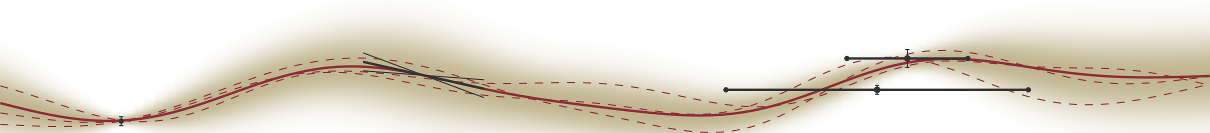
Probabilistic Machine Learning
Lecture 03
Continuous Variables

Philipp Hennig

24 April 2023

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Faculty of Science
Department of Computer Science
Chair for the Methods of Machine Learning



– Warning –
Theory incoming

We need to talk about

- Derived quantities: Functions of elementary events
- Continuous quantities: Measures on the continuum
- Functions of continuous quantities

Recap from Lecture 1

Plausibility as a Measure

[A.N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*, 1933]

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Definition (σ-algebra, measurable sets & spaces)

Let E be a space of *elementary events*. Consider the power set 2^E, and let $\mathfrak{F} \subset 2^E$ be a set of subsets of E. Elements of $\mathfrak{F}$ are called *random events*. If $\mathfrak{F}$ satisfies the following properties, it is called a σ-**algebra**.

1. $E \in \mathfrak{F}$ II.
2. $(A \in \mathfrak{F}) \Rightarrow (\neg A := E - A \in \mathfrak{F})$ I.
3. $(A_1, A_2, \dots \in \mathfrak{F}) \Rightarrow \bigcup_{i=1}^{\mathbb{N}} A_i \in \mathfrak{F}$ I.

(this implies, by de Morgan's Laws, $\bigcap_{i=1}^{\mathbb{N}} A_i \in \mathfrak{F}$, and thus also $\varnothing \in \mathfrak{F}$. If E is countable, then 2^E is a σ-algebra). If $\mathfrak{F}$ is a σ-algebra, its elements are called **measurable sets**, and $(E, \mathfrak{F})$ is called a **measurable space** (or **Borel space**).

If Ω is countable, then 2^Ω is a σ-algebra, and everything is easy.

Recap from Lecture 1

Plausibility as a Measure

[A.N. Kolmogorov. *Grundbegriffe der Wahrscheinlichkeitsrechnung*, 1933]

Definition (Measure & Probability Measure)

Let $(\Omega, \mathfrak{F})$ be a **measurable space** (aka. Borel space). A nonnegative real function $P : \mathfrak{F} \to \mathbb{R}_{0,+}$ (III.) is called a **measure** if it satisfies the following properties:

1. $P(\varnothing) = 0$
2. For any countable sequence $\{A_i \in \mathfrak{F}\}_{i=1,\dots,}$ of pairwise disjoint sets ($A_i \cap A_j = \varnothing$ if $i \neq j$), P satisfies **countable additivity** (aka. σ-**additivity**):

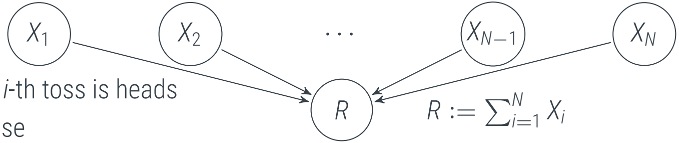
$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i). \qquad \text{(V.)}$$

The measure P is called a **probability measure** if $P(\Omega) = 1$. IV.
(For probability measures, 1. is unnecessary). Then, $(\Omega, \mathfrak{F}, P)$ is called a **probability space**.

What about derived quantities?

Random variables

A bent coin has probability f of coming up heads. The coin is tossed N times. What is the probability distribution of the number of heads r?



$$X_i := \begin{cases} 1 & \text{if } i\text{-th toss is heads} \\ 0 & \text{else} \end{cases}$$

$$R := \sum_{i=1}^{N} X_i$$

- For $\boldsymbol{X} = [X_1, \dots, X_N]$, we have $\Omega = \{0, 1\}^N$.
- But what about $R \in [0, \dots, N] \subset \mathbb{N}$? It's not part of Ω.

Building new Probability Distributions from old ones

Random Variables

Definition (Measurable Functions, Random Variables)

Let $(\Omega, \mathfrak{F})$ and $(\Gamma, \mathfrak{G})$ be two measurable spaces (i.e. spaces with σ-algebras). A function $X : \Omega \rightharpoonup \Gamma$ is called **measurable** if $X^{-1}(G) \in \mathfrak{F}$ for all $G \in \mathfrak{G}$. If there is, additionally, a probability measure P on $(\Omega, \mathfrak{F})$, then X is called a **random variable**.

Definition (Distribution Measure)

Let $X : \Omega \rightharpoonup \Gamma$ be a random variable. Then the **distribution measure** (or **law**) P_X of X is defined for any $G \subset \mathfrak{G}$ as

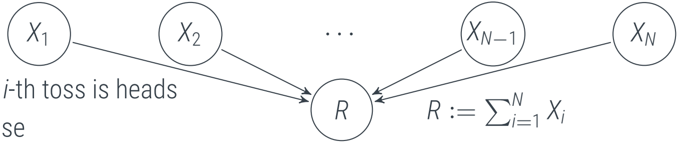
$$P_X(G) = P(X^{-1}(G)) = P(\{\omega \mid X(\omega) \in G\}).$$

Note: We will call essentially every variable X over which we construct probability measures random variables, even if the measure is directly constructed on X, because every variable can be thought of as the output of a function (if necessary, the identity).

Example: the Binomial Distribution

statistics of accumulated Bernoulli experiments

A bent coin has probability f of coming up heads. The coin is tossed N times. What is the probability distribution of the number of heads r?



$$X_i := \begin{cases} 1 & \text{if } i\text{-th toss is heads} \\ 0 & \text{else} \end{cases}$$

$$R := \textstyle\sum_{i=1}^N X_i$$

$$P(R = r) = \sum_{\omega \in \{\boldsymbol{X} | R = r\}} \prod_{i=1}^N P(X_i) = \sum_{\omega \in \{\boldsymbol{X} | R = r\}} f^r \cdot (1-f)^{N-r} := P(r \mid f, N)$$

- original space: $\Omega = \{0; 1\}^N$ (countably finite)
- σ-algebra: 2^Ω (the power set)
- random variable $R = \sum_{i=1}^N X_i \in [0, \dots, N] =: \Gamma \subset \mathbb{N}$.
- distribution (measure) / law of R: …

