(x;\beta_1,\lambda_1)}{\mathrm{Gam}(x;\beta_2,\lambda_2)}=\mathrm{Gam}(x;\boxed{\beta_1-\beta_2,\lambda_1-\lambda_2})\cdot\boxed{\frac{\mathrm{Gam}(1;\beta_1,\lambda_1)}{\mathrm{Gam}(1;\beta_1-\beta_2,\lambda_1-\lambda_2)\cdot\mathrm{Gam}(1;\beta_2,\lambda_2)}}$$

Subtractive updates!

Summary

- **Probability in Machine Learning**
  - Probability is not a physical quantity but a mathematical model of uncertainty
  - Two different axiomatic justifications of the same math: one for data and one for parameters!
- **Probability Theory**
  - Two key rules of probability theory: (1) Total probability rule and (2) Bayes' rule
  - Independence is a concept of probability; it does not require random variables!
  - A random variable is a real-valued function of the outcome of the experiment
- **Exponential Family Distributions**
  - Density is proportional to an exponential of a *fixed* d-dimensional linear mapping of the random variable (*sufficient statistic*) and a (*natural*) *parameter* vector
  - Taking products and divisions of densities is an $\mathcal{O}(d)$ operation!
  - Key distributions are categorical, Dirichlet, Gaussian and Gamma distributions

Hasso Plattner Institut
Digital Engineering • Universität Potsdam

See you next week!