Probability and Statistics

Michele Caprio

Department of Computer Science, University of Manchester
Manchester Centre for AI Fundamentals

COMP 64101 – Reasoning and Learning under Uncertainty
Lecture 1

Topics: Probability

"I know it's Tuesday. It's a good day for math!"

Max Mintz

Topics: Probability

"I know it's Tuesday. It's a good day for math!"

Max Mintz

- Discrete vs Continuous Distributions (Murphy, 2023, § 2.1.2 - 2.1.3)
- Bayes' Rule (Murphy, 2023, § 2.1.5 - 2.1.6)
- Some Common Probability Distributions (Murphy, 2023, § 2.2 - 2.3)
  - Mixture of Gaussians (Murphy, 2023, § 28.2.1)
- Markov Chains (Murphy, 2023, § 2.6)
- On Your Own: The Exponential Family (Murphy, 2023, § 2.4) and Divergences between probabilities (Murphy, 2023, § 2.7)

Topics: Statistics

"I know it's Monday. It's a good day for math!"

Max Mintz

- (Some Concepts of) Bayesian Statistics (Murphy, 2023, § 3.2)
- (Some Concepts of) Frequentist Statistics (Murphy, 2023, § 3.3)
- Maximum Likelihood Estimator and the EM Algorithm (Murphy, 2023, § 6.5.3)
- On Your Own: Hypotheses Testing (Murphy, 2023, § 3.10)

Probability Space

- To talk about probability, we need to introduce the probability space

$$(\Omega, \mathcal{F}, \mathbb{P})$$

Probability Space

- To talk about probability, we need to introduce the probability space

$$(\Omega, \mathcal{F}, \mathbb{P})$$

- Ω is the sample space (possible outcomes from an experiment)
- $\mathcal{F}$ is the event space (σ-algebra), a collection of subsets of Ω
- $\mathbb{P} : \mathcal{F} \to [0, 1]$ is the probability measure

(Kolmogorovian) Probability Axioms

- $\mathbb{P}(E) \geq 0$, for all $E \in \mathcal{F}$
- $\mathbb{P}(\Omega) = 1$
- $\mathbb{P}(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} \mathbb{P}(E_i)$

(Kolmogorovian) Probability Axioms

- $\mathbb{P}(E) \geq 0$, for all $E \in \mathcal{F}$
- $\mathbb{P}(\Omega) = 1$
- $\mathbb{P}(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} \mathbb{P}(E_i)$
  - Only Finite Additivity: Subjectivist Approach to Probability de Finetti (1974, 1975)
  - Super/Subadditivity: Imprecise Approach to Probability Walley (1991); Augustin et al. (2014)

Discrete Random Variables

- Outcomes of the experiment constitute a countable set
- **Example:** We flip a coin twice

Discrete Random Variables

- Outcomes of the experiment constitute a countable set
- **Example:** We flip a coin twice
- $\Omega = \{\omega_1 = (H, H), \omega_2 = (H, T), \omega_3 = (T, H), \omega_4 = (T, T)\}$
- $\mathcal{F} = 2^{\Omega}$, so $|\mathcal{F}| = 2^4 = 16$
- $\mathbb{P}(\{\omega_i\}) = 1/4$, $i \in \{1, \ldots, 4\}$, and the probability of the other sets in $\mathcal{F}$ follows by additivity

Discrete Random Variables

- Can we assign a number to each element of Ω (i.e. to each outcome of our experiment of interest)?
- Yes, via a Random Variable (rv)

$$X : \Omega \to \mathbb{R}$$

  - In this case is discrete because Ω is countable

Discrete Random Variables

- Can we assign a number to each element of Ω (i.e. to each outcome of our experiment of interest)?
- Yes, via a Random Variable (rv)

$$X : \Omega \to \mathbb{R}$$

  - In this case is discrete because Ω is countable
- **Example cont'd:** Number of Heads via rv

$$X(\omega_1) = 2, \quad X(\omega_2) = X(\omega_3) = 1, \quad X(\omega_4) = 0$$

Discrete Random Variables

- Random Variable need not assign only numbers to the outcomes of the experiment
- In general, we call state space $\mathcal{X}$ the range[1] of rv X, i.e. $\mathcal{X} = X(\Omega)$

[1] The image of the domain Ω of X under X.

Discrete Random Variables

- Random Variable need not assign only numbers to the outcomes of the experiment
- In general, we call state space $\mathcal{X}$ the range[1] of rv X, i.e. $\mathcal{X} = X(\Omega)$
- We immediately obtain the probability of any given state in $a \in \mathcal{X}$ as

$$p_X(a) \equiv \mathbb{P}(X = a) = \mathbb{P}[X^{-1}(a)], \quad X^{-1}(a) := \{\omega \in \Omega : X(\omega) = a\}$$

  - p_X is called the probability mass function (pmf) for rv X
  - Can be represented by a histogram or some parametric function

[1] The image of the domain Ω of X under X.

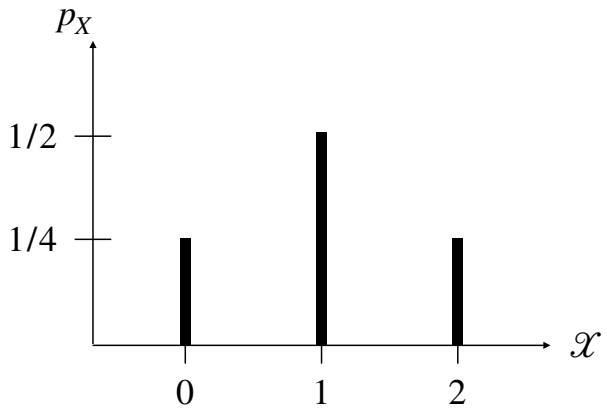
Discrete Random Variables

- Random Variable need not assign only numbers to the outcomes of the experiment
- In general, we call state space $\mathcal{X}$ the range[1] of rv X, i.e. $\mathcal{X} = X(\Omega)$
- We immediately obtain the probability of any given state in $a \in \mathcal{X}$ as

$$p_X(a) \equiv \mathbb{P}(X = a) = \mathbb{P}[X^{-1}(a)], \quad X^{-1}(a) := \{\omega \in \Omega : X(\omega) = a\}$$

  - p_X is called the probability mass function (pmf) for rv X
  - Can be represented by a histogram or some parametric function

- **Example cont'd:** pmf is
  - $p_X(0) = \mathbb{P}(\{(H, H)\}) = 1/4$
  - $p_X(1) = \mathbb{P}(\{(H, T), (T, H)\}) = 1/2$
  - $p_X(2) = \mathbb{P}(\{(T, T)\}) = 1/4$



[1] The image of the domain Ω of X under X.