Example: the Binomial Distribution

statistics of accumulated Bernoulli experiments

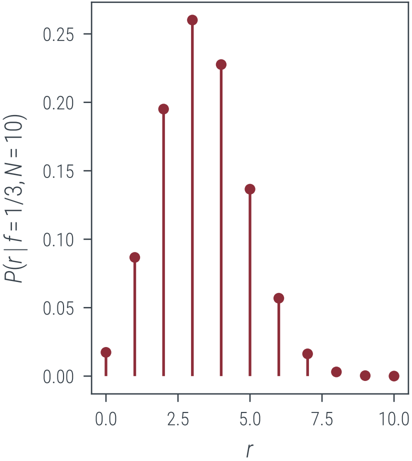
Definition: Let $R : \Omega \rightarrow \Gamma$ be a random variable. Then the **distribution measure** (or **law**) P_R of R is defined for any $G \subset \Gamma$ as

$$P_R(G) = P(R^{-1}(G)) = P(\{\omega \mid R(\omega) \in G\}).$$

The **distribution measure** of $R := \sum_{i=1}^N X_i$ is

$$\begin{aligned}P(r \mid f, N) &= (\#\text{ ways to choose } r \text{ from } N) \cdot f^r \cdot (1-f)^{N-r}\\ &= \frac{N!}{(N-r)! \cdot r!} \cdot f^r \cdot (1-f)^{N-r}\\ &= \binom{N}{r} \cdot f^r \cdot (1-f)^{N-r}\end{aligned}$$

Note: In the remainder of the course, will often **abuse notation** by writing $P(r)$ instead of $P(R=r)$ (recall again that $P(X) \neq P(Y)$!)



Now for the Real case …

some complications for continuous spaces

- in a countable space Ω, even 2^{Ω} is a σ-algebra.
- But in continuous spaces, such as $\Omega = \mathbb{R}^d$, not all sets are measurable.

Now for the Real case …

some complications for continuous spaces

- in a countable space Ω, even 2^Ω is a σ-algebra.
- But in continuous spaces, such as $\Omega = \mathbb{R}^d$, not all sets are measurable.

Example: Consider the set $S : \{e^{ix\pi} : x \in [0, 1) \subset \mathbb{R}\}$ of all points on the unit circle, and the group G (also a set) of rotations of $s \in S$ by **rational** numbers $se^{iq\pi} : q \in \mathbb{Q}$

- We note S is uncountable, but G is countable. Hence, S breaks up into *uncountably* many orbits $\{se^{iq\pi} : q \in \mathbb{Q}\}$ under G.
- If we have the *axiom of choice*, we can pick one point from each orbit, forming an *uncountable* subset $Z \subset S$ with the property that, for **all sets** of the form $\{e^i q\pi z : z \in Z\}$ for $q \in \mathbb{Q}$ are pairwise disjoint from each other.
- Thus, $S = \{\{e^{iq\pi}z : z \in Z\} : q \in \mathbb{Q}\}$, hence we have partitioned the circle into a **countable collection of disjoint sets**, which are also **pairwise disjoint**.

Proposition: Z is not measurable.
Proof: If Z has zero measure, then, by sigma-additivity, S has zero measure. If Z has non-zero measure, then, by sigma-additivity, S has infinite measure. □

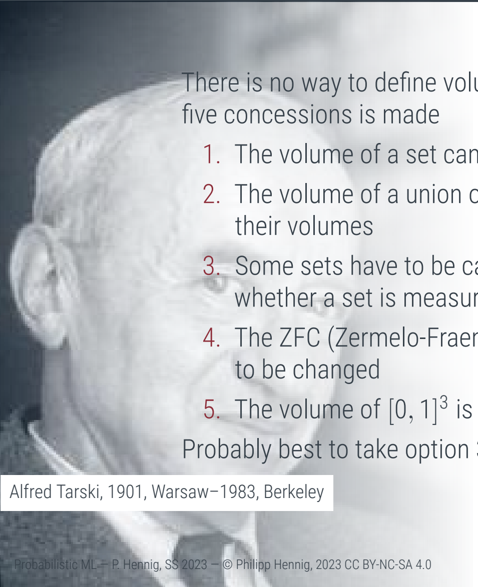
The Banach-Tarski Paradox

You thought you would get away without ever thinking about countability again after your BSc, didn’t you?

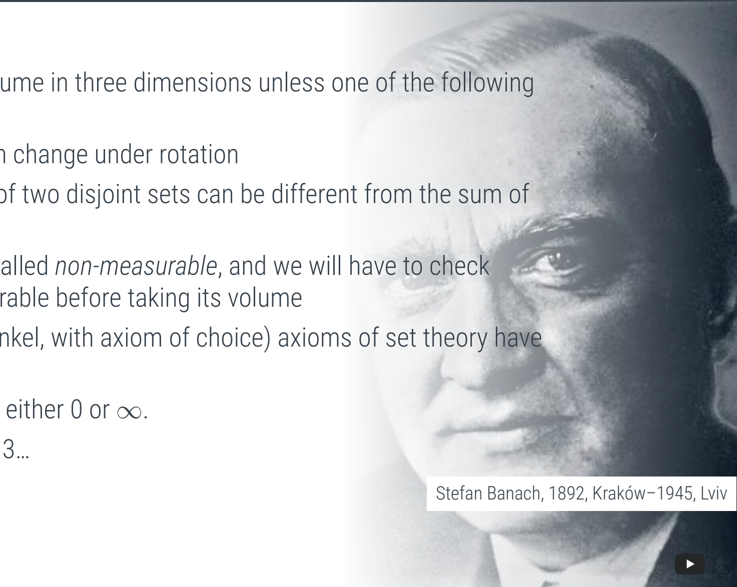
There is no way to define volume in three dimensions unless one of the following five concessions is made

1. The volume of a set can change under rotation
2. The volume of a union of two disjoint sets can be different from the sum of their volumes
3. Some sets have to be called *non-measurable*, and we will have to check whether a set is measurable before taking its volume
4. The ZFC (Zermelo-Fraenkel, with axiom of choice) axioms of set theory have to be changed
5. The volume of $[0, 1]^3$ is either 0 or ∞.

Probably best to take option 3…



Alfred Tarski, 1901, Warsaw–1983, Berkeley



Stefan Banach, 1892, Kraków–1945, Lviv

Taking the easy way out

Accepting that not everything is measurable

- in a countable space Ω, even 2^Ω is a σ-algebra.
- But in continuous spaces, such as $\Omega = \mathbb{R}^d$, not all sets are measurable.
- However, $\mathbb{R}^d$ is a *topological space*

Definition (Topology)

Let Ω be a space and τ be a collection of sets. We say τ is a **topology** on Ω if

- $\Omega \in \tau$, and $\varnothing \in \tau$
- *any* union of elements of τ is in τ
- any intersection of *finitely many* elements of τ is in τ.

