Introduction to Bayesian statistics: exercises

Exercise 1

- Implement your own version of the line-fitting procedure, using the same data.
- Now try it with the data in `lectures/data/linear_fits/data_1.txt`
  - First plot the data. What has changed?
  - Try the same model and likelihood on the new data. You might want to adjust the prior on m and b for this new data set.
  - What if you use the provided uncertainty per data point σ_{y_i}, instead of assuming a constant variance σ_y for all data points?
  - Instead use the actual likelihood of the data:
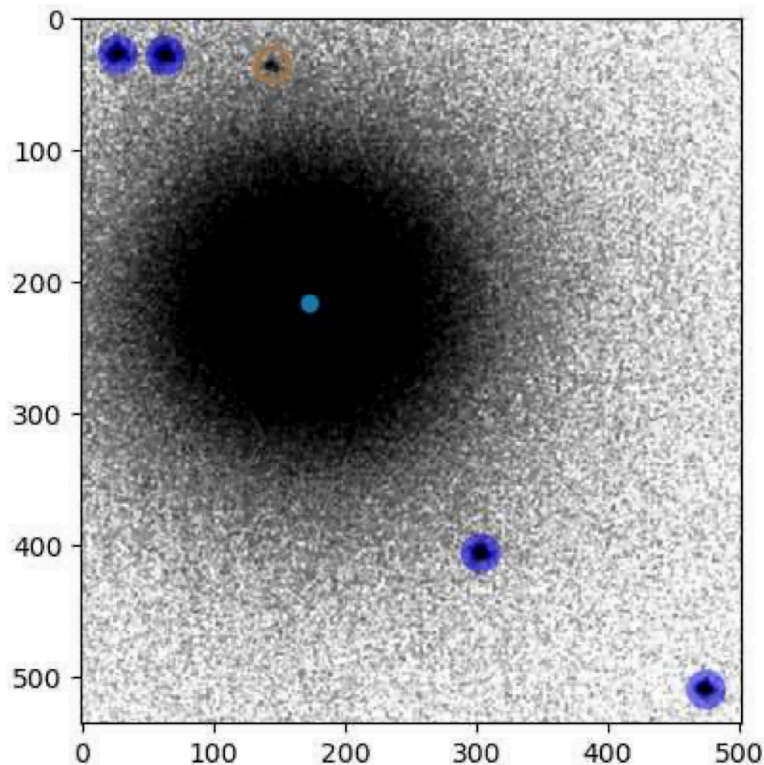
$$\mu(x) = mx + b \tag{1}$$
$$\sigma(x_i) = \sigma_{y_i} + f\mu(x_i)^2 \quad f > 0 \text{ a parameter} \tag{2}$$
$$y_i \sim \mathcal{N}(\mu(x_i), \sigma(x_i)^2) \tag{3}$$

  Careful with the normalisation of the Gaussian likelihood. Because we vary the variance, this matters now!

Exercise 2

The data in the previous example were a synthetic version of real observations that were obtained in an AstroWoche project in 2024 by Jonas Spiller, Theo Lequy, Clara Bleich.

The real observations can be found in `data/jovian_moons/{moon}.dat` :

```
moon = "Io"

t, distance, distance_alt, distance_err = np.loadtxt(
    f"./data/jovian_moons/{moon}.dat", unpack=True)
```

1. Repeat the analysis with the real data set. What do you find?

There is another distance estimate provided. The difference between the distance estimates comes from different approaches to converting the measured distance between Jupiter and its moons from the number of pixels to a physical distance in km.

2. How does the analysis look when using this other distance estimate (what I call `distance_alt`)? What about the other moons?

Because there are two distance estimates that do not agree, both of which are based on reasonable approaches, this suggests that there might be a *systematic* error in the measurements. Systematic errors are usually not random (which means they do not get smaller as you obtain more data). But we could pretend the offset between the two measurements is an estimate of an unaccounted statistical uncertainty.

3. How does the analysis look when we assume `sigma_d**2 = distance_err**2 + (distance - distance_alt)**2` ?

4. Instead of assuming that the uncertainty is underestimated by this offset, we can also model it. Does the model work better if the variance is a free parameter?