Bayesian posterior estimation with classification networks

Christoph Weniger, GRAPPA

University of Amsterdam

28 September 2020

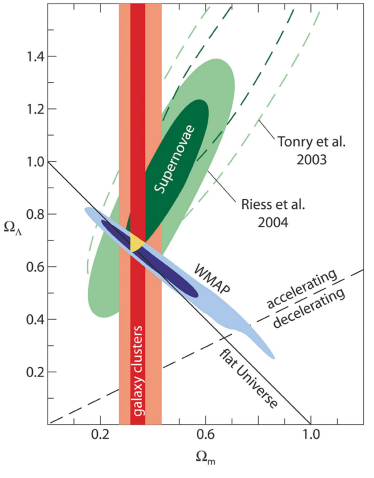
BSc Course Machine learning for Physics and Astronomy, 2020

All measurements are uncertain

Causes for uncertainties

- Random fluctuations
  - photon shot noise, ...
- Detector systematics
  - calibration uncertainties, ...
- Uncertainties in theoretical models
  - cosmological constant, ...
- Imperfect/simplistic models
  - modeling complex astrophysical data
- Limitations of the analysis framework
  - assumption about well-behaved noise

Example: Cosmological parameters



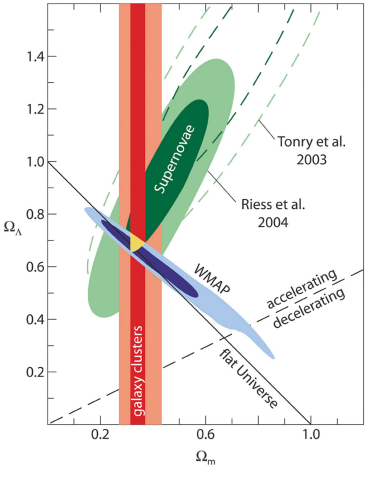
Credit: ESO

All measurements are uncertain

Causes for uncertainties

- Random fluctuations
  - photon shot noise, ...
- Detector systematics
  - calibration uncertainties, ...
- Uncertainties in theoretical models
  - cosmological constant, ...
- Imperfect/simplistic models
  - modeling complex astrophysical data
- Limitations of the analysis framework
  - assumption about well-behaved noise

Example: Cosmological parameters



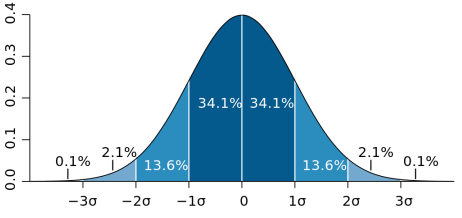
How can we describe these uncertainties mathematically?

Credit: ESO

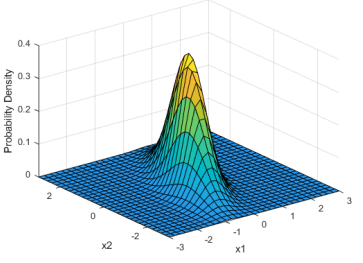
Probability distributions

Continuous random numbers are described by probability density functions

1-dim standard normal distribution



$$\int_{-\infty}^{\infty} dx\, P(x) = 1$$



$$\int_{\mathbb{R}^2} dx_1 dx_2\, P(x_1, x_2) = 1$$

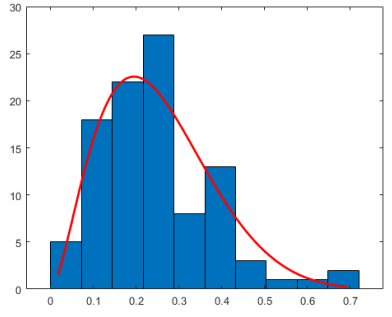
Any non-negative function normalized to one can act as probability distribution.