The elements of the topology τ are called **open sets**. In the Euclidean vector space $\mathbb{R}^d$, the canonical topology is that of all sets U that satisfy $x \in U :\Rightarrow \exists \varepsilon > 0 : ((\|y - x\| < \varepsilon) \Rightarrow (y \in U))$.

From topologies to σ-algebras

for topological spaces, it's easy to define σ-algebras

Note that a topology is *almost* a σ-algebra:

Definition (σ-algebra, measurable sets & spaces)

Let E be a space, and $\mathfrak{F} \subset 2^E$ a set of subsets. We say $\mathfrak{F}$ is a σ**-algebra** if

1. $E \in \mathfrak{F}$ II.
2. $(A \in \mathfrak{F}) \Rightarrow (\neg A := E - A \in \mathfrak{F})$ I.
3. $(A_1, A_2, \dots \in \mathfrak{F}) \Rightarrow \bigcup_{i=1}^{\mathbb{N}} A_i \in \mathfrak{F}$ I.

(this implies, by de Morgan's Laws, $\bigcap_{i=1}^{\mathbb{N}} A_i \in \mathfrak{F}$, and thus also $\varnothing \in \mathfrak{F}$. If E is countable, then 2^E is a σ-algebra). If $\mathfrak{F}$ is a σ-algebra, its elements are called **measurable sets**, and $(E, \mathfrak{F})$ is called a **measurable space** (or **Borel space**).

Definition (Topology)

Let Ω be a space and τ be a collection of sets. We say τ is a **topology** on Ω if

- $\Omega \in \tau$, and $\varnothing \in \tau$
- *any* union of elements of τ is in τ
- any intersection of **finitely many** elements of τ is in τ.

(this implies $A \in \tau \Rightarrow$) The elements of the topology τ are called **open sets**. In the Euclidean vector space $\mathbb{R}^d$, the canonical topology is that of all sets U that satisfy

$x \in U :\Rightarrow \exists \varepsilon > 0 : ((\|y - x\| < \varepsilon) \Rightarrow (y \in U))$.

Almost done!

Standard settings

Definition (Borel algebra)

Let (Ω, τ) be a topological space. The **Borel σ-algebra** is the σ-algebra *generated* by τ. That is by taking τ and completing it to include infinite intersections of elements from τ and all complements in Ω to elements of τ.

- In this lecture, we will almost exclusively consider (random) variables defined on discrete or Euclidean spaces. In the latter case, the σ-algebra will not be mentioned but assumed to be the Borel σ-algebra.
- Consider $(\Omega, \mathfrak{F})$ and $(\Gamma, \mathfrak{G})$. If both $\mathfrak{F}$ and $\mathfrak{G}$ are Borel σ-algebras, then any continuous function X is measurable (and can thus be used to define a random variable). This is because, for continuous functions, pre-images of open sets are open sets.

Now that we can define (Borel) σ-algebras on continous spaces, we can define probability distribution measures. They might just be a bit unwieldy.

- **Random Variables** allow us to define derived quantities from atomic events
- **Borel σ-algebras** can be defined on all topological spaces, allowing us to define probabilities if the elementary space is continuous.

Note the connection to computability theory: ***measurable*** *functions* and ***computable*** *functions*. "Not all sets are **measurable**", and "not all languages are **computable**".

Cumulative Distributions

Connecting probabilities to integration

Definition (Cumulative Distribution Function (CDF))

Let $\mathfrak{B}$ be the Borel σ-algebra in $\mathbb{R}^d$. For probability measures P on $(\mathbb{R}^d, \mathfrak{B})$, the **cumulative distribution function** is the function

$$F(\boldsymbol{x}) = P\left(\prod_{i=1}^{d}(X_i < x_i)\right).$$

In particular for the univariate case $d = 1$, we have $F(x) = P\left((-\infty, x]\right)$.

Probability Densities

a convenient way to write things down

Definition (Probability Density Functions (pdf's))

A probability measure P on $(\mathbb{R}^d, \mathfrak{B})$ has a **density** p if p is a non-negative (Borel) measurable function on $\mathbb{R}^d$ satisfying, for all $B \in \mathfrak{B}$

$$P(B) = \int_B p(x)\, dx =: \int_B p(x_1, \dots, x_d)\, dx_1 \, \dots \, dx_d$$

In particular, if the CDF F of P is sufficiently differentiable, then P has a density, given by

$$p(\boldsymbol{x}) = \left. \frac{\partial^d F}{\partial x_1 \cdots \partial x_d} \right|_{\boldsymbol{x}} .$$

and, for $d = 1$,

$$P(a \leq X < b) = F(b) - F(a) = \int_a^b f(x)\, dx.$$

Densities Satisfy the Laws of Probability Theory

because integrals are linear operators

proof in Math for ML lecture

- For probability densities p on $(\mathbb{R}^d, \mathfrak{B})$ we have

$$P(E) \overset{(IV)}{=} 1 = \int_{\mathbb{R}^d} p(x)\, dx.$$

- Let $X = (X_1, X_2) \in \mathbb{R}^2$ be a random variable with density p_X on $\mathbb{R}^2$. Then the **marginal densities** of X_1 and X_2 are given by the **sum rule**

$$p_{X_1}(x_1) = \int_{\mathbb{R}} p_X(x_1, x_2)\, dx_2, \qquad p_{X_2}(x_2) = \int_{\mathbb{R}} p_X(x_1, x_2)\, dx_1$$

- The **conditional density** $p(x_1 \mid x_2)$ (for $p(x_2) > 0$) is given by the **product rule**

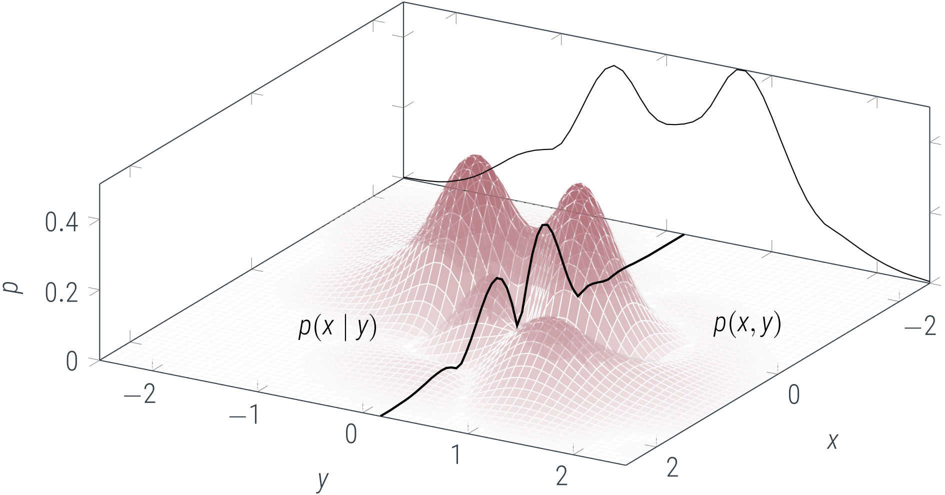
$$p(x_1 \mid x_2) = \frac{p(x_1, x_2)}{p(x_2)}$$