Continuous Random Variables

- Experiments with continuous outcome
- $\Omega \subseteq \mathbb{R}$, and $X(\omega) = \omega$, so that $\mathcal{X} = \Omega$
- **Example:** The duration of some event (in seconds), so $\Omega = \{t \in \mathbb{R}_+ : t \leq T_{\text{max}}\}$
- Letting event space $\mathcal{F}$ be a power set when Ω is uncountable may be too large
  - Use Borel σ-algebra
  - On $\mathbb{R}$, it is the σ-algebra generated by semi-closed intervals $(-\infty, b]$

σ-algebra

A collection $\mathcal{F}$ of subsets of Ω such that

- $\emptyset, \Omega \in \mathcal{F}$
- $E \in \mathcal{F} \implies E^c \in \mathcal{F}$
- $\{E_n\}_{n=1}^{\infty} \subset \mathcal{F} \implies \cup_n E_n, \cap_n E_n \in \mathcal{F}$

Probability Density and Cumulative Distribution Functions (pdf, cdf)

- Let $\Omega = \mathbb{R}$, and call ℓ the Lebesgue measure
  - ℓ measures the length of intervals on the Real line

Probability Density and Cumulative Distribution Functions (pdf, cdf)

- Let $\Omega = \mathbb{R}$, and call ℓ the Lebesgue measure
  - ℓ measures the length of intervals on the Real line
- We assume $\mathbb{P} \ll \ell$, i.e. $\ell(E) = 0 \implies \mathbb{P}(E) = 0$, $E \in \mathcal{F}$
- The Radon-Nykodim derivative $\mathrm{d}\mathbb{P}/\mathrm{d}\ell = p$ is the pdf of our continuous rv X
  - RND is a function on $\mathcal{X} = \Omega = \mathbb{R}$ such that, for any $E \in \mathcal{F}$,

$$\mathbb{P}(E) = \int_E p(x)\ell(\mathrm{d}x) = \int_E p(x)\mathrm{d}x$$

  - E is an interval, e.g. $E = [a, b]$, $a < b$

Probability Density and Cumulative Distribution Functions (pdf, cdf)

- Let $\Omega = \mathbb{R}$, and call ℓ the Lebesgue measure
  - ℓ measures the length of intervals on the Real line
- We assume $\mathbb{P} \ll \ell$, i.e. $\ell(E) = 0 \implies \mathbb{P}(E) = 0$, $E \in \mathcal{F}$
- The Radon-Nykodim derivative $\mathrm{d}\mathbb{P}/\mathrm{d}\ell = p$ is the pdf of our continuous rv X
  - RND is a function on $\mathcal{X} = \Omega = \mathbb{R}$ such that, for any $E \in \mathcal{F}$,

$$\mathbb{P}(E) = \int_E p(x)\ell(\mathrm{d}x) = \int_E p(x)\mathrm{d}x$$

  - E is an interval, e.g. $E = [a, b]$, $a < b$
- cdf is defined as

$$F(x) \equiv \mathbb{P}(X \leq x) = \int_{-\infty}^{x} p\left(x'\right) \mathrm{d}x'$$

Probability Density and Cumulative Distribution Functions (pdf, cdf)

- Let $\Omega = \mathbb{R}$, and call ℓ the Lebesgue measure
  - ℓ measures the length of intervals on the Real line
- We assume $\mathbb{P} \ll \ell$, i.e. $\ell(E) = 0 \implies \mathbb{P}(E) = 0$, $E \in \mathcal{F}$
- The Radon-Nykodim derivative $\mathrm{d}\mathbb{P}/\mathrm{d}\ell = p$ is the pdf of our continuous rv X
  - RND is a function on $\mathcal{X} = \Omega = \mathbb{R}$ such that, for any $E \in \mathcal{F}$,

$$\mathbb{P}(E) = \int_E p(x)\ell(\mathrm{d}x) = \int_E p(x)\mathrm{d}x$$

  - E is an interval, e.g. $E = [a, b]$, $a < b$
- cdf is defined as

$$F(x) \equiv \mathbb{P}(X \leq x) = \int_{-\infty}^{x} p\left(x'\right) \mathrm{d}x'$$

- These notions can be generalized to $\mathbb{R}^n$, and to more complex sample spaces

Conditional Probability

- Consider events E_1 and E_2, and suppose $\mathbb{P}(E_2) > 0$. Then, conditional probability of E_1 given E_2 is

$$\mathbb{P}(E_1 \mid E_2) := \frac{\mathbb{P}(E_1 \cap E_2)}{\mathbb{P}(E_2)}$$

- In turn, $\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1 \mid E_2)\mathbb{P}(E_2) = \mathbb{P}(E_2 \mid E_1)\mathbb{P}(E_1)$

Conditional Probability

- Consider events E_1 and E_2, and suppose $\mathbb{P}(E_2) > 0$. Then, conditional probability of E_1 given E_2 is

$$\mathbb{P}(E_1 \mid E_2) := \frac{\mathbb{P}(E_1 \cap E_2)}{\mathbb{P}(E_2)}$$

- In turn, $\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1 \mid E_2)\mathbb{P}(E_2) = \mathbb{P}(E_2 \mid E_1)\mathbb{P}(E_1)$
- Conditional probability measures how likely an event E_1 is, given that event E_2 has happened

Conditional Probability

- Consider events E_1 and E_2, and suppose $\mathbb{P}(E_2) > 0$. Then, conditional probability of E_1 given E_2 is

$$\mathbb{P}(E_1 \mid E_2) := \frac{\mathbb{P}(E_1 \cap E_2)}{\mathbb{P}(E_2)}$$

- In turn, $\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1 \mid E_2)\mathbb{P}(E_2) = \mathbb{P}(E_2 \mid E_1)\mathbb{P}(E_1)$
- Conditional probability measures how likely an event E_1 is, given that event E_2 has happened
- Law of total probability: if $\{A_i\}_{i=1}^n$ is a partition of Ω, – i.e. $\sqcup_{i=1}^n A_i = \Omega$ – then for all $B \subseteq \Omega$,

