# Posterior estimation with neural networks

BSc Machine learning course 2020

Christoph Weniger, GRAPPA

University of Amsterdam

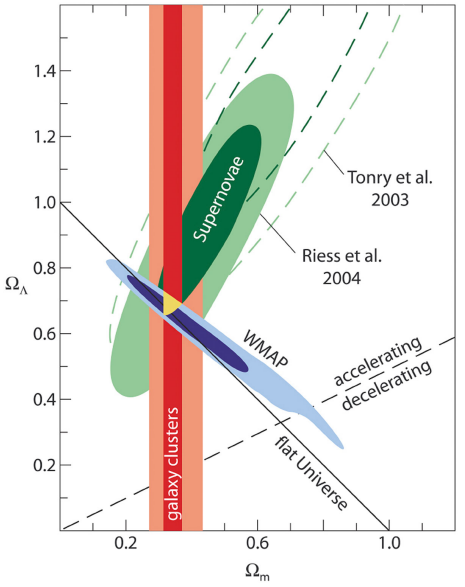
28 September 2020

# All measurements are uncertain

**Cosmological parameters**

- Matter content vs dark energy in the Universe
- Different measurements constrain the parameters in complementary ways

How can we describe these uncertainties mathematically?

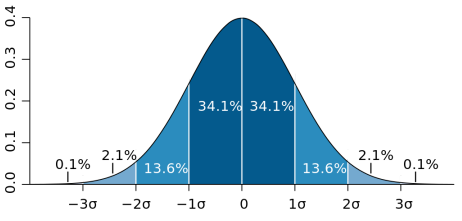



Credit: ESO

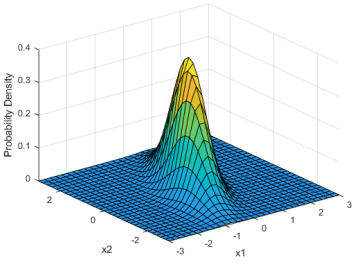
# Probability distributions

The distribution of continuous random numbers is described by probability density functions

1-dim standard normal distribution



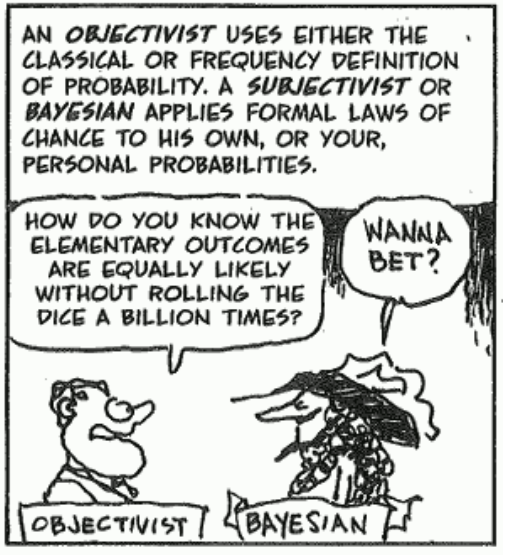

Example for a 2-dim distribution




Bayesian statistics: Those distributions can also describe our belief (plausibility/probability) that a certain parameter has a certain value.

Examples: Mass of the higgs particle, age of the Universe, average temperature of Earth atmosphere in 1900 (reconstruction) or in 2100 (extrapolation).

# Two schools: Bayesian* vs Frequentist




*what we do here

Source: The Cartoon Guide to Statistics

# Bayes' theorem

Bayes' theorem provides a clear rule for how to update believes with new data

**Likelihood** $P(D|H)$

How probable is the data $D$ given that our hypothesis $H$ is true?

**Prior** $P(H)$

How probable was our hypothesis $H$) before observing the evidence?

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

# Bayes' theorem

Bayes' theorem provides a clear rule for how to update believes with new data

**Likelihood** $P(D|H)$

How probable is the data $D$ given that our hypothesis $H$ is true?

**Prior** $P(H)$

How probable was our hypothesis $H$) before observing the evidence?

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

**Posterior** $P(D|H)$

How probably is our hypothesis $H$ given the observed data $D$?