- **Bayes' Theorem** holds:

$$p(x_1 \mid x_2) = \frac{p(x_1) \cdot p(x_2 \mid x_1)}{\int p(x_1) \cdot p(x_2 \mid x_1)\, dx_1}.$$

A Graphical View

sketch



Change of Measure

The transformation law

Theorem (Change of Variable for Probability Density Functions)

Let X be a continuous random variable with PDF $p_X(x)$ over $c_1 < x < c_2$. And, let $Y = u(X)$ be a monotonic differentiable function with inverse $X = v(Y)$. Then the PDF of Y is

$$p_Y(y) = p_X(v(y)) \cdot \left|\frac{dv(y)}{dy}\right| = p_X(v(y)) \cdot \left|\frac{du(x)}{dx}\right|^{-1}.$$

Proof: for $u'(X) > 0$: $\forall\, d_1 = u(c_1) < y < u(c_2) = d_2$

$$F_Y(y) = P(Y \leq y) = P(u(X) \leq y) = P(X \leq v(y)) = \int_{c_1}^{v(y)} p_X(x)\, dx$$

$$p_Y(y) = \frac{\mathrm{d}F_Y(y)}{dy} = p_X(v(y)) \cdot \frac{dv(y)}{dy} = p_X(v(y)) \cdot \left|\frac{dv(y)}{dy}\right|$$

Change of Measure

The transformation law

Theorem (Change of Variable for Probability Density Functions)

Let X be a continuous random variable with PDF $p_X(x)$ over $c_1 < x < c_2$. And, let $Y = u(X)$ be a monotonic differentiable function with inverse $X = v(Y)$. Then the PDF of Y is

$$p_Y(y) = p_X(v(y)) \cdot \left|\frac{dv(y)}{dy}\right| = p_X(v(y)) \cdot \left|\frac{du(x)}{dx}\right|^{-1}.$$

Proof: for $u'(X) < 0$: $\forall\, d_2 = u(c_2) < y < u(c_1) = d_1$

$$F_Y(y) = P(Y \leq y) = P(u(X) \leq y) = P(X \geq v(y)) = 1 - P(X \leq v(y)) = 1 - \int_{c_1}^{v(y)} p_X(x)\,dx$$

$$p_Y(y) = \frac{\mathrm{d}F_Y(y)}{dy} = -p_X(v(y)) \cdot \frac{dv(y)}{dy} = p_X(v(y)) \cdot \left|\frac{dv(y)}{dy}\right|$$

Change of Measure

The transformation law

Theorem (Transformation Law, general)

Let $X = (X_1, \dots, X_d)$ have a joint density p_X. Let $g : \mathbb{R}^d \rightarrow \mathbb{R}^d$ be continously differentiable and injective, with non-vanishing Jacobian J_g. Then $Y = g(X)$ has density

$$p_Y(y) = \begin{cases} p_X(g^{-1}(y)) \cdot |J_{g^{-1}}(y)| & \textit{if } y \textit{ is in the range of } g, \\ 0 & \textit{otherwise.} \end{cases}$$

The Jacobian J_g is the $d \times d$ matrix with

$$[J_g(x)]_{ij} = \frac{\partial g_i(x)}{\partial x_j}.$$

DEMO

live: `streamlit cloud`

local:

```
▶ git clone https://github.com/philipphennig/ProbML_Apps.git
▶ cd ProbML_Apps/03
▶ pip install -r requirements.txt
▶ streamlit run Lecture_03.py
```

- **Probability Density Functions (pdf's)** distribute probability across continuous domains.
- they satisfy "the rules of probability":

$$\int_{\mathbb{R}^d} p(x)\, dx = 1$$

$$p_{X_1}(x_1) = \int_{\mathbb{R}} p_X(x_1, x_2)\, dx_2 \qquad \textbf{sum rule}$$

$$p(x_1 \mid x_2) = \frac{p(x_1, x_2)}{p(x_2)} \qquad \textbf{product rule}$$

$$p(x_1 \mid x_2) = \frac{p(x_1) \cdot p(x_2 \mid x_1)}{\int p(x_1) \cdot p(x_2 \mid x_1)\, dx_1} \qquad \textbf{Bayes' Theorem.}$$

- Not every measure has a density, but all pdfs define measures
- Densities transform under continuously differentiable, injective functions $g : x \mapsto y$ with non-vanishing Jacobian as

$$p_Y(y) = \begin{cases} p_X(g^{-1}(y)) \cdot |J_{g^{-1}}(y)| & \text{if } y \text{ is in the range of } g, \\ 0 & \text{otherwise.} \end{cases}$$

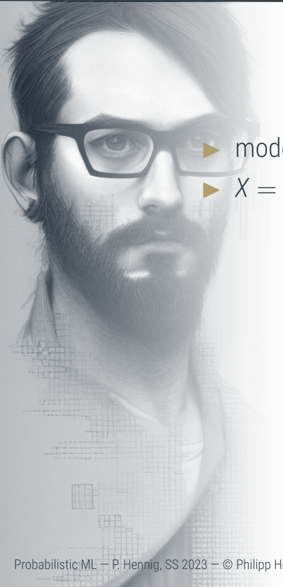
An example

Pierre Simon Laplace, 1814

What is the probability π for a person to be wearing glasses?

An example

Pierre Simon Laplace, 1814

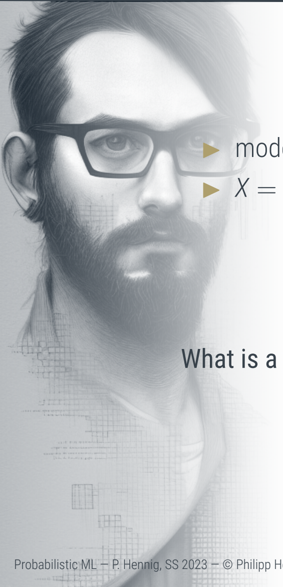


What is the probability π for a person to be wearing glasses?