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B \mid A_i)\mathbb{P}(A_i)$$

A Note on Independent Events

- E_1 and E_2 are independent events if

$$\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1)\mathbb{P}(E_2)$$

- If both $\mathbb{P}(E_1) > 0$ and $\mathbb{P}(E_2) > 0$, this is equivalent to

$$\mathbb{P}(E_1 \mid E_2) = \mathbb{P}(E_1), \quad \mathbb{P}(E_2 \mid E_1) = \mathbb{P}(E_2)$$

A Note on Independent Events

- E_1 and E_2 are independent events if

$$\mathbb{P}(E_1 \cap E_2) = \mathbb{P}(E_1)\mathbb{P}(E_2)$$

- If both $\mathbb{P}(E_1) > 0$ and $\mathbb{P}(E_2) > 0$, this is equivalent to

$$\mathbb{P}(E_1 \mid E_2) = \mathbb{P}(E_1), \quad \mathbb{P}(E_2 \mid E_1) = \mathbb{P}(E_2)$$

- E_1 and E_2 are conditionally independent given E_3 if

$$\mathbb{P}(E_1 \cap E_2 \mid E_3) = \mathbb{P}(E_1 \mid E3)\mathbb{P}(E_2 \mid E3)$$

Bayes' Theorem

- Consider events E_1 and E_2, and suppose $\mathbb{P}(E_1), \mathbb{P}(E_2) > 0$. Then, Bayes' rule is

$$\mathbb{P}(E_1 \mid E_2) = \frac{\mathbb{P}(E_2 \mid E_1)\mathbb{P}(E_1)}{\mathbb{P}(E_2)}$$

- Discrete case with $|\mathcal{X}| = K$,

$$p(X = k \mid E) = \frac{p(E \mid X = k)p(X = k)}{\sum_{k'=1}^{K} p(E \mid X = k')p(X = k')}$$

- Continuous case, e.g. with $\mathcal{X} = \mathbb{R}$,

$$p(x \mid E) = \frac{p(E \mid x)p(x)}{\int_{\mathcal{X}} p(E \mid x')p(x')\mathrm{d}x'}$$

Common Discrete Distributions

- Let $\mathcal{X} = \{1, \ldots, K\}$
- Binomial: $X \sim \text{Bin}(N, \mu)$, $p(x) = \binom{N}{x} \mu^x (1-\mu)^{N-x}$
  - $\binom{N}{x} := \frac{N!}{(N-x)!x!}$ and $\mu \in [0, 1]$
  - Number x of successes in a sequence of N independent experiments, each asking a yes–no question, and having success probability μ

Common Discrete Distributions

- Let $\mathcal{X} = \{1, \ldots, K\}$
- Binomial: $X \sim \text{Bin}(N, \mu)$, $p(x) = \binom{N}{x} \mu^x (1-\mu)^{N-x}$
  - $\binom{N}{x} := \frac{N!}{(N-x)!x!}$ and $\mu \in [0, 1]$
  - Number x of successes in a sequence of N independent experiments, each asking a yes–no question, and having success probability μ
- Categorical: $X \sim \text{Cat}(\theta)$, $p_X(k) = \theta_k$
  - θ is a probability vector, and hence belongs to unit simplex $\Delta^{K-1} \subset \mathbb{R}^K$
  - $p_X(k) = \theta_k \rightarrow$ probability that X is equal to k; such probability is the k^{th} entry of parameter θ
  - Fundamental for Classification Problems

Common Discrete Distributions

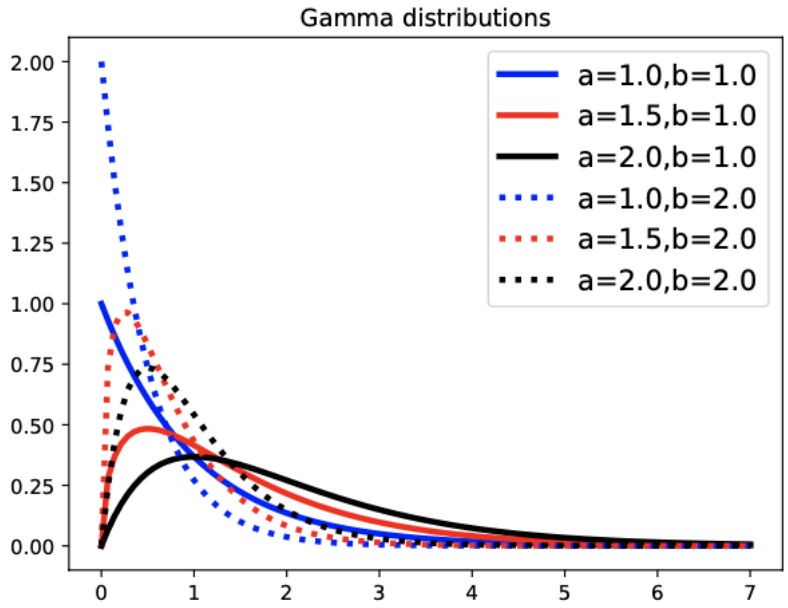
- Let $\mathcal{X} = \{1, \ldots, K\}$
- Binomial: $X \sim \text{Bin}(N, \mu)$, $p(x) = \binom{N}{x} \mu^x (1-\mu)^{N-x}$
  - $\binom{N}{x} := \frac{N!}{(N-x)!x!}$ and $\mu \in [0, 1]$
  - Number x of successes in a sequence of N independent experiments, each asking a yes–no question, and having success probability μ
- Categorical: $X \sim \text{Cat}(\theta)$, $p_X(k) = \theta_k$
  - θ is a probability vector, and hence belongs to unit simplex $\Delta^{K-1} \subset \mathbb{R}^K$
  - $p_X(k) = \theta_k \rightarrow$ probability that X is equal to k; such probability is the k^{th} entry of parameter θ
  - Fundamental for Classification Problems
- Poisson: $X \sim \text{Po}(\lambda)$, $p(x) = e^{-\lambda} \frac{\lambda^x}{x!}$
  - $\lambda > 0$; here $\mathcal{X} = \mathbb{N}_0$
  - Probability of a given number x of events occurring in a fixed interval of time if these events occur with a known constant mean rate λ and independently of the time since the last event