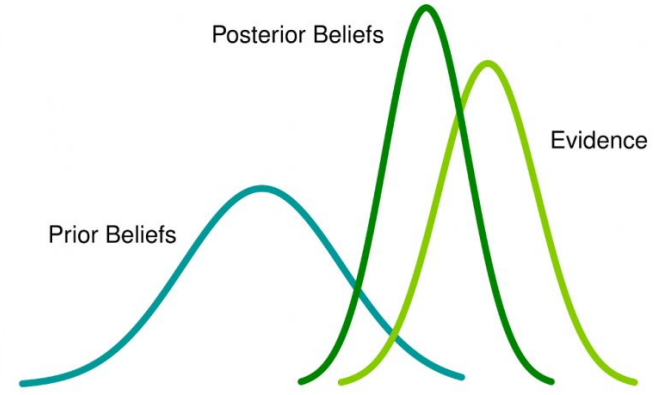
**Marginal likelihood** $P(D)$

How probable is the new data $D$ under all possible hypothesis $H$?

$$P(D) \equiv \sum_H P(D|H)P(H)$$

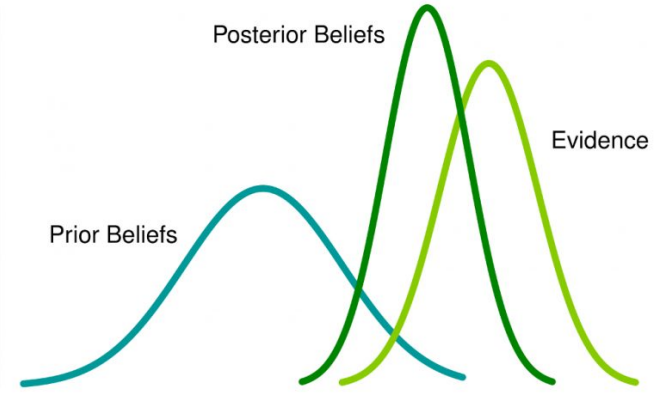
# Update rule in practice

- Prior Beliefs: $P(H)$
- Evidence/likelihood: $P(D|H)$
- Posterior beliefs: $P(H|D)$



# Update rule in practice

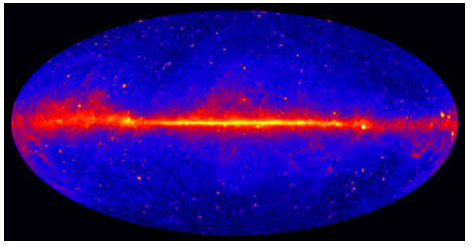
- Prior Beliefs: $P(H)$
- Evidence/likelihood: $P(D|H)$
- Posterior beliefs: $P(H|D)$



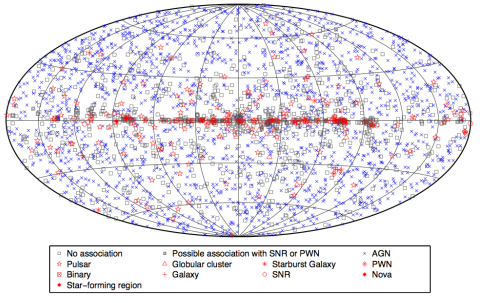

However, in many cases the likelihood of observing $D$ given hypothesis $H$, $P(D|H)$ is not actually known, or very hard to calculate.