- model probability as random variable π ranging in $[0, 1]$
- $X =$ person is wearing glasses. How do we do inference?

An example

Pierre Simon Laplace, 1814



What is the probability π for a person to be wearing glasses?

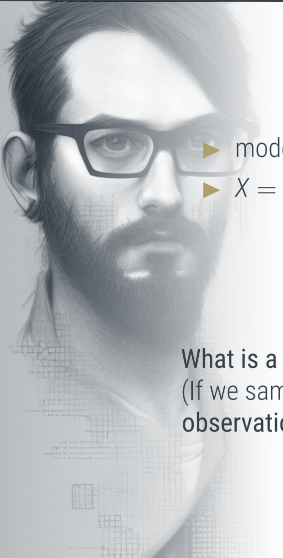
- model probability as random variable π ranging in $[0, 1]$
- $X =$ person is wearing glasses. How do we do inference?

$$p(\pi \mid X) = \frac{p(X \mid \pi)\, p(\pi)}{p(X)} = \frac{p(X \mid \pi)\, p(\pi)}{\int p(X \mid \pi)\, p(\pi)\, d\pi}$$

What is a good prior?

An example

Pierre Simon Laplace, 1814



What is the probability π for a person to be wearing glasses?

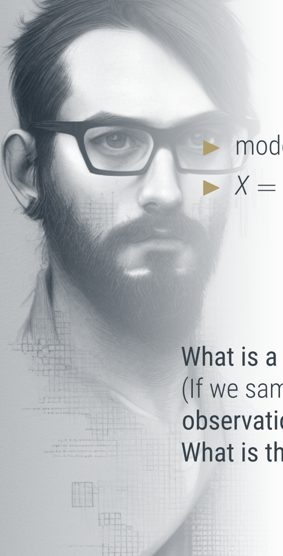
- model probability as random variable π ranging in $[0, 1]$
- $X =$ person is wearing glasses. How do we do inference?

$$p(\pi \mid X) = \frac{p(X \mid \pi)\, p(\pi)}{p(X)} = \frac{p(X \mid \pi)\, p(\pi)}{\int p(X \mid \pi)\, p(\pi)\, d\pi}$$

What is a good prior? uniform for $\pi \in [0, 1]$, i.e. $p(\pi) = 1$, zero elsewhere.
(If we sample independently) **What is the likelihood for a positive or a negative observation?**

An example

Pierre Simon Laplace, 1814



What is the probability π for a person to be wearing glasses?

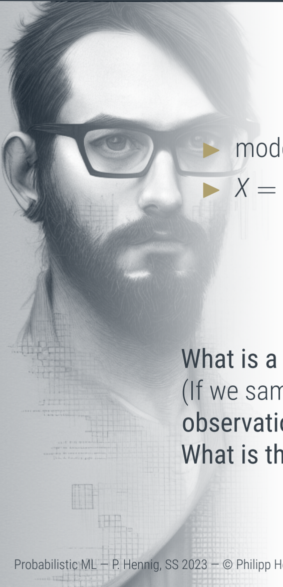
- model probability as random variable π ranging in $[0,1]$
- $X =$ person is wearing glasses. How do we do inference?

$$p(\pi \mid X) = \frac{p(X \mid \pi)\, p(\pi)}{p(X)} = \frac{p(X \mid \pi)\, p(\pi)}{\int p(X \mid \pi)\, p(\pi)\, d\pi}$$

What is a good prior? uniform for $\pi \in [0,1]$, i.e. $p(\pi) = 1$, zero elsewhere.
(If we sample independently) **What is the likelihood for a positive or a negative observation?** $p(X = 1 \mid \pi) = \pi$; $p(X = 0 \mid \pi) = 1 - \pi$.
What is the posterior after n **positive,** m **negative observations?**

An example

Pierre Simon Laplace, 1814



What is the probability π for a person to be wearing glasses?

- model probability as random variable π ranging in $[0, 1]$
- $X =$ person is wearing glasses. How do we do inference?

$$p(\pi \mid X) = \frac{p(X \mid \pi)\, p(\pi)}{p(X)} = \frac{p(X \mid \pi)\, p(\pi)}{\int p(X \mid \pi)\, p(\pi)\, d\pi}$$

What is a good prior? uniform for $\pi \in [0, 1]$, i.e. $p(\pi) = 1$, zero elsewhere.
(If we sample independently) **What is the likelihood for a positive or a negative observation?** $p(X = 1 \mid \pi) = \pi;$ $p(X = 0 \mid \pi) = 1 - \pi.$
What is the posterior after n positive, m negative observations?

$$p(\pi \mid n, m) = \frac{\pi^n (1-\pi)^m \cdot 1}{\int \pi^n (1-\pi)^m \cdot 1\, d\pi} = \frac{\pi^n (1-\pi)^m}{B(n+1, m+1)}$$

DEMO

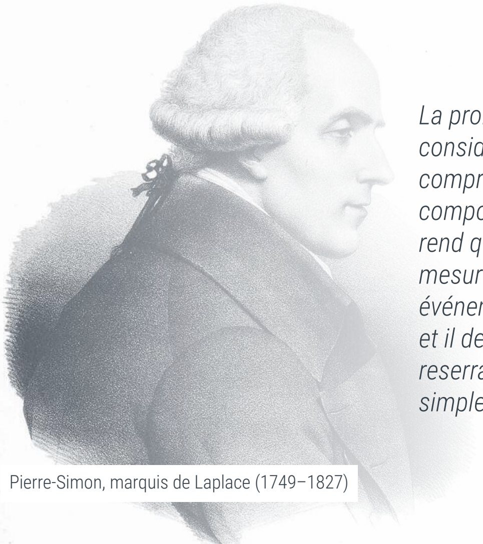
live: `streamlit cloud`

local:

```
▶ git clone https://github.com/philipphennig/ProbML_Apps.git
▶ cd ProbML_Apps/03
▶ pip install -r requirements.txt
▶ streamlit run Lecture_03.py
```

The Invention of Inference

It's actually "Laplace's Theorem"



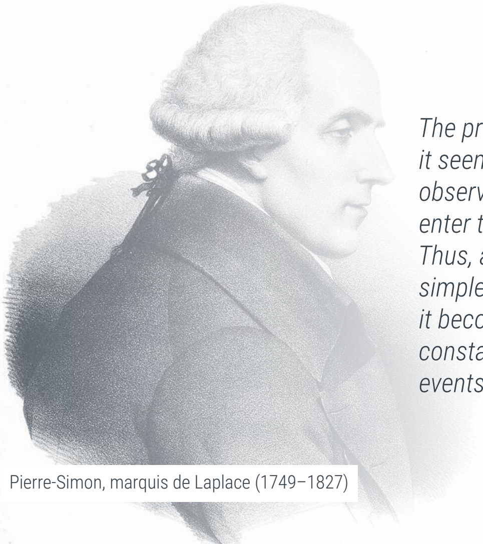
Pierre-Simon, marquis de Laplace (1749–1827)