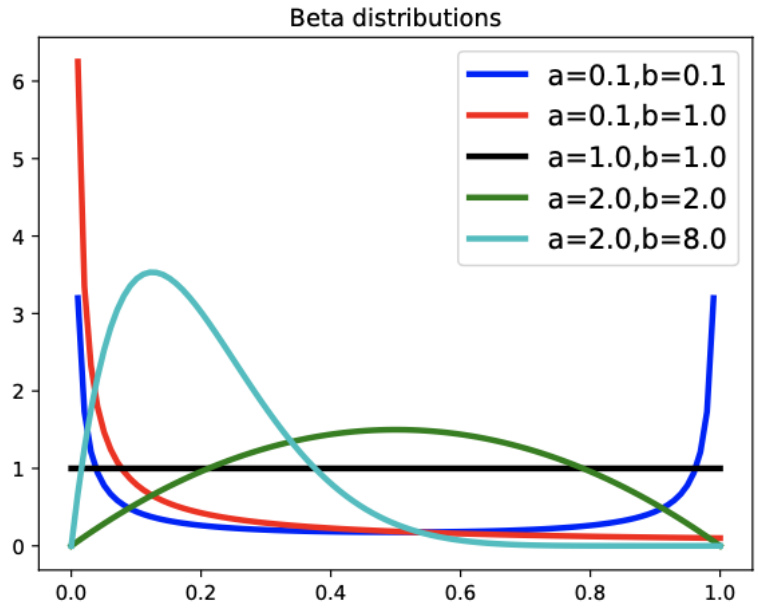
Common Continuous Distributions on subsets of $\mathbb{R}$

- Gamma: $X \sim \text{Ga}(a, b)$, $p(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-xb}$
  - $a, b > 0$ and $\mathcal{X} = \mathbb{R}_+$
  - $\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} \mathrm{d}t$, which is equal to $(a-1)!$ if $a \in \mathbb{N}_0$

Common Continuous Distributions on subsets of $\mathbb{R}$

- Gamma: $X \sim \text{Ga}(a, b)$, $p(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-xb}$
  - $a, b > 0$ and $\mathcal{X} = \mathbb{R}_+$
  - $\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} \mathrm{d}t$, which is equal to $(a-1)!$ if $a \in \mathbb{N}_0$
- Beta: $X \sim \text{Beta}(a, b)$, $p(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}$
  - $a, b > 0$ and $\mathcal{X} = [0, 1]$
  - $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$



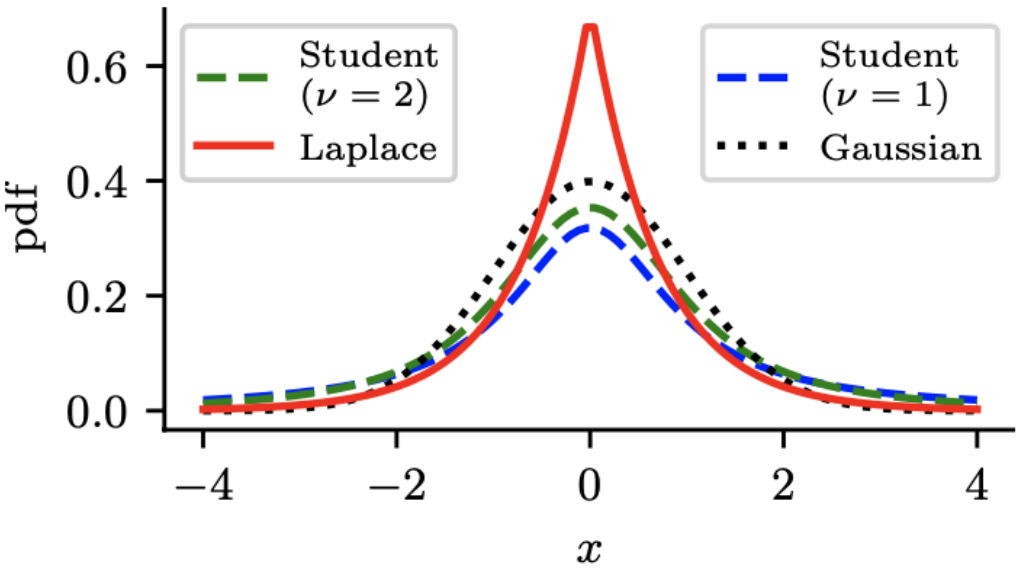


Common Continuous Distributions on $\mathbb{R}$

- Gaussian: $X \sim \mathcal{N}(\mu, \sigma^2)$, $p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
  - $\mu \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}_+$
  - Ubiquitous in science; parly because of the Central Limit Theorem
  - Standard Normal: $\mu = 0$, $\sigma = 1$

Common Continuous Distributions on $\mathbb{R}$

- Gaussian: $X \sim \mathcal{N}(\mu, \sigma^2)$, $p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
  - $\mu \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}_+$
  - Ubiquitous in science; parly because of the Central Limit Theorem
  - Standard Normal: $\mu = 0$, $\sigma = 1$
- Problem with the Gaussian distribution: sensitive to outliers
  - Probability decays exponentially fast with the (squared) distance from the center
  - Solution: more robust distributions, Laplace, Student t, Cauchy