# Physics examples - gamma rays

Gamma-ray sky

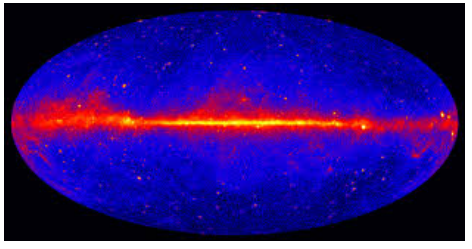


Observed gamma-ray sources

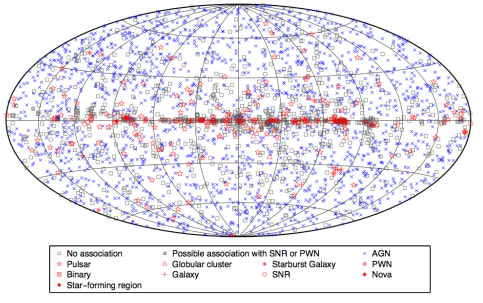


# Physics examples - gamma rays

Gamma-ray sky



Observed gamma-ray sources

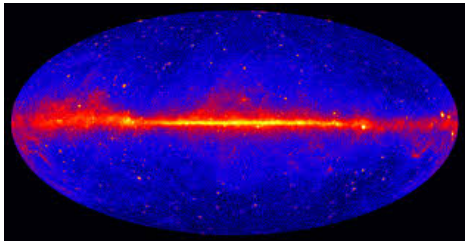



**Difficult to determine properties of source populations? E.g.:**

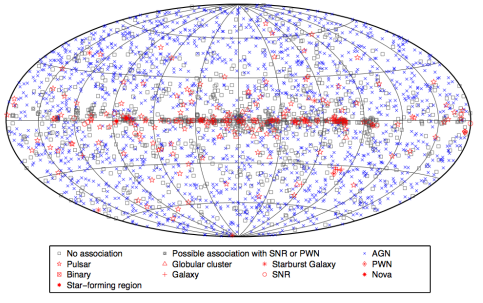
- What is the spatial distribution of pulsars in the Galactic disk?

# Physics examples - gamma rays

Gamma-ray sky



Observed gamma-ray sources

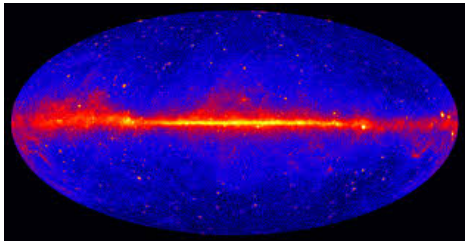



**Difficult to determine properties of source populations? E.g.:**

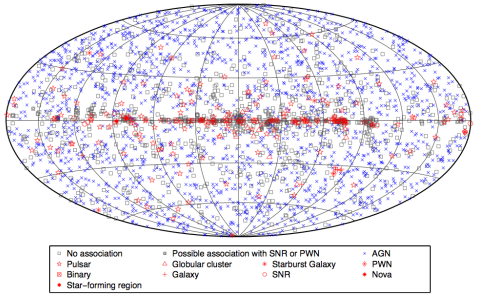
- What is the spatial distribution of pulsars in the Galactic disk?
- What is the luminosity function of blazars at high Galactic latitudes?

# Physics examples - gamma rays

Gamma-ray sky



Observed gamma-ray sources

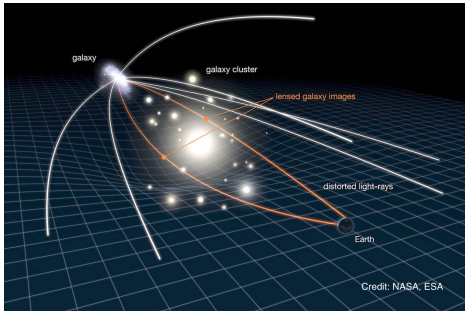



**Difficult to determine properties of source populations? E.g.:**

- What is the spatial distribution of pulsars in the Galactic disk?
- What is the luminosity function of blazars at high Galactic latitudes?
- Are gaps in the source map physical or due to too large noise in the image?

# Physics examples - strong lensing

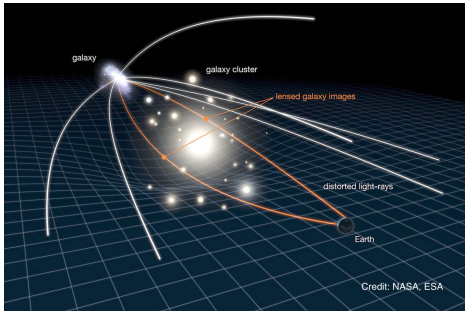
Strong gravitational lensing due to dark and visible matter




Example: Distant galaxy lensed by red lens galaxy along the line-of-sight

# Physics examples - strong lensing

Strong gravitational lensing due to dark and visible matter



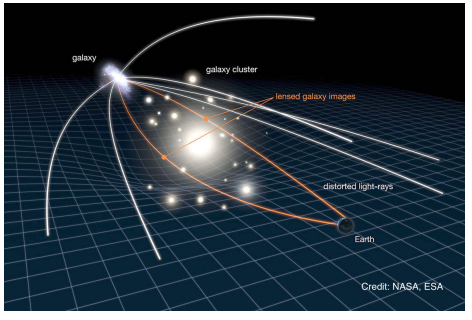

Example: Distant galaxy lensed by red lens galaxy along the line-of-sight

## Questions

- How does the unperturbed source look like and how the lens?