La probabilité de la plupart des événemens simples, est inconnue; en la considérant à priori, elle nous paraît susceptible de toutes les valeurs comprises entre zéro et l'unité; mais sie l'on a observé un résultat composé de plusieurs de ces événemens, la manière dont ils y entrent, rend quelques-unes de ces valeurs plus probables que les autres. Ainsi à mesure que les résultat observé se compose par le développement des événemens simples, leur vraie possibilité se fait de plus en plus connaître, et il devient de plus en plus probable qu'elle tombe dans des limites qui se reserrant sans cesse, finiraient par coïncider, si le nombre des événemens simples devenait infini. *Theorie Analytique des Probabilités*, 1814, p. 363

Translated by a Deep Network, assisted by a human

The Invention of Inference

It's actually "Laplace's Theorem"



Pierre-Simon, marquis de Laplace (1749–1827)

The probability of most simple events is unknown. Considering it a priori, it seems susceptible to all values between zero and unity. But if one has observed a result composed of several of these events, the way they enter them makes some of these values more probable than the others. Thus, as the observed results are composed by the development of simple events, their real possibility becomes more and more known, and it becomes more and more probable that it falls within limits that constantly tighten, would end up coinciding if the number of simple events became infinite. *Theorie Analytique des Probabilités*, 1814, p. 363

Translated by a Deep Network, assisted by a human

Let’s be more careful with notation!
(but only once more, then we’ll be sloppy)

Example — inferring probability of wearing glasses (2)

Step 1: Construct σ-algebra

Represent all unknowns as random variables (RVs)

- probability to wear glasses is represented by RV Y
- five observations are represented by RVs X_1, X_2, X_3, X_4, X_5

Example — inferring probability of wearing glasses (2)

Step 1: Construct σ-algebra

Represent all unknowns as random variables (RVs)

- probability to wear glasses is represented by RV Y
- five observations are represented by RVs X_1, X_2, X_3, X_4, X_5

Possible values of the RVs

- Y takes values $\pi \in [0, 1]$
- X_1, X_2, X_3, X_4, X_5 are binary, i.e. values 0 and 1

Example — inferring probability of wearing glasses (2)

Step 1: Construct σ-algebra

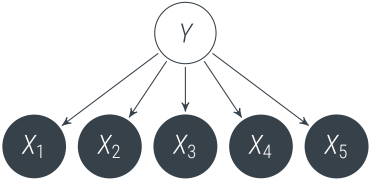
Represent all unknowns as random variables (RVs)

- probability to wear glasses is represented by RV Y
- five observations are represented by RVs X_1, X_2, X_3, X_4, X_5

Possible values of the RVs

- Y takes values $\pi \in [0, 1]$
- X_1, X_2, X_3, X_4, X_5 are binary, i.e. values 0 and 1

Graphical representation



Example — inferring probability of wearing glasses (2)

Step 1: Construct σ-algebra

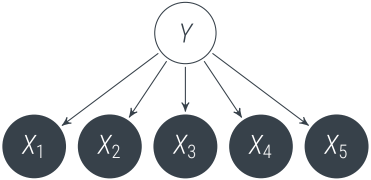
Represent all unknowns as random variables (RVs)

- probability to wear glasses is represented by RV Y
- five observations are represented by RVs X_1, X_2, X_3, X_4, X_5

Possible values of the RVs

- Y takes values $\pi \in [0, 1]$
- X_1, X_2, X_3, X_4, X_5 are binary, i.e. values 0 and 1

Graphical representation



Generative model and joint probability

- we abbreviate $Y = \pi$ as π, $X_i = x_i$ as x_i
- $p(\pi)$ is the prior of Y, written fully $p(Y = \pi)$
- $p(x_i | \pi)$ is the likelihood of observation x_i
- note that the likelihood is a function of π

Example — inferring probability of wearing glasses (3)

Step 2: Define probability space, taking care of conditional independence

Probability of wearing glasses without observations

$$p(\pi|\text{“nothing”}) = p(\pi)$$

Example — inferring probability of wearing glasses (3)

Step 2: Define probability space, taking care of conditional independence

Probability of wearing glasses without observations

$$p(\pi|\text{“nothing”}) = p(\pi)$$

Probability of wearing glasses after one observation

$$p(\pi|x_1) = \frac{p(x_1|\pi)p(\pi)}{\int p(x_1|\pi)p(\pi)\,d\pi} = Z_1^{-1} p(x_1|\pi)p(\pi)$$

Example — inferring probability of wearing glasses (3)

Step 2: Define probability space, taking care of conditional independence

Probability of wearing glasses without observations

$$p(\pi|\text{“nothing”}) = p(\pi)$$

Probability of wearing glasses after one observation

$$p(\pi|x_1) = \frac{p(x_1|\pi)p(\pi)}{\int p(x_1|\pi)p(\pi)\,d\pi} = Z_1^{-1} p(x_1|\pi)p(\pi)$$

Probability of wearing glasses after two observations

$$p(\pi|x_1,x_2) = Z_2^{-1} p(x_2|x_1,\pi)p(x_1|\pi)p(\pi) = Z_2^{-1} p(x_2|\pi)p(x_1|\pi)p(\pi)$$

Example – inferring probability of wearing glasses (3)

Step 2: Define probability space, taking care of conditional independence

Probability of wearing glasses without observations

$$p(\pi | \text{“nothing”}) = p(\pi)$$

Probability of wearing glasses after one observation

$$p(\pi | x_1) = \frac{p(x_1 | \pi) p(\pi)}{\int p(x_1 | \pi) p(\pi) \, d\pi} = Z_1^{-1} p(x_1 | \pi) p(\pi)$$

Probability of wearing glasses after two observations

$$p(\pi | x_1, x_2) = Z_2^{-1} p(x_2 | x_1, \pi) p(x_1 | \pi) p(\pi) = Z_2^{-1} p(x_2 | \pi) p(x_1 | \pi) p(\pi)$$

...

