Chapter 5, Part 2: Bayesian Modelling

Advanced Topics in Statistical Machine Learning

Tom Rainforth

Hilary 2024

rainforth@stats.ox.ac.uk

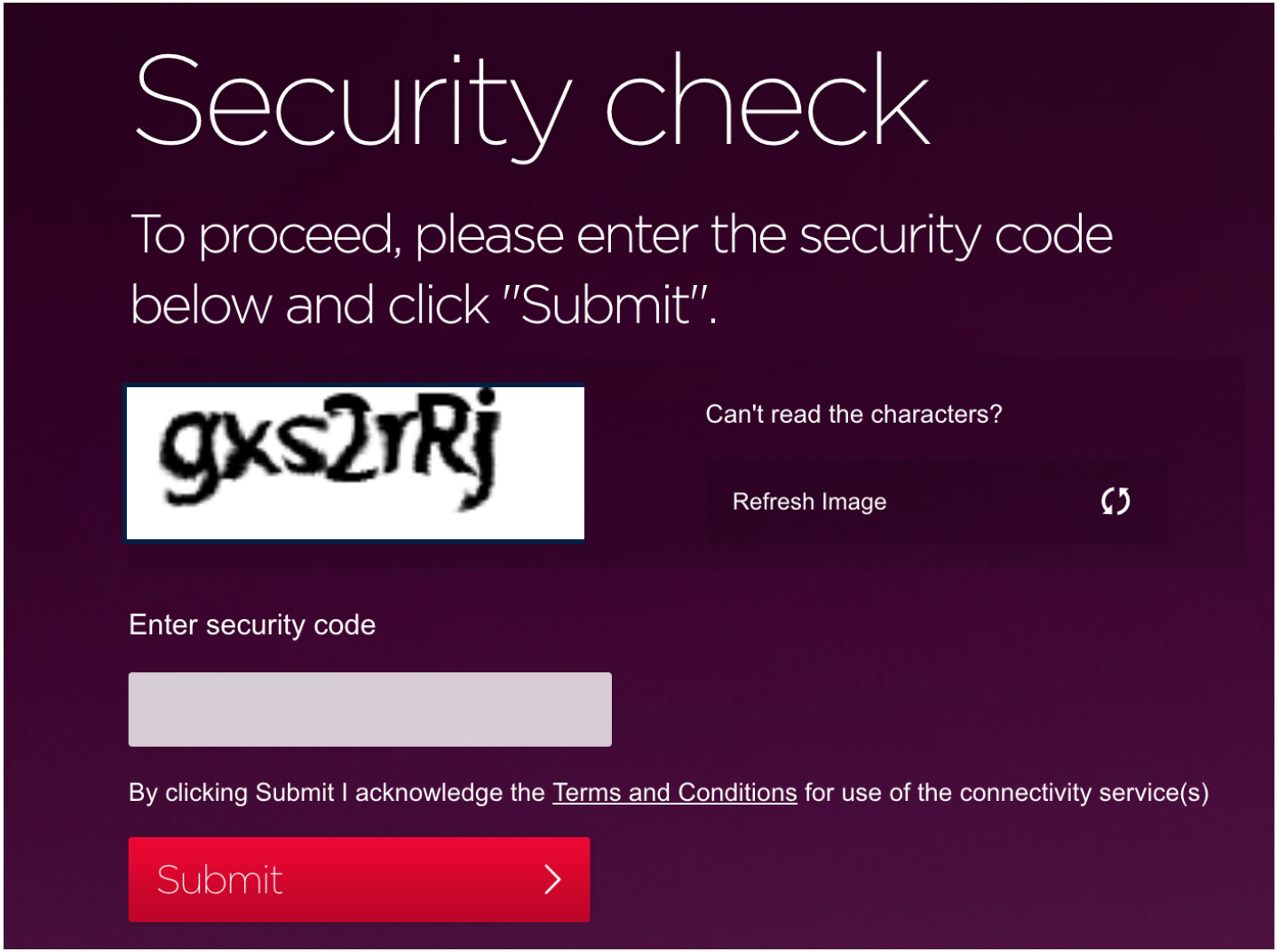
What is a Model?