# Physics examples - strong lensing

Strong gravitational lensing due to dark and visible matter




Example: Distant galaxy lensed by red lens galaxy along the line-of-sight

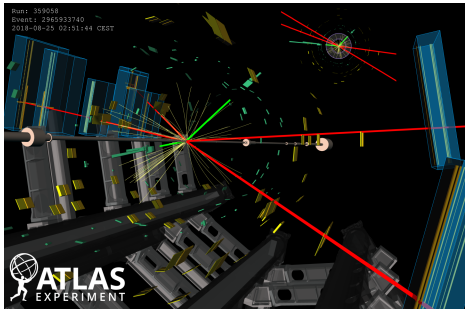
## Questions

- How does the unperturbed source look like and how the lens?
- What can we learn about the nature of dark matter? Small clumps of dark matter would lead to characteristic distortions in the image.

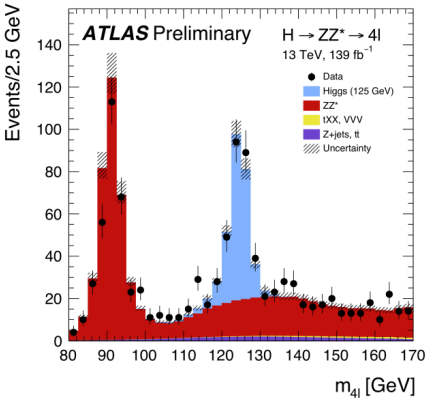
# Observations are often the autcome of a LARGE number of random processes

# Physics examples - collider physics

Illustration of collition at ALTAS detector

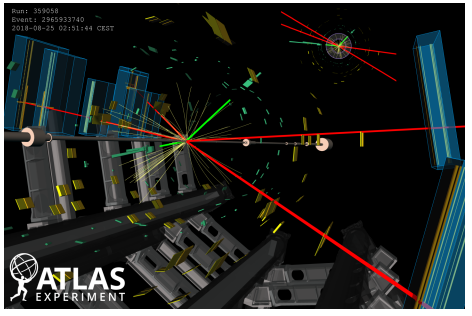



Invariant mass of 4-lepton channel

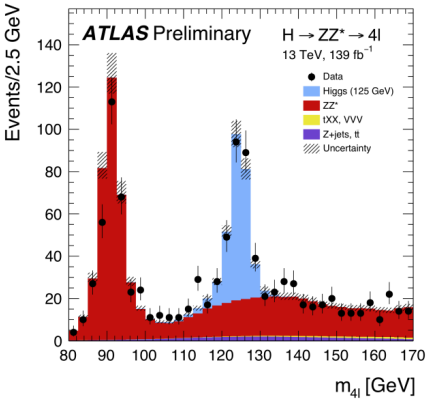


# Physics examples - collider physics

Illustration of collition at ALTAS detector




Invariant mass of 4-lepton channel

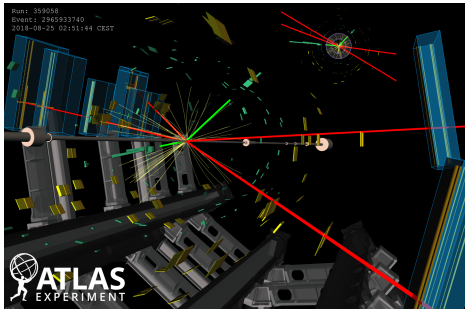



**Typical questions are:**

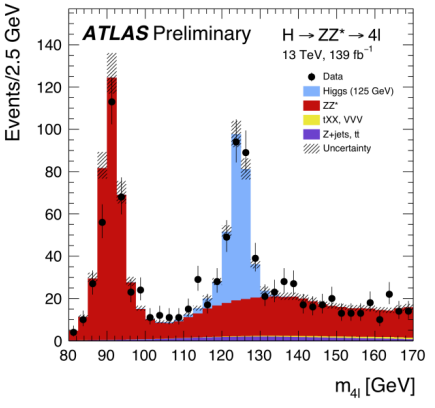
- What processes have most likely contributed to the 4-lepton signal?