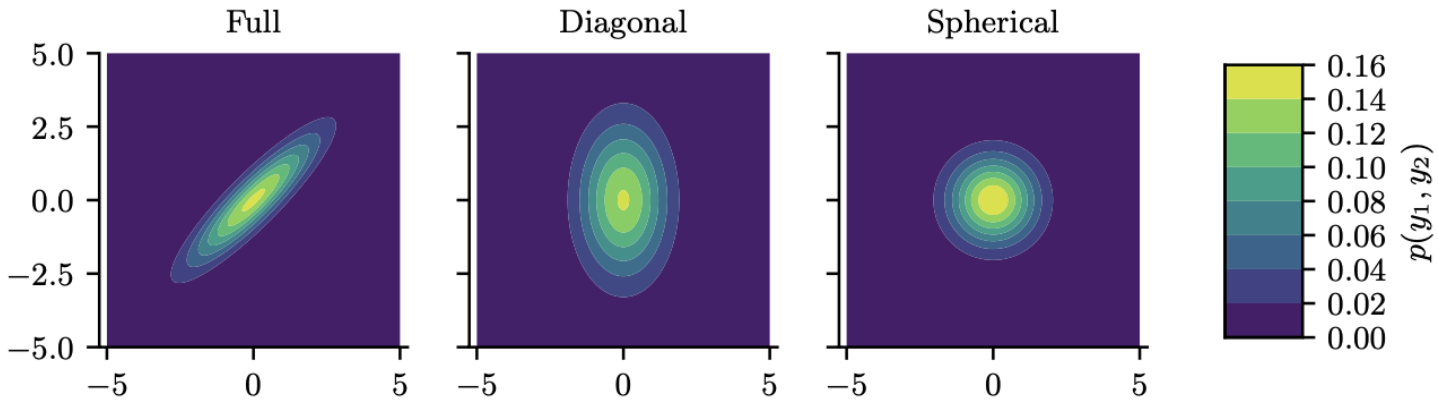


The Holy Grail

- Multivariate Normal: $X \sim \mathcal{N}(\mu, \Sigma)$,
  $p(x) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right]$
  - $\mu \in \mathbb{R}^D$, $\Sigma \in \mathbb{R}^{D \times D}$

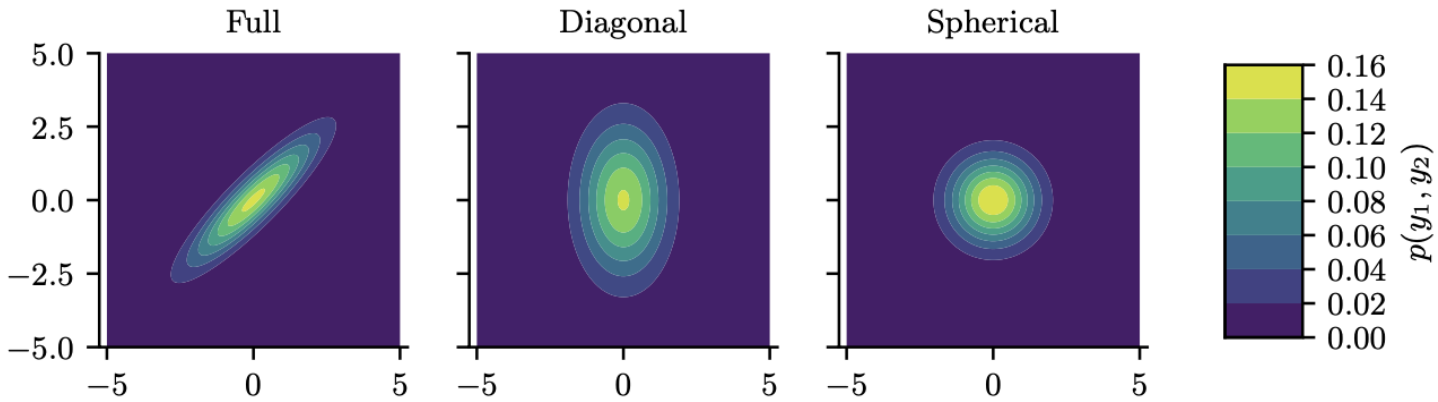
The Holy Grail

- Multivariate Normal: $X \sim \mathcal{N}(\mu, \Sigma)$,
  $p(x) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right]$
  - $\mu \in \mathbb{R}^D$, $\Sigma \in \mathbb{R}^{D \times D}$
  - Full Covariance Matrix: $D(D+1)/2$ parameters; we divide by 2 since Σ is symmetric
  - Diagonal covariance matrix: D parameters, and 0s in the off-diagonal terms
  - Spherical covariance matrix: $\Sigma = \sigma^2 I_D$, so only one free parameter σ



The Holy Grail

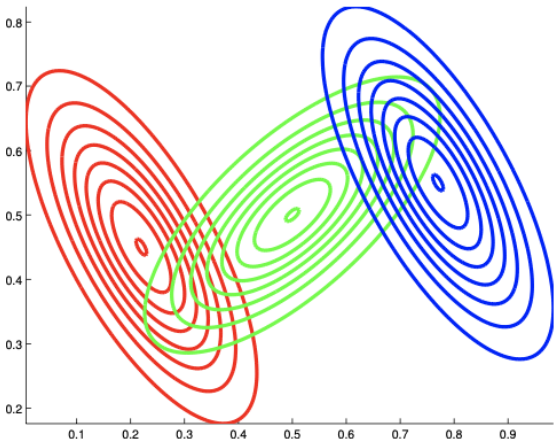
- Multivariate Normal: $X \sim \mathcal{N}(\mu, \Sigma)$,
  $p(x) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right]$
  - $\mu \in \mathbb{R}^D$, $\Sigma \in \mathbb{R}^{D\times D}$
  - Full Covariance Matrix: $D(D+1)/2$ parameters; we divide by 2 since Σ is symmetric
  - Diagonal covariance matrix: D parameters, and 0s in the off-diagonal terms
  - Spherical covariance matrix: $\Sigma = \sigma^2 I_D$, so only one free parameter σ



- Strong suggestion: Read chapter 2.3

Mixture of Gaussians

- Gaussian mixture model (GMM): $p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mu_k, \Sigma_k)$
  - If we let the number of mixture components grow sufficiently large, a GMM can approximate any smooth distribution over $\mathbb{R}^D$
  - GMMs are often used for unsupervised clustering of real-valued data samples $x_n \in \mathbb{R}^D$



Matrix Distributions (Not Necessarily Square Matrices)