- Models are mechanisms for **reasoning** about the world
- Examples: Newtonian mechanics, simulators, internal models our brain constructs
- Good models balance **fidelity**, **predictive power** and **tractability**
    - Quantum mechanics is a more accurate model than Newtonian mechanics, but it is actually less useful for everyday tasks

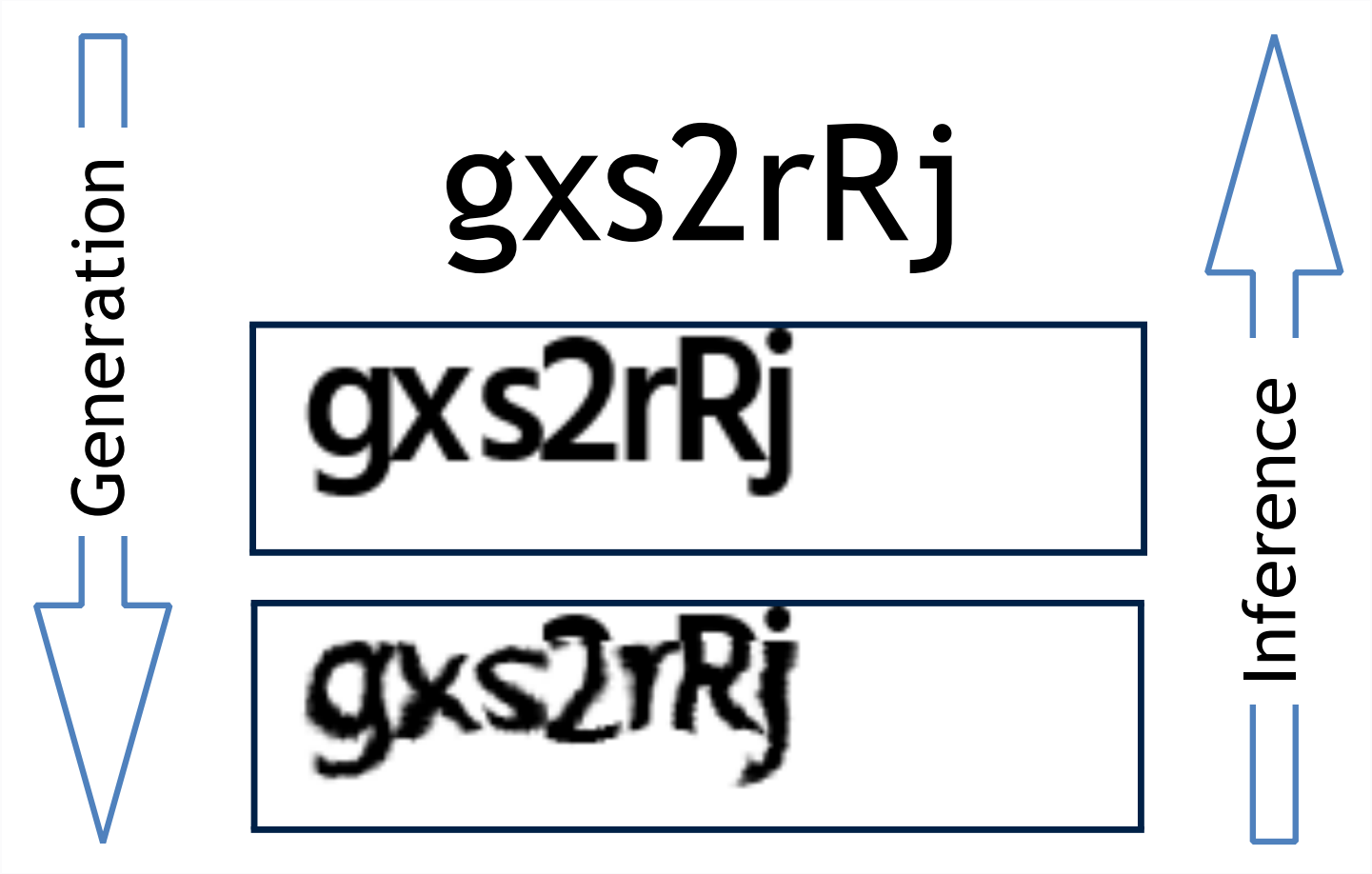
What is Bayesian Model?