Probability of wearing glasses after five observations

$$p(\pi | x_1, x_2, x_3, x_4, x_5) = Z_5^{-1} \left(\prod_{i=1}^{5} p(x_i | \pi) \right) p(\pi)$$

Example — inferring probability of wearing glasses (4)

Step 3: Define analytic forms of generative model

What is the likelihood?

$$p(x_1|\pi) = \begin{cases} \pi & \text{for } x_1 = 1 \\ 1 - \pi & \text{for } x_1 = 0 \end{cases}$$

Example — inferring probability of wearing glasses (4)

Step 3: Define analytic forms of generative model

What is the likelihood?

$$p(x_1|\pi) = \begin{cases} \pi & \text{for } x_1 = 1 \\ 1-\pi & \text{for } x_1 = 0 \end{cases}$$

More helpful RVs:

- RV N for the number of observations being 1 (with values n)
- RV M for the number of observations being 0 (with values m)

Example — inferring probability of wearing glasses (4)

Step 3: Define analytic forms of generative model

What is the likelihood?

$$p(x_1|\pi) = \begin{cases} \pi & \text{for } x_1 = 1 \\ 1-\pi & \text{for } x_1 = 0 \end{cases}$$

More helpful RVs:

- RV N for the number of observations being 1 (with values n)
- RV M for the number of observations being 0 (with values m)

Probability of wearing glasses after five observations

$$\begin{aligned} p(\pi|x_1, x_2, x_3, x_4, x_5) &= Z_5^{-1}\left(\prod_{i=1}^{5} p(x_i|\pi)\right) p(\pi) \\ &= Z_5^{-1}\pi^n(1-\pi)^m p(\pi) \\ &= p(\pi|n, m) \end{aligned}$$

Example — inferring probability of wearing glasses (5)

Step 4: make computationally convenient choices. Here: a **conjugate** prior

Posterior after seeing five observations:

$$p(\pi|n,m) = Z_5^{-1}\pi^n(1-\pi)^m p(\pi)$$

Example — inferring probability of wearing glasses (5)

Step 4: make computationally convenient choices. Here: a **conjugate** prior

Posterior after seeing five observations:

$$p(\pi|n,m) = Z_5^{-1} \pi^n (1-\pi)^m p(\pi)$$

What prior $p(\pi)$ would make the calculations easy?

Example — inferring probability of wearing glasses (5)

Step 4: make computationally convenient choices. Here: a **conjugate** prior

Posterior after seeing five observations:

$$p(\pi|n,m) = Z_5^{-1}\pi^n(1-\pi)^m p(\pi)$$

What prior $p(\pi)$ would make the calculations easy?

$$p(\pi) = Z^{-1}\pi^{a-1}(1-\pi)^{b-1}$$

with parameters $a > 0, b > 0$

the Beta ***distribution*** *with parameter a and b*

Example — inferring probability of wearing glasses (5)

Step 4: make computationally convenient choices. Here: a **conjugate** prior

Posterior after seeing five observations:

$$p(\pi|n,m) = Z_5^{-1}\pi^n(1-\pi)^m p(\pi)$$

What prior $p(\pi)$ would make the calculations easy?

$$p(\pi) = Z^{-1}\pi^{a-1}(1-\pi)^{b-1}$$ with parameters $a > 0, b > 0$

the Beta ***distribution*** *with parameter a and b*

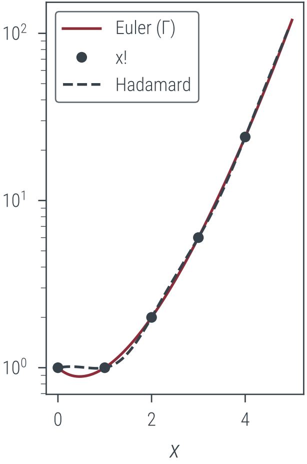
Let's give the normalization factor Z of the beta distribution a name!

$$B(a,b) = \int_0^1 \pi^{a-1}(1-\pi)^{b-1} d\pi$$

the Beta ***function*** *with parameters a and b*

Aside: The Eulerian Integrals

For an evening at the fireplace: [Philip J. Davis. *Leonhard Euler's Integral: A Historical Profile of the Gamma Function*, 1959]



For $m, n \in \mathbb{N}$ and $x, y, z \in \mathbb{C}$:

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\,dt$$

$$= \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} \text{ if } x+\bar{x}, y+\bar{y} > 0$$

$$B(m,n) = \frac{(m-1)!\,(n-1)!}{(m+n-1)!}$$

$$\Gamma(z) = \int_0^\infty e^{-t}t^{z-1}\,dt = \int_0^1 (-\log x)^{z+1}\,dx$$

$$\Gamma(n) = (n-1)!$$

Hadamard:

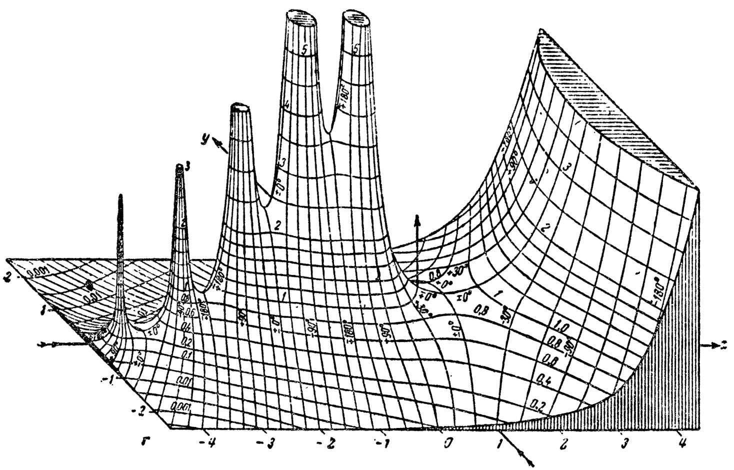
$$H(x) = \frac{1}{\Gamma(1-x)}\frac{d}{dx}\log\left[\Gamma\left(\frac{1-x}{2}\right)/\Gamma\left(1-\frac{x}{2}\right)\right]$$

Leonhard Euler (1707–1783)

The Gamma / Beta Functions

intractable integrals, for Laplace

[graph: E. Jahnke & F. Emde, *Tafeln höherer Funktionen, 4.ed., Leipzig 1948*; image: public domain]

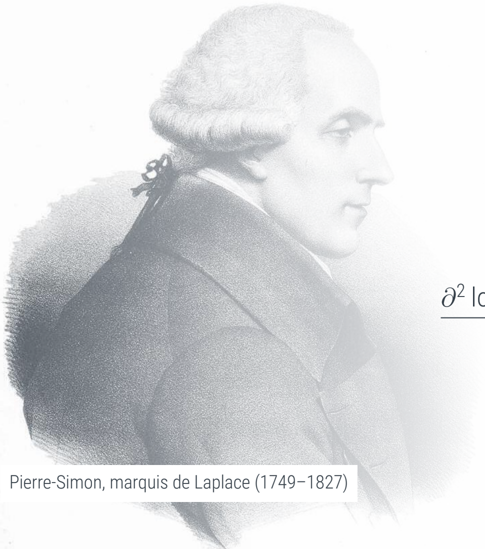


Jacques Salomon Hadamard (1865–1963)