Sampling from a probability distribution

- Random draws x from $P(x)$ can be written as $x \sim P(x)$

Sampling from a probability distribution

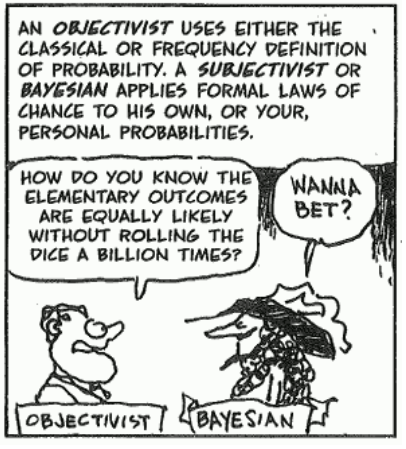
- Random draws x from $P(x)$ can be written as $x \sim P(x)$
- Making a histogram of many draws recovers the shape of $P(x)$



red line: $\propto P(x)$

blue bars: histogram of many draws $x \sim P(x)$

Frequentist vs Bayesian* interpretation



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

Source: The Cartoon Guide to Statistics

Conditional probabilities

1) Probability of x and y:

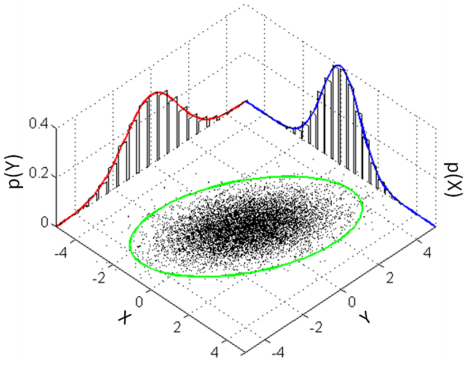
$$P(x, y)$$

2) Probability of y, obtained by "marginalizing" x:

$$P(y) = \int_{-\infty}^{\infty} dx\, P(x, y)$$

3) Conditional probability of x given y:

$$P(x|y) = \frac{P(x, y)}{P(y)}$$



Conditional probabilities

1) Probability of x and y:

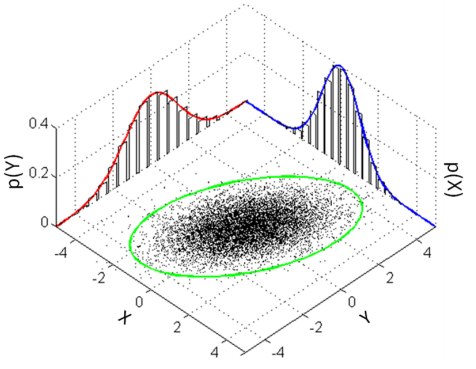
$$P(x, y)$$

2) Probability of y, obtained by "marginalizing" x:

$$P(y) = \int_{-\infty}^{\infty} dx\, P(x, y)$$

3) Conditional probability of x given y:

$$P(x|y) = \frac{P(x, y)}{P(y)}$$



Per definition

$$P(x|y)P(y) = P(y|x)P(x)$$

Updating beliefs with Bayes' theorem

Bayes' theorem provides a rule for updating believes in light of 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?

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

The Galton board / Bean machine

An example for random sampling from the normal distribution

Updating beliefs with Bayes' theorem

Bayes' theorem provides a rule for updating believes in light of 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?