- A Bayesian model is a **probabilistic generative model** $p(\theta, \mathcal{D})$ over **latents** θ and **data** $\mathcal{D}$
- It forms a probabilistic "simulator" for generating data that we **might** have seen
    - It is an approximation of the unknown true data generating process
- Pretty much any stochastic simulator can be used as a Bayesian model

Example: How Might we Write a System to Break Captchas?



Simulating Captchas is Much Easier



Breaking Captchas with Bayesian Models

https://youtu.be/ZTKx4TaqNrQ?t=9

[1] TA Le, A G Baydin, and F Wood. "Inference Compilation and Universal Probabilistic Programming". In: **AISTATS**. 2017.

All models are wrong, but some are useful

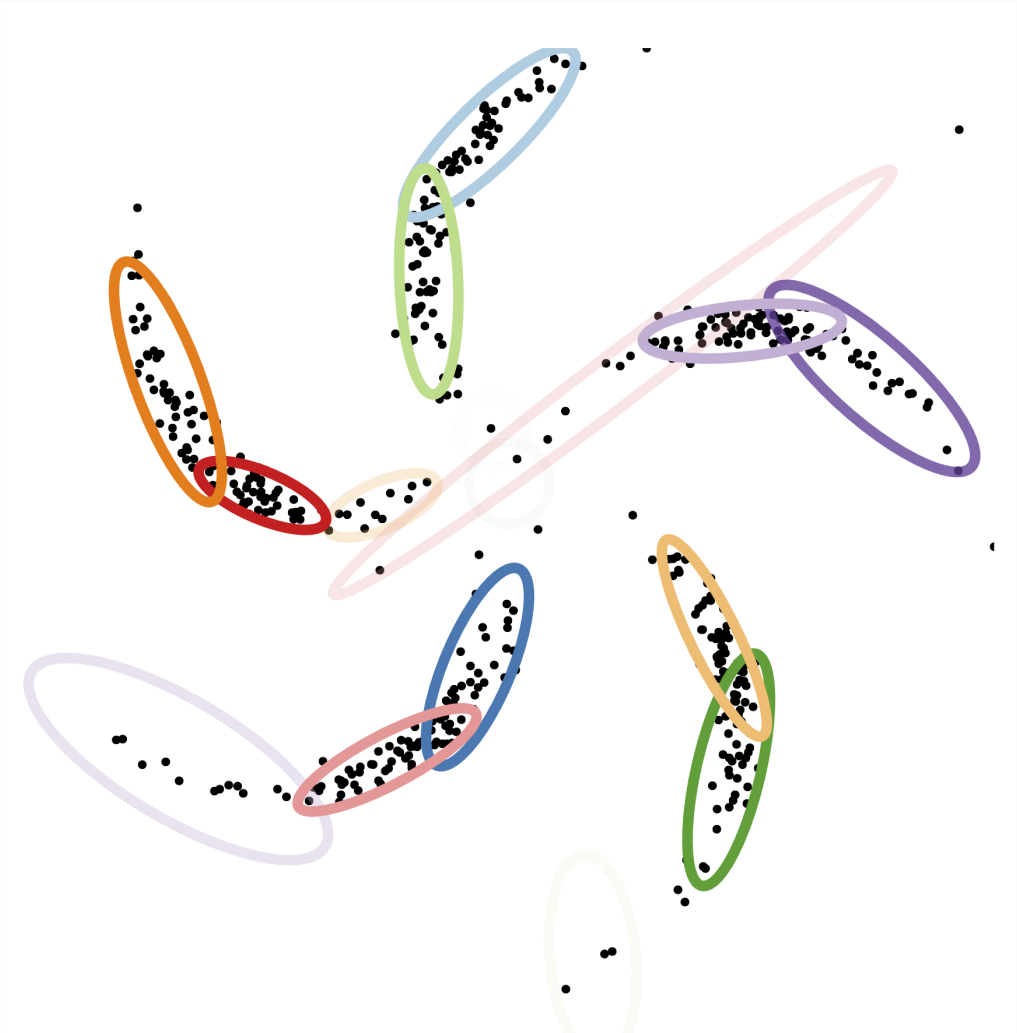