- Matrix Normal: $X \sim \mathcal{MN}(M, U, V)$; property:

$$\text{vec}(X) \sim \mathcal{N}(\text{vec}(M), V^{-1} \otimes U)$$

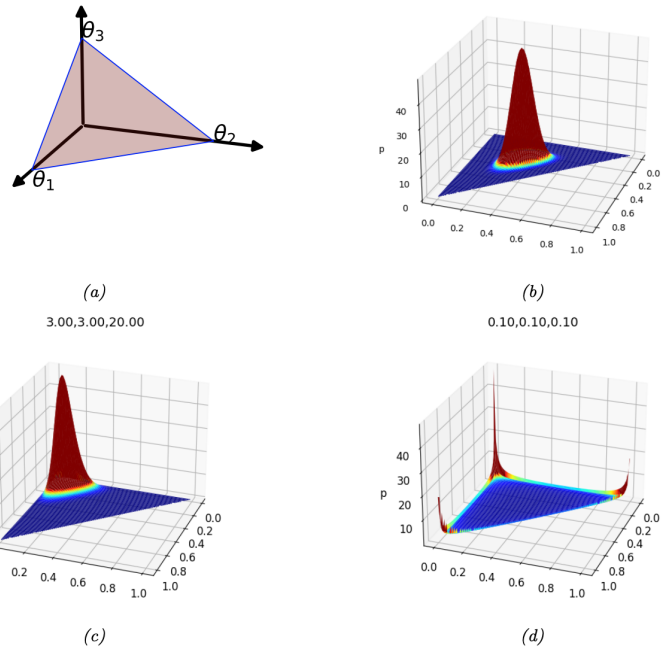
- mean value: $M \in \mathbb{R}^{n \times p}$
- covariance among rows: $U \in \mathbb{R}^{n \times n}$ positive definite
- precision among columns: $V \in \mathbb{R}^{p \times p}$ positive definite
- $\otimes$ is the Kroneker product, so $V^{-1} \otimes U = \begin{bmatrix} v_{11}^{\text{inv}} U & \cdots & v_{1p}^{\text{inv}} U \\ \vdots & \ddots & \vdots \\ v_{p1}^{\text{inv}} U & \cdots & v_{pp}^{\text{inv}} U \end{bmatrix}$
- Useful when dealing with black-and-white pictures

Matrix Distributions (Square Matrices)

- Wishart: $X \sim \text{Wi}(S, \nu)$
- Scale matrix: $S \in \mathbb{R}^{D \times D}$ positive definite
- Degrees of freedom: $\nu > D - 1$
- Property: $x_n \sim \mathcal{N}(0, \Sigma)$ i.i.d. $\implies S = \sum_{n=1}^{N} x_n x_n^\top \sim \text{Wi}(\Sigma, N)$
- Useful to model our uncertainty when estimating covariance matrices

Dirichlet Distribution (Unit Simplex)

- $X \sim \text{Dir}(\alpha)$, $p(x) = \frac{\Gamma(\alpha_0)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} x_k^{\alpha_k - 1}$
- $\alpha \in \Delta^{K-1}$ and $\alpha_0 = \sum_{k=1}^{K} \alpha_k$
- α_0 controls the strength of the distribution (how peaked it is), and the α_k's control where the peak occurs
- Mean: $\mathbb{E}(x_k) = \alpha_k / \alpha_0$



- Useful to quantify epistemic and aleatoric uncertainties in classification problems

Markov Chains

- Let $(x_t)_{t\in\mathbb{N}}$ be a sequence of elements of $\mathbb{R}^D$
- Markov property: $p(x_{t+\tau} \mid x_1, \ldots, x_t) = p(x_{t+\tau} \mid x_t)$

Markov Chains

- Let $(x_t)_{t \in \mathbb{N}}$ be a sequence of elements of $\mathbb{R}^D$
- Markov property: $p(x_{t+\tau} \mid x_1, \ldots, x_t) = p(x_{t+\tau} \mid x_t)$
- In turn, we have that $p(x_1, \ldots, x_T) = p(x_1) \prod_{t=2}^{T} \underbrace{p(x_t \mid x_{t-1})}_{\text{transition function}}$
  - This is a Markov model (MM)
- If $p(x_t \mid x_{t-1})$ is indep. of time, the MM is called stationary

Markov Chains

- Let $\mathcal{X} = \{1, \ldots, K\}$, so that the MM is a finite-state Markov chain
- Transition function $p(X_t \mid X_{t-1})$ can be written as a $K \times K$ transition matrix A
  - $A_{ij} = p(X_t = j \mid X_{t-1} = i)$
  - Stochastic matrix: Each row i is such that $\sum_j A_{ij} = 1$

Markov Chains

- Let $\mathcal{X} = \{1, \ldots, K\}$, so that the MM is a finite-state Markov chain
- Transition function $p(X_t \mid X_{t-1})$ can be written as a $K \times K$ transition matrix A
  - $A_{ij} = p(X_t = j \mid X_{t-1} = i)$
  - Stochastic matrix: Each row i is such that $\sum_j A_{ij} = 1$
- n-step transition matrix $A_{ij}(n) := p(X_{t+n} = j \mid X_t = i)$
- Chapman-Kolmogorov:

$$A_{ij}(m+n) = \sum_{k=1}^{K} A_{ik}(m) A_{kj}(n)$$

or equivalently, $A(m+n) = A(m)A(n)$, and so $A(n) = A^n$

Markov Chains

- Let $\mathcal{X} = \{1, \ldots, K\}$, so that the MM is a finite-state Markov chain
- Transition function $p(X_t \mid X_{t-1})$ can be written as a $K \times K$ transition matrix A
  - $A_{ij} = p(X_t = j \mid X_{t-1} = i)$
  - Stochastic matrix: Each row i is such that $\sum_j A_{ij} = 1$
- n-step transition matrix $A_{ij}(n) := p(X_{t+n} = j \mid X_t = i)$
- Chapman-Kolmogorov:

$$A_{ij}(m+n) = \sum_{k=1}^{K} A_{ik}(m) A_{kj}(n)$$

  or equivalently, $A(m+n) = A(m)A(n)$, and so $A(n) = A^n$
- MLE for this MM: (Murphy, 2023, § 2.6.3.1)

Markov Chains

- Stationary distribution: intuitively, it is the long term distribution over states
- A: one-step transition matrix
- $\pi_t(j) = p(X_t = j)$: prob. of being in state j at time t
- π_0: initial distribution over states

Markov Chains

- Stationary distribution: intuitively, it is the long term distribution over states
- A: one-step transition matrix
- $\pi_t(j) = p(X_t = j)$: prob. of being in state j at time t
- π_0: initial distribution over states
- Then, $\pi_1(j) = \sum_i \pi_0(i) A_{ij}$, or in matrix notation, $\pi_1 = \pi_0 A$
- Stationary distribution: π such that $\pi = \pi A$
  - To compute it, solve the eigenvector equation $A^\top v = v$, and put $\pi = v$,
  - v is an eigenvector with eigenvalue 1
  - More general way of finding π: (Murphy, 2023, § 2.6.4.1)
- Stationary distributions need not always exist (Murphy, 2023, § 2.6.4.3 - 2.6.4.4)