# Physics examples - collider physics

Illustration of collition at ALTAS detector




Invariant mass of 4-lepton channel




**Typical questions are:**

- What processes have most likely contributed to the 4-lepton signal?
- How does this constraint the Higgs production cross section?

# The likelihood-to-evidence ratio

In order to evaluate the probability of any outcome, we often have to sum or integrate over a very large number of random variables, here called $z$, that we are not actually interested in.

$$P(D|H) = \underbrace{\int dz}_{\text{intractable}} \; P(D|H,z)P(z|H)$$

# The likelihood-to-evidence ratio

In order to evaluate the probability of any outcome, we often have to sum or integrate over a very large number of random variables, here called $z$, that we are not actually interested in.

$$P(D|H) = \underbrace{\int dz}_{\text{intractable}} \; P(D|H,z)P(z|H)$$

**Examples**

- Position of individual point sources on the sky

# The likelihood-to-evidence ratio

In order to evaluate the probability of any outcome, we often have to sum or integrate over a very large number of random variables, here called $z$, that we are not actually interested in.

$$P(D|H) = \underbrace{\int dz}_{\text{intractable}} \; P(D|H,z)P(z|H)$$

**Examples**

- Position of individual point sources on the sky
- Position of individual dark matter halos, individual galaxies images

# The likelihood-to-evidence ratio

In order to evaluate the probability of any outcome, we often have to sum or integrate over a very large number of random variables, here called $z$, that we are not actually interested in.

$$P(D|H) = \underbrace{\int dz}_{\text{intractable}} \; P(D|H,z)P(z|H)$$

**Examples**

- Position of individual point sources on the sky
- Position of individual dark matter halos, individual galaxies images
- Physical mechanisms that lead to specific 4-lepton invariant mass

# The likelihood-to-evidence ratio

In order to evaluate the probability of any outcome, we often have to sum or integrate over a very large number of random variables, here called $z$, that we are not actually interested in.

$$P(D|H) = \underbrace{\int dz}_{\text{intractable}} \; P(D|H,z)P(z|H)$$

**Examples**

- Position of individual point sources on the sky
- Position of individual dark matter halos, individual galaxies images
- Physical mechanisms that lead to specific 4-lepton invariant mass

There are many ways to solve the integral approximately. Here, we will use neural networks to help us out.

# Neural likelihood-free inference

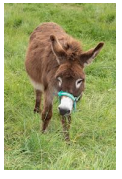
**Starting point**: for any pair of observation $x$ and model parameter $\theta$, the goal is to estimate the probability that this pair belongs one of the following classes:

| | |
|---|---|
| $H_0$: Data $x$ comes from model $\theta$ | $(x, \theta) \sim P(x, \theta) = P(x\|\theta)P(\theta)$ |
| $H_1$: Data $x$ and model $\theta$ are unrelated | $(x, \theta) \sim P(x)P(\theta)$ |

# Joint vs marginal samples

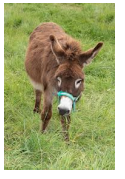
Examples for $H_0$, jointly sampled from $D, H \sim P(D|H)P(H)$

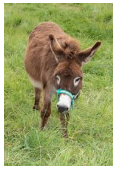
Cat 

Donkey 

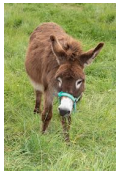
Cat 

Cat 

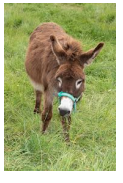
Donkey 

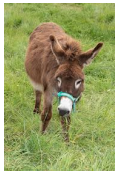
Donkey 

Examples for $H_1$, marginally sampled from $D, H \sim P(D)P(H)$

Cat 

Cat 

Cat 

Donkey 

Donkey 

Cat 

# Loss function

**Strategy:** We train a neural network $d_\phi(x, \theta)$ as binary classifier to estimate the probability of hypothesis $H_0$ ($y = 0$) or $H_1$ ($y = 1$).