—George Box

What is the Purpose of a Model?

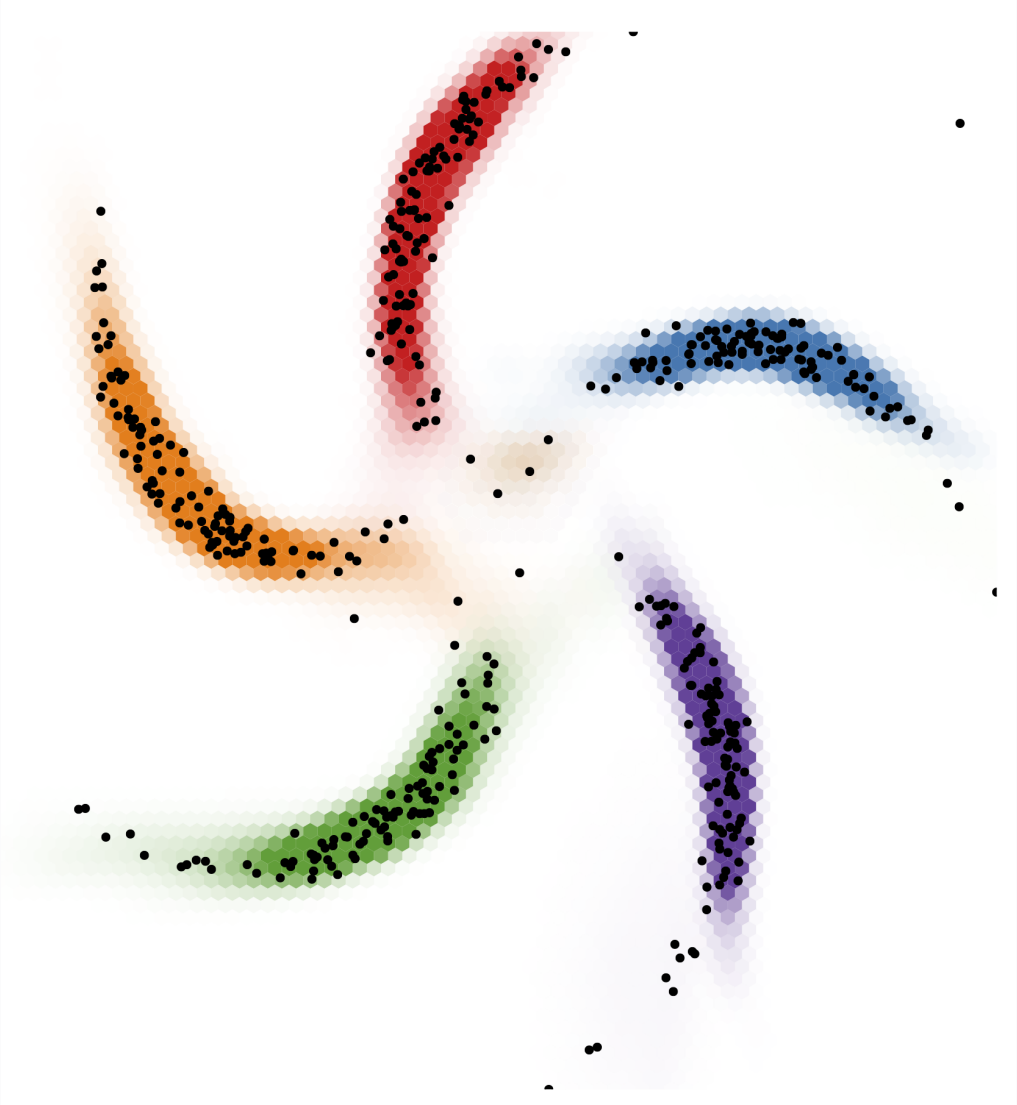
- The purpose of a model is to help provide insights into a target problem or data and sometimes to further use these insights to make predictions
- Its purpose is **not** to try and fully encapsulate the "true" generative process or perfectly describe the data
- There are infinite different ways to generate any given dataset
  - Trying to uncover the "true" generative process is not even a well-defined problem
- In any real–world scenario, no Bayesian model can be "correct"
  - The posterior is inherently subjective
- It is still important to criticize—models can be very wrong!
  - E.g. we can use frequentist methods to falsify the likelihood