Laplace’s Approximation

What you do when it’s 1814 and you don’t have a computer

P.-S. Laplace, 1814



Pierre-Simon, marquis de Laplace (1749–1827)

$$
\begin{aligned}
p(x \mid a, b) &= \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)} \\
\log p(x \mid a, b) &= (a-1)\log x + (b-1)\log(1-x) - \text{const.} \\
\frac{\partial \log p(x \mid a, b)}{\partial x} &= \frac{a-1}{x} - \frac{b-1}{1-x} \qquad \Rightarrow \hat{x} := \frac{a-1}{a+b-2} \\
\left.\frac{\partial^2 \log p(x \mid a, b)}{\partial x^2}\right|_{x=\hat{x}} &= -\frac{a-1}{\hat{x}^2} - \frac{b-1}{(1-\hat{x})^2} \\
&= -(a+b-2)^2 \left(\frac{1}{a-1} + \frac{1}{b-1}\right) =: \Psi \\
\log p(x) &\approx \log p(\hat{x}) + 0 + \frac{1}{2}(x-\hat{x})^2 \Psi \\
\int p(x)\, dx &\approx p(\hat{x}) \cdot \int \exp\left(-\frac{(x-\hat{x})^2}{2(-\Psi^{-1})}\right) dx = 1
\end{aligned}
$$

A little help from a friend

what a nice chap, I wonder whether we'll ever encounter him again …

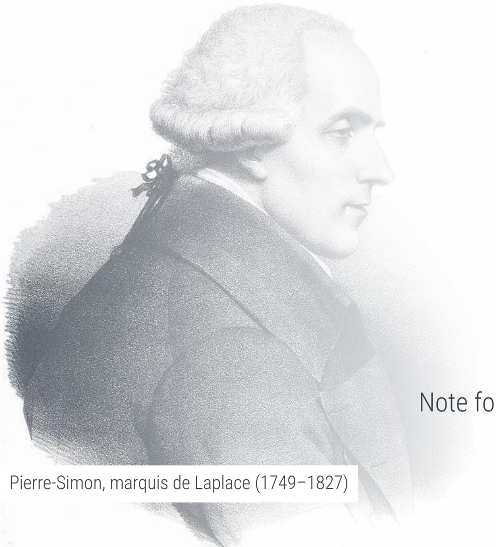
Carl Friedrich Gauss (1777–1855)

$$\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\frac{(x-m)^2}{v}\right)\,dx = \sqrt{2\pi v}$$

Laplace’s Approximation

What you do when it’s 1814 and you don’t have a computer

P.-S. Laplace, 1814



Pierre-Simon, marquis de Laplace (1749–1827)

$$p(x \mid a, b) = \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)} \qquad \hat{x} := \frac{a-1}{a+b-2}$$

$$\Psi := -(a+b-2)^2 \left(\frac{1}{a-1} + \frac{1}{b-1} \right)$$

$$\int p(x)\, dx \approx p(\hat{x}) \cdot \int \exp\left(-\frac{(x-\hat{x})^2}{2(-\Psi^{-1})} \right) dx = 1$$

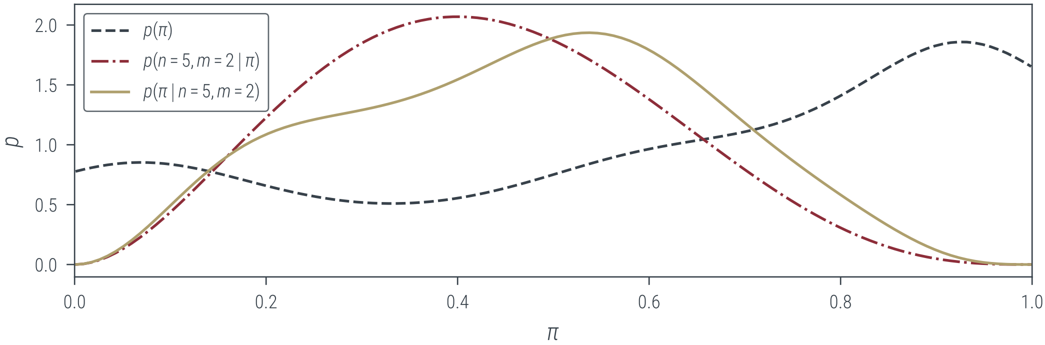
$$B(a,b) \approx \hat{x}^{a-1} \hat{x}^{b-1} \cdot \sqrt{2\pi(-\Psi^{-1})}$$

Note for $N = a + b \to \infty$ we have $\Psi \approx -\frac{(a+b)^3}{ab} = -\frac{N^2}{N-a} = -\frac{N^2}{N-b}$, so

$$\sqrt{-\Psi^{-1}} \to \frac{1}{N}\sqrt{N-b} = \frac{1}{N}\sqrt{N-b} \approx \frac{\sqrt{f}}{\sqrt{N}}$$

Choose your prior, if you’re willing to pay for it

inference is never “intractable” if you have enough compute



```
N = 120
xp = jnp.linspace(0, 1, N)
posterior = prior(xp) * likelihood(xp)
posterior = posterior / posterior.sum() * (N - 1)
```

- **Random Variables** allow us to define derived quantities from atomic events
- **Borel σ-algebras** can be defined on all topological spaces, allowing us to define probabilities if the elementary space is continuous.
- **Probability Density Functions (pdf's)** distribute probability across continuous domains.
  - they satisfy "the rules of probability" (integrate to one, sum rule, product rule, hence Bayes' Theorem)
  - Not every measure has a density, but all pdfs define measures
  - Densities transform under continuously transformations
- Probabilistic inference can even be used to infer probabilities!
- With the right prior, the posterior might be possible to compute simply by adding `floats`, if you know how to compute one special integral. If you don't, just use a **Laplace approximation**
- If you don't like that prior, *we* can use another one (Laplace couldn't!), because **we've got computers!**

Please cite this course, as

```
@techreport{Tuebingen_ProbML23,
    title =
    {Probabilistic Machine Learning},
    author = {Hennig, Philipp},
    series = {Lecture Notes
         in Machine Learning},
    year = {2023},
    institution = {Tübingen AI Center}}
```