Statistics

- Probability theory: modeling the distribution over observed data outcomes D given known parameters θ by computing $p(D \mid \theta)$
- Statistics: inverse problem. Infer the unknown parameters θ given observations, i.e. compute $p(\theta \mid D)$

Bayesian Statistics: Basic Concepts

- Parameter θ as unknown (rv), and data D as fixed and known
- Represent uncertainty about θ, after seeing data D, by computing the posterior distribution via Bayes' rule

$$p(\theta \mid D) = \frac{p(\theta)p(D \mid \theta)}{\int_{\Theta} p(\theta')p(D \mid \theta')\mathrm{d}\theta'} \propto p(\theta)p(D \mid \theta)$$

  - Prior: $p(\theta)$, represents beliefs about parameter before seeing the data
  - Likelihood: $p(D \mid \theta)$, represents beliefs about what data we expect to see, for each setting of the parameters
  - Marginal likelihood: $p(D) = \int_{\Theta} p(\theta')p(D \mid \theta')\mathrm{d}\theta'$, normalization constant, crucial in Bayesian Model Selection (BMS)
- Example: see (Murphy, 2023, § 3.2.1)
  - If we assume iid data, then $p(D \mid \theta) = \prod_{y \in D} p(y \mid \theta)$
  - $p(y \mid \theta)$: distributions we introduced before, e.g. a Binomial

Bayesian Statistics: MAP

- Maximum A Posteriori estimate (MAP):

$$\hat{\theta}_{\text{map}} = \arg\max_{\theta \in \Theta} p(\theta \mid D) = \arg\max_{\theta \in \Theta} [\log p(\theta) + \log p(D \mid \theta)]$$

- It is the posterior mode (most probable value)
- Confront it with the MLE

$$\arg\max_{\theta \in \Theta} p(D \mid \theta) = \arg\max_{\theta \in \Theta} \log p(D \mid \theta)$$

  - An extra component coming from the prior $p(\theta)$
  - If we use uniform prior $p(\theta) \propto 1$, MAP = MLE

Bayesian Statistics: MAP

- Maximum A Posteriori estimate (MAP):

$$\hat{\theta}_{\text{map}} = \arg\max_{\theta \in \Theta} p(\theta \mid D) = \arg\max_{\theta \in \Theta} [\log p(\theta) + \log p(D \mid \theta)]$$

- It is the posterior mode (most probable value)
- Confront it with the MLE

$$\arg\max_{\theta \in \Theta} p(D \mid \theta) = \arg\max_{\theta \in \Theta} \log p(D \mid \theta)$$

  - An extra component coming from the prior $p(\theta)$
  - If we use uniform prior $p(\theta) \propto 1$, MAP $=$ MLE
- In the case of probability models having missing data and/or hidden variables, we can compute MAP or MLE via the expectation maximization (EM) algorithm (Murphy, 2023, § 6.5.3)

Bayesian Statistics: HDRs

- First find threshold $p^\star$ such that

$$1 - \alpha = \int_{\{\theta \in \Theta : p(\theta|D) > p^\star\}} p(\theta \mid D) \mathrm{d}\theta$$

  - $\alpha \in [0, 1]$ chosen by the user
- Then define the Highest Density Region (HDR) as:

$$C_\alpha(D) = \{\theta \in \Theta : p(\theta \mid D) \geq p^\star\}$$

- It is the narrowest parameter region containing $100 \times (1 - \alpha)\%$ of the posterior probability mass
- Crucial in Uncertainty Quantification research Coolen (1992)

Bayesian Statistics: Posterior Predictive

- Posterior Predictive Distribution:

$$p(y \mid D) = \int_{\Theta} p(y \mid \theta) p(\theta \mid D) \mathrm{d}\theta \tag{1}$$

- Given the data D we observed, it tells us what is the probability that the next observation is some value y
- Laplace's rule of succession: see (Murphy, 2023, § 3.2.1.8)
- A Bayesian approach allows us to disentangle between EU and AU

Bayesian Statistics: Posterior Predictive

- In ML: interested in predicting outcomes y given input features x
- Use conditional probability of the form $p(y \mid x, \theta)$ (e.g. coming from a neural network)
- (Conditional) likelihood is $p(D \mid \theta) = \prod_{(x,y) \in D} p(y \mid x, \theta)$

Bayesian Statistics: Posterior Predictive

- In ML: interested in predicting outcomes y given input features x
- Use conditional probability of the form $p(y \mid x, \theta)$ (e.g. coming from a neural network)
- (Conditional) likelihood is $p(D \mid \theta) = \prod_{(x,y)\in D} p(y \mid x, \theta)$
- Eq. (1) then becomes

$$p(y \mid x, D) = \int_{\Theta} p(y \mid x, \theta) p(\theta \mid D) \mathrm{d}\theta$$

- By integrating out the unknown parameters, we reduce the chance of overfitting
  - We are computing the weighted average of predictions from an infinite number of models

Frequentist Statistics: Basic Concepts

- Frequentist statistics: uncertainty by calculating how a quantity estimated from data (e.g. a parameter) would change if the data were changed
  - Captured by the sampling distribution of an estimator
- This notion of variation across repeated trials: uncertainty modeling by the frequentist approach

Frequentist Statistics: Sampling Distribution

- Estimator: decision procedure that specifies what action to take given some observed data D
  - Parameter estimation: the action space is to return a parameter vector via function δ, so $\hat{\theta} = \delta(D)$, e.g. the MLE
- Sampling distribution of an estimator: distribution of results we would see if we applied the estimator multiple times to different datasets sampled from some distribution
  - Parameter estimation: it is the distribution of $\hat{\theta}$, viewed as a random variable that depends on the random sample D

Frequentist Statistics: Sampling Distribution

- Sample S different datasets, each of size N, from the true model $p(x \mid \theta^\star)$ to generate

$$\tilde{D}^{(s)} = \{x_n \sim p(x_n \mid \theta^\star)\}_{n=1}^{N}, \quad s \in \{1, \ldots, S\}$$