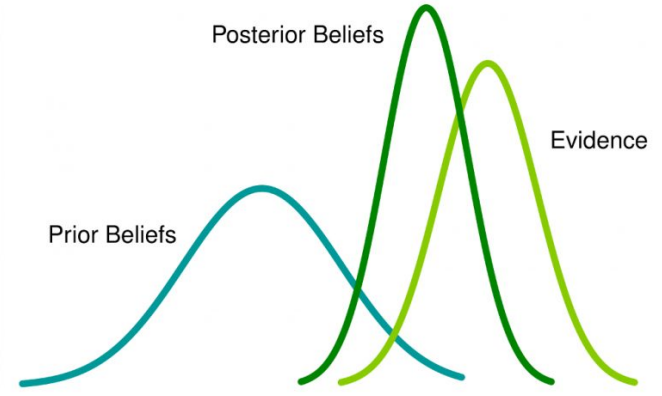
Marginal likelihood $P(D)$

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

$$P(D) \equiv \int dH \, P(D|H)P(H)$$

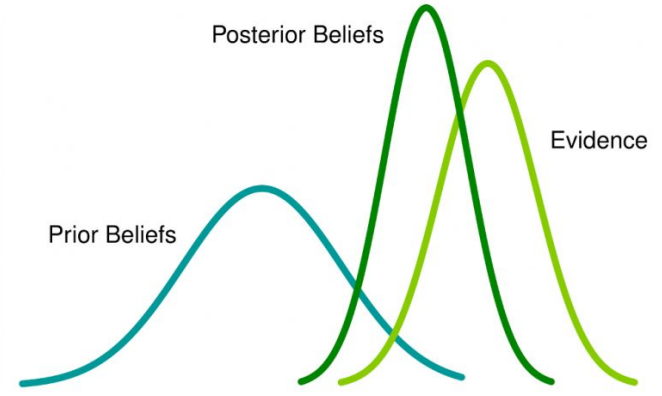
Update rule in practice

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



Update rule in practice

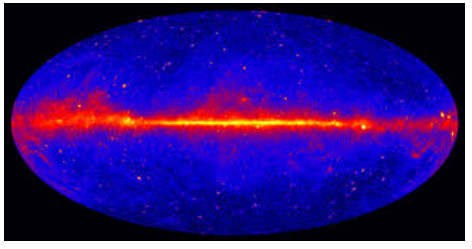
- Prior Beliefs: $P(H)$
- Evidence/likelihood: $P(D|H)$
- Posterior beliefs: $P(H|D)$



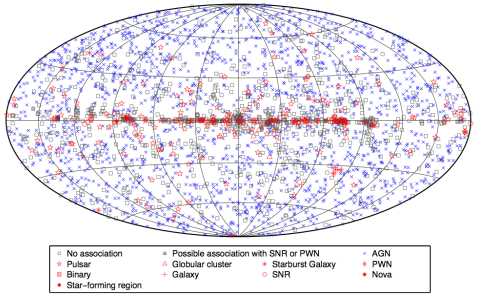
However, in many cases the likelihood of observing D given hypothesis H, $P(D|H)$ is not actually known, or very hard to calculate.

Physics examples - gamma rays

Gamma-ray sky

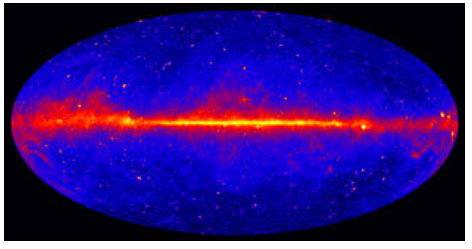


Observed gamma-ray sources

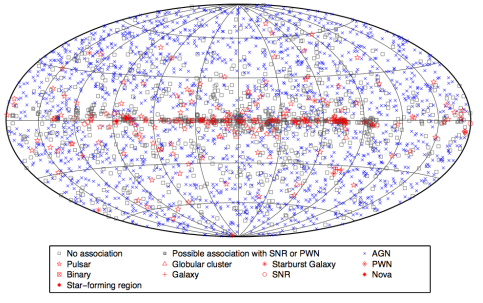


Physics examples - gamma rays

Gamma-ray sky



Observed gamma-ray sources

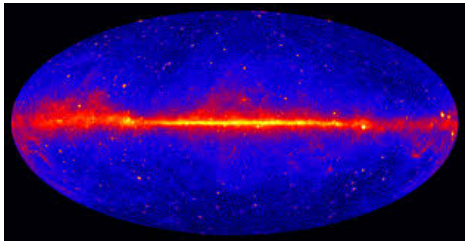


Difficult to determine properties of source populations? E.g.:

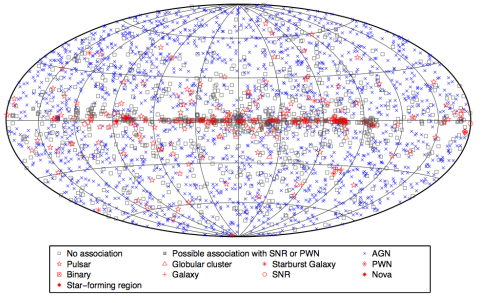
- What is the spatial distribution of pulsars in the Galactic disk?