Some Models are Much Better than Others

Some Models are Much Better than Others



Some Models are Much Better than Others



Bayesian Modelling as Multiple Hypotheses

Bayesian models are rooted in **hypotheses**:

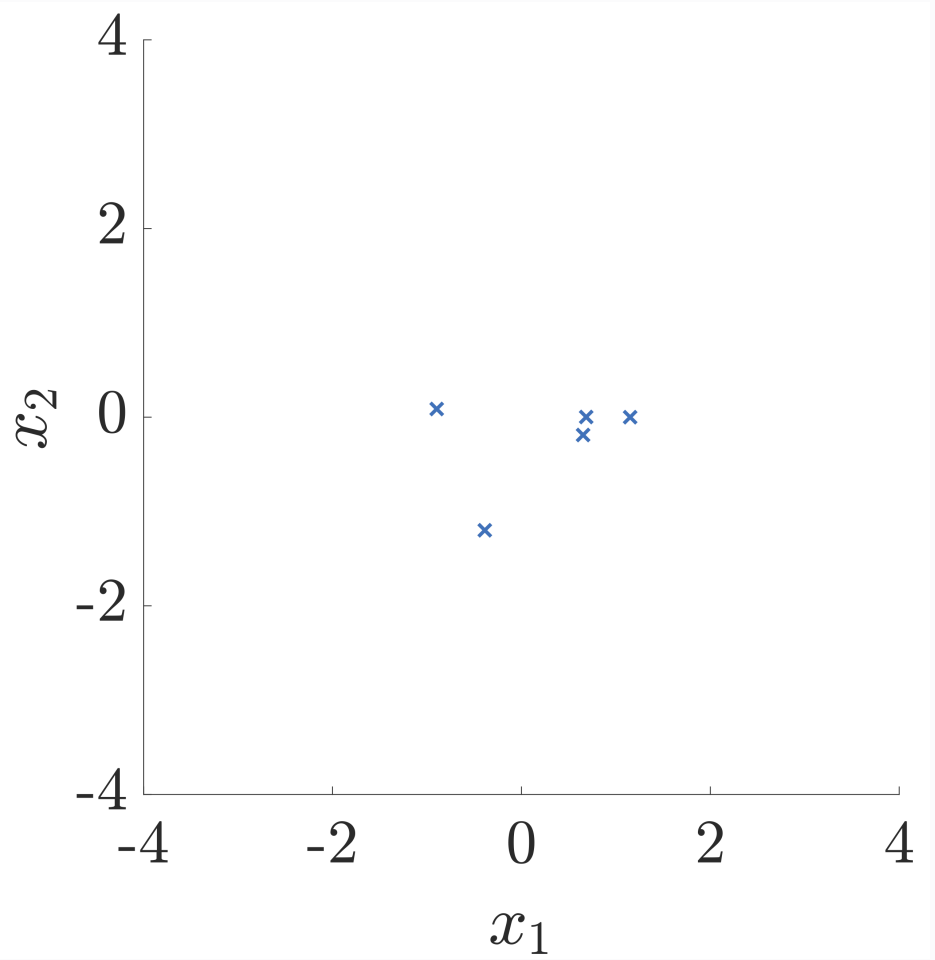
- Each possible instance of our parameters θ represents a hypothesis
- The prior $p(\theta)$ represents our initial (subjective) belief that each such hypothesis is true
- The likelihood $p(\mathcal{D}|\theta)$ is the probability of generating data $\mathcal{D}$ if we assume hypothesis θ
- The posterior $p(\theta|\mathcal{D})$ is our belief that each hypothesis is true given that the actual dataset generated
- The posterior predictive is a posterior–weighted sum of the predictive models from all possible hypotheses