The corresponding loss function (see logistic regression example) is

$$L\left[d(x,\theta)\right] = \int dx d\theta \left[p(x,\theta) \ln\left(d(x,\theta)\right) + p(x)p(\theta) \ln\left(1 - d(x,\theta)\right)\right]$$

# Loss function

**Strategy:** We train a neural network $d_\phi(x, \theta)$ as binary classifier to estimate the probability of hypothesis $H_0$ ($y = 0$) or $H_1$ ($y = 1$).

The corresponding loss function (see logistic regression example) is

$$L\left[d(x,\theta)\right] = \int dx d\theta \left[p(x,\theta)\ln\left(d(x,\theta)\right) + p(x)p(\theta)\ln\left(1 - d(x,\theta)\right)\right]$$

Minimizing that function (blackboard!) yields

$$d(x,\theta) = \frac{P(x,\theta)}{P(x,\theta) + P(x)P(\theta)}$$

# Loss function

**Strategy:** We train a neural network $d_\phi(x, \theta)$ as binary classifier to estimate the probability of hypothesis $H_0$ ($y = 0$) or $H_1$ ($y = 1$).

The corresponding loss function (see logistic regression example) is

$$L\left[d(x,\theta)\right] = \int dx d\theta \left[p(x,\theta)\ln\left(d(x,\theta)\right) + p(x)p(\theta)\ln\left(1 - d(x,\theta)\right)\right]$$

Minimizing that function (blackboard!) yields

$$d(x,\theta) = \frac{P(x,\theta)}{P(x,\theta) + P(x)P(\theta)}$$

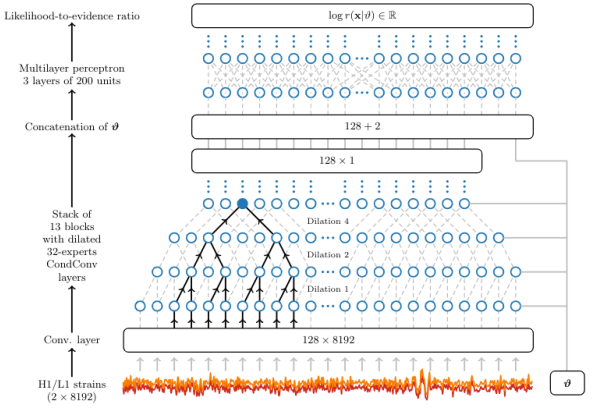
Since

$$r(x,\theta) \equiv \frac{P(x|\theta)}{P(x)} = \frac{P(\theta|x)}{P(\theta)} = \frac{1}{d(x,\theta)} - 1.$$

we are actually performing ratio estimation.

# Goal is to train something like this

Read world example for the analysis of gravitational wave signals




- Goal is to train NN as function: $d(D, H) \in (0, 1)$
- $d(D, H)$ is the probability that $D$ and $H$ are jointly drawn.

# Exercise: Neural posterior estimation

Your task is to take the solution to exercise 3 and replace point estimation with posterior estimation. To this end you have to

- Adopt the convolutional neural network such that it takes an additional input, the radius $r$, and such that the output is between zero and one. A good way to do that is to replace the last layer with

  $$\texttt{(INPUT, r)} \to \texttt{FC} \to \texttt{RELU} \to \texttt{FC} \to \texttt{SIGMOID} \to \texttt{OUTPUT}$$

  Here, `INPUT` is the output of the second-to-last `FC`, which is concatenated with the input variable $r$.
- Replace the loss function by the above binary cross-entropy loss function $L[d(x, \theta)]$.

Once this is done

- Train the network and show that the posterior (which is $\propto \exp(-r(x, \theta))$), peaks usually close to the true value
- Explore how the posterior becomes broaders when making the signal model more complicated (e.g. by masking random regions).