Physics examples - gamma rays

Gamma-ray sky



Observed gamma-ray sources

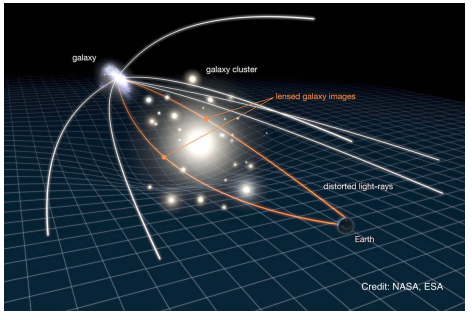


Difficult to determine properties of source populations? E.g.:

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

Physics examples - strong lensing

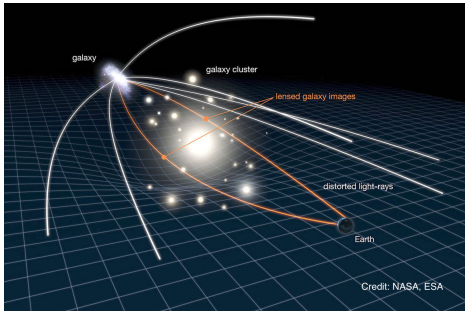
Strong gravitational lensing due to dark and visible matter



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



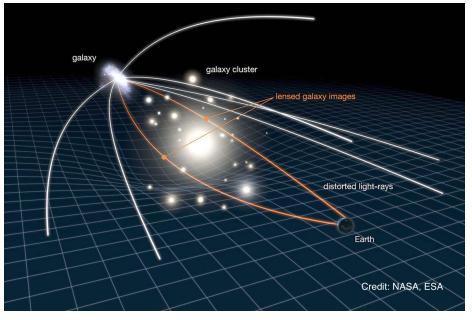
Example: Distant galaxy lensed by red lens galaxy along the line-of-sight

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



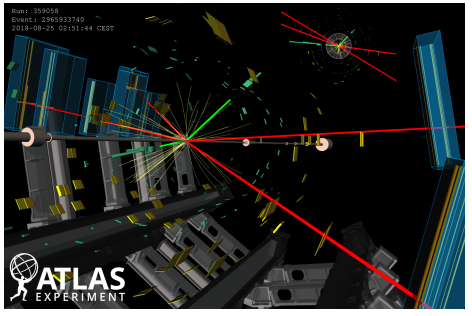
Example: Distant galaxy lensed by red lens galaxy along the line-of-sight

Questions

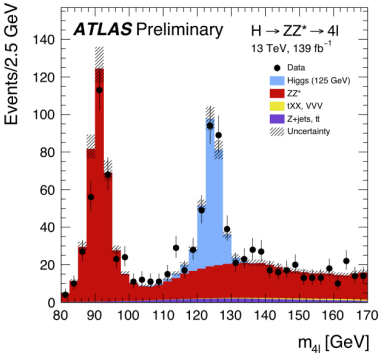
- How does the unperturbed source look like and how the lens?
- What can we learn about the nature of dark matter? Small clumps of dark matter would lead to characteristic distortions in the image.

Physics examples - collider physics

Illustration of collition at ALTAS detector

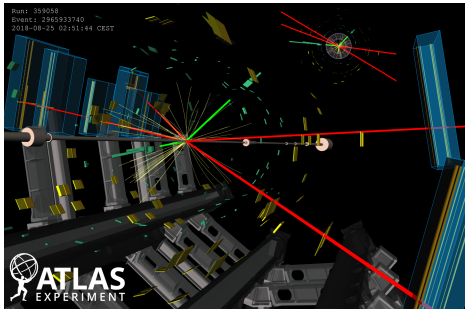


Invariant mass of 4-lepton channel

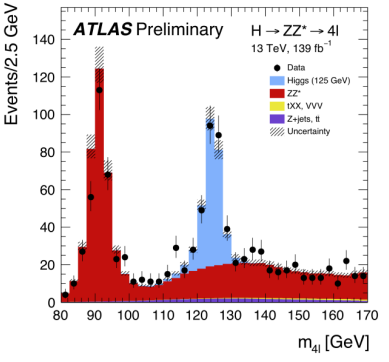


Physics examples - collider physics

Illustration of collition at ALTAS detector



Invariant mass of 4-lepton channel

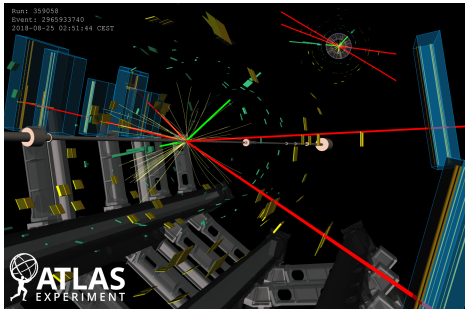


Typical questions are:

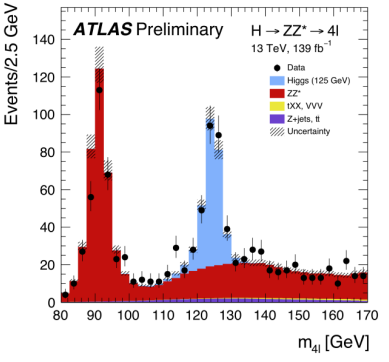
- What processes have most likely contributed to the 4-lepton signal?

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?

Neural likelihood-free inference

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

H_0: Data $\vec{x}$ corresponds to model parameters $\vec{z}$

H_1: Data $\vec{x}$ and model $\vec{z}$ are unrelated

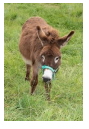
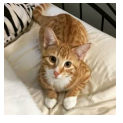
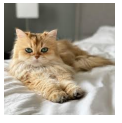
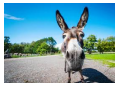
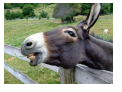
$$(\vec{x}, \vec{z}) \sim P(\vec{x}, \vec{z}) = P(\vec{x}|\vec{z})P(\vec{z})$$

$$(\vec{x}, \vec{z}) \sim P(\vec{x})P(\vec{z})$$

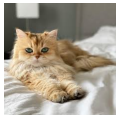
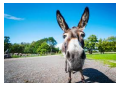
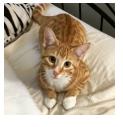
See Louppe+Hermanns 2019

Joint vs marginal samples

1) Examples for H_0, jointly sampled from $\vec{x}, z \sim P(\vec{x}|z)P(z)$

Cat  Donkey  Cat  Cat  Donkey  Donkey 

2) Examples for H_1, marginally sampled from $\vec{x}, z \sim P(\vec{x})P(z)$

Donkey  Cat  Cat  Cat  Donkey  Donkey 

Data: $\vec{x} = \text{Image}$; Label: $z \in \text{Cat}, \text{Donkey}$

Loss function

Strategy: We train a neural network $d_\phi(\vec{x}, z) \in [0, 1]$ as binary classifier to estimate the probability of hypothesis H_0 or H_1. The network output can be interpreted, for a given input pair $\vec{x}$ and z, as probability that H_0 is true.

- H0 is true: $d_\phi(\vec{x}, z) \simeq 1$
- H1 is true: $d_\phi(\vec{x}, z) \simeq 0$

The corresponding loss function is (so-called "binary cross-entroy")

$$L\left[d(\vec{x}, z)\right] = \int dx dz \left[p(\vec{x}, z) \ln\left(d(\vec{x}, z)\right) + p(\vec{x}) p(z) \ln\left(1 - d(\vec{x}, z)\right)\right]$$

Loss function

Strategy: We train a neural network $d_\phi(\vec{x}, z) \in [0, 1]$ as binary classifier to estimate the probability of hypothesis H_0 or H_1. The network output can be interpreted, for a given input pair $\vec{x}$ and z, as probability that H_0 is true.

- H0 is true: $d_\phi(\vec{x}, z) \simeq 1$
- H1 is true: $d_\phi(\vec{x}, z) \simeq 0$

The corresponding loss function is (so-called "binary cross-entroy")

$$L\left[d(\vec{x}, z)\right] = \int dx dz \left[p(\vec{x}, z) \ln\left(d(\vec{x}, z)\right) + p(\vec{x})p(z) \ln\left(1 - d(\vec{x}, z)\right)\right]$$