  - Denote this $\tilde{D}^{(s)} \sim \theta^\star$ for brevity
- Apply the estimator δ to each $\tilde{D}^{(s)}$ to get a set of estimates $\{\hat{\theta}^{(s)} = \delta(\tilde{D}^{(s)})\}_{s=1}^{S}$
- Sampling distribution:

$$p_{\text{sample}}(\theta \mid \{\tilde{D}^{(s)}\}_{s=1}^{S}) = \frac{1}{S} \sum_{s=1}^{S} \text{Dirac}(\theta - \hat{\theta}^{(s)})$$

  - $p_{\text{sample}}(\theta \mid \{\tilde{D}^{(s)}\}_{s=1}^{S}) \xrightarrow[S \to \infty]{} p(\delta(\tilde{D}) = \theta \mid \tilde{D} \sim \theta^\star)$

Frequentist Statistics: Parametric Bootstrap

- When estimator is a complex function of the data, or when sample size small: approximate sampling distribution using the bootstrap
- Since $\theta^\star$ is unknown, generate each sampled dataset using $\hat{\theta} = \delta(D)$ instead of $\theta^\star$,

$$\tilde{D}^{(s)} = \{x_n \sim p(x_n \mid \hat{\theta})\}_{n=1}^{N}, \quad s \in \{1, \ldots, S\}$$

- The rest stays the same

Frequentist Statistics: Asymptotic Normality of sampling distribution of MLE

Theorem

Under various technical conditions, we have

$$\sqrt{N}(\hat{\theta}_{\text{MLE}} - \theta^\star) \xrightarrow[N\to\infty]{\text{dist}} \mathcal{N}(0, F(\theta^\star)^{-1}),$$

where $F(\theta^\star)$ is the Fisher information matrix (Murphy, 2023, § 3.3.4)

- As the sample size goes to infinity, the sampling distribution of the MLE will converge to a Gaussian centered on the true parameter, with a precision equal to the Fisher information

Frequentist Statistics: Drawbacks

- Frequentist Statistics has some counterintuitive properties (Murphy, 2023, § 3.3.5 - 3.3.6)
- Popular because easy, taught at UG level, sometimes faster to implement than Bayesian
- "Inside every Non-Bayesian, there is a Bayesian struggling to get out", D. Lindley, cf. Jaynes (2002)
- CAREFUL: Bayesian approach is only as correct as its modeling assumptions
  - Check sensitivity of the conclusions to the choice of prior (and likelihood): BMS

Selecting the Prior: Conjugate Priors

- A prior $p(\theta) \in \mathcal{F}$ is a conjugate prior for a likelihood function $p(D \mid \theta)$ if the posterior is in the same parameterized family as the prior, i.e. $p(\theta \mid D) \in \mathcal{F}$
- That is, $\mathcal{F}$ is closed under Bayesian updating
- Conjugate priors simplify the computation of the posterior (Murphy, 2023, § 3.4)

Selecting the Prior: Noninformative Priors

- When we have little or no domain specific knowledge, desirable to use a noninformative prior, to "let the data speak for itself"
- No unique way to define such priors, and they all encode some kind of knowledge
  - Better to use the term minimally informative prior (Murphy, 2023, § 3.5)

Selecting the Prior: Hierarchical Priors

- Parameters ξ of the prior distribution $p(\theta)$ are called hyperparameters
- If unknown, put a prior on them; this defines a hierarchical Bayesian model $\xi \to \theta \to D$
- Assume the prior on the hyper-parameters is fixed, so the joint distribution has the form

$$p(\xi, \theta, D) = p(\xi)p(\theta \mid \xi)p(D \mid \theta)$$

- Use hierarchical approach when we have $J > 1$ related datasets D_j, each with their own parameters θ_j
- Inferring $p(\theta_j \mid D_j)$ independently for each group j can give poor results if D_j is a small dataset
- A hierarchical Bayesian model lets us borrow statistical strength from groups with lots of data to help groups with little data
  - (Murphy, 2023, § 3.6)

Selecting the Prior: Empirical Bayes

- Computationally convenient approximation of hierarchical Bayes
- First compute a point estimate of the hyperparameters, $\hat{\xi}$, and then compute the conditional posterior, $p(\theta \mid \hat{\xi}, D)$
- For example, type II MLE

$$\hat{\xi}_{\text{MML}}(D) = \arg\max_{\xi} p(D \mid \xi) = \arg\max_{\xi} \int_{\theta} p(D \mid \theta) p(\theta \mid \xi) \mathrm{d}\theta$$

- Once we have estimated $\hat{\xi}$, we compute the posterior $p(\theta \mid \hat{\xi}, D)$ in the usual way (Murphy, 2023, § 3.7)
- Notice that we are "cheating": we use data twice, once to form our prior belief, and once again for our likelihood

Other Statistical Concepts

- Model selection: we have a set of different models $\mathcal{M}$, each of which may fit the data to different degrees, and each of which may make different assumptions
  - How to pick the best model from this set, or to average over all of them
  - Assumed that the "true" model is in $\mathcal{M}$ (Murphy, 2023, § 3.8)

- Model checking: Bayesian inference is "optimal", but only if the modeling assumptions are correct. How to assess if a model is reasonable?
  - We assume that we do not have a specific alternative model in mind
  - We see if the data we observe is "typical" of what we might expect if our model were correct (Murphy, 2023, § 3.9)

References I

Thomas Augustin, Frank P. A. Coolen, Gert de Cooman, and Matthias C. M. Troffaes. *Introduction to imprecise probabilities*. John Wiley & Sons,, West Sussex, England, 2014.

Frank P. A. Coolen. Imprecise highest density regions related to intervals of measures. *Memorandum COSOR*, 9254, 1992.

Bruno de Finetti. *Theory of Probability*, volume 1. New York : Wiley, 1974.

Bruno de Finetti. *Theory of Probability*, volume 2. New York : Wiley, 1975.

Edwin T. Jaynes. *Probability Theory. The Logic of Science*. Cambridge University Press: Cambridge, 2002.

Kevin P. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, 2023. URL `http://probml.github.io/book2`.

References II

Peter Walley. *Statistical reasoning with imprecise probabilities*, volume 42 of *Monographs on Statistics and Applied Probability*. London : Chapman and Hall, 1991.