Minimizing that function (see next slide) w.r.t. the network parameters ϕ yields

$$d(\vec{x}, z) \approx \frac{p(\vec{x}, z)}{p(\vec{x}, z) + p(\vec{x})p(z)}$$

Properties of optimized network

Some analytical estimates

$$\frac{\partial}{\partial \phi} L\left[d_\phi(\vec{x}, z)\right] = \int dx dz \left[p(\vec{x}, z) \ln\left(d(\vec{x}, z)\right) + p(\vec{x}) p(z) \ln\left(1 - d(\vec{x}, z)\right)\right]$$

$$= \int dx dz \left[\frac{p(\vec{x}, z)}{d(\vec{x}, z)} + \frac{p(\vec{x}) p(z)}{1 - d(\vec{x}, z)}\right] \frac{\partial d(\vec{x}, z)}{\partial \phi}$$

Setting the part in square brackets to zero yields that the network is optimized once

$$d(\vec{x}, z) \approx \frac{p(\vec{x}, z)}{p(\vec{x}, z) + p(\vec{x}) p(z)}$$

Likelihood-to-evidence ratio

Training binary classification networks yield true Bayesian posterior estimates!

With a bit of math one can show that

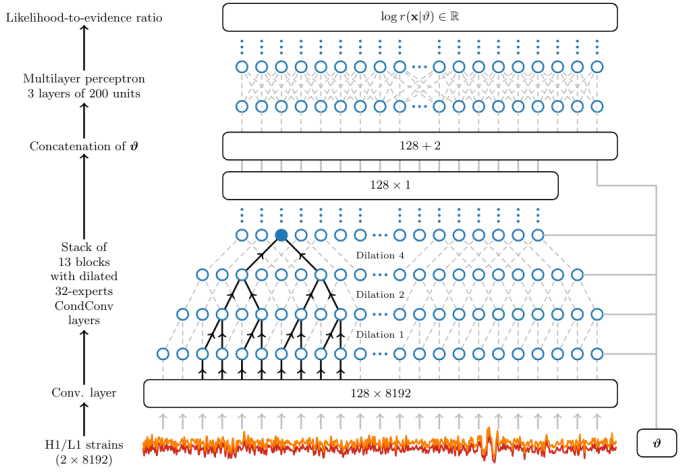
$$r(\vec{x}, z) \equiv \frac{1}{d(\vec{x}, z)} - 1 \simeq \frac{p(\vec{x}|z)}{p(\vec{x})} = \frac{p(z|\vec{x})}{p(z)}$$

Once we have trained the network $d_\phi(\vec{x}, z)$, we can estimate the posterior

$$p(z|\vec{x}) \simeq r(\vec{x}, z)p(z)$$

Example network

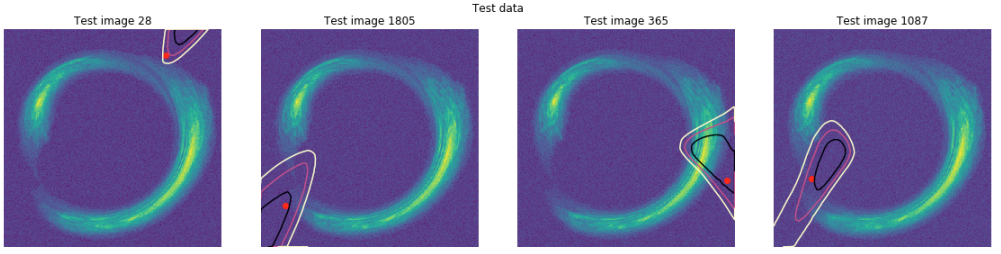
An example network implementing $d_\phi(\vec{x}, z)$



(our $\vec{x}$ corresponds to the noisy 1-dim data; our z corresponds to ϑ here)

Example: Strong lensing image analysis

Here we trained a neural network to recognizing subhalos in an image (indicated by a red dot). The contours show the most likely regions guessed by the network.



The shown posteriors are effectively marginalized over thousands of source and lens parameters. Those marginal posteriors can be evaluated in seconds once the network is trained (and training takes maybe 20 min).

Exercise: Neural posterior estimation

The overall goal of the exercise is to perform posterior estimation with classification networks. This is broken down in several tasks.

1. Training of a parameter regression network
   - Point estmation of ring radii in an image.
2. Training of a classification network
   - Train network to predict the probability of an image containing different number of rings.
3. Training of a posterior estimation network
   - Posterior estimation of ring radii in an image.

Enjoy!