Example: Density Estimation

Presume that we decide to use an isotropic Gaussian likelihood with unknown mean θ to model the data on the right:

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{n} \mathcal{N}(x_i; \theta, I)$$

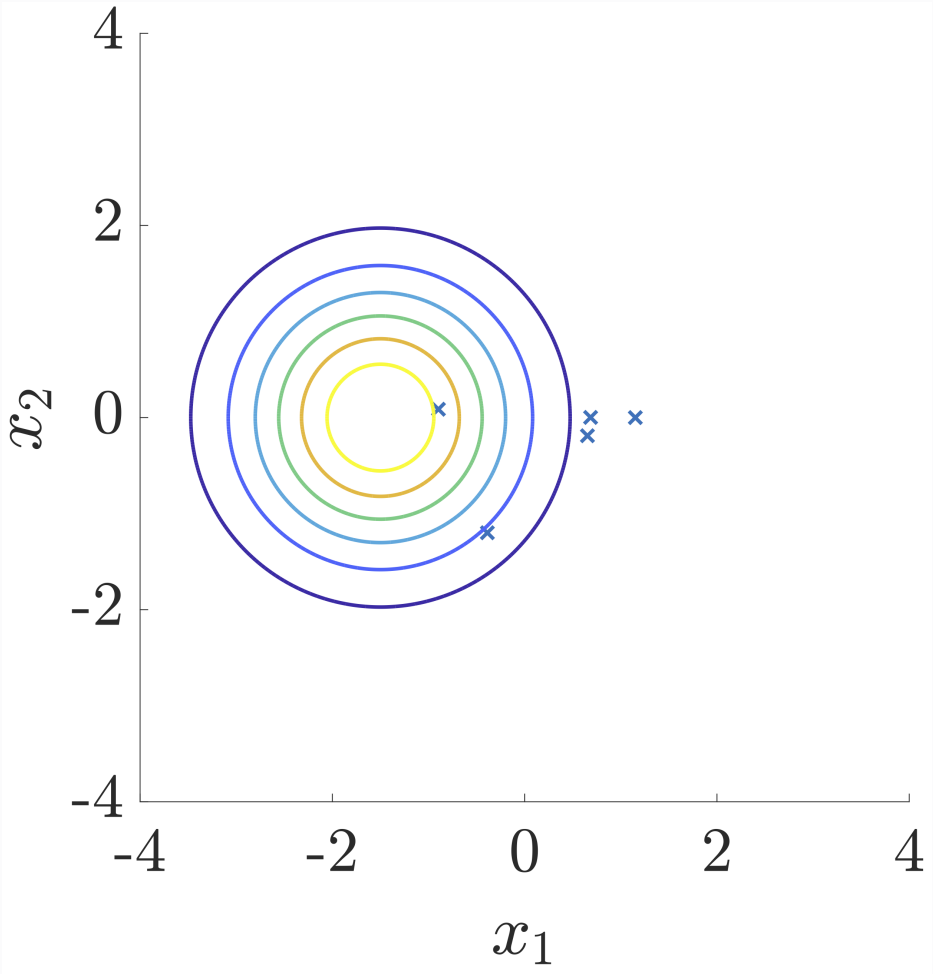
where I is a two-dimensional identity matrix



Example: Density Estimation

Hypothesis 1: $\theta = [-2, 0]$

$$p(\mathcal{D}|\theta = [-2, 0]) = 0.00059 \times 10^{-5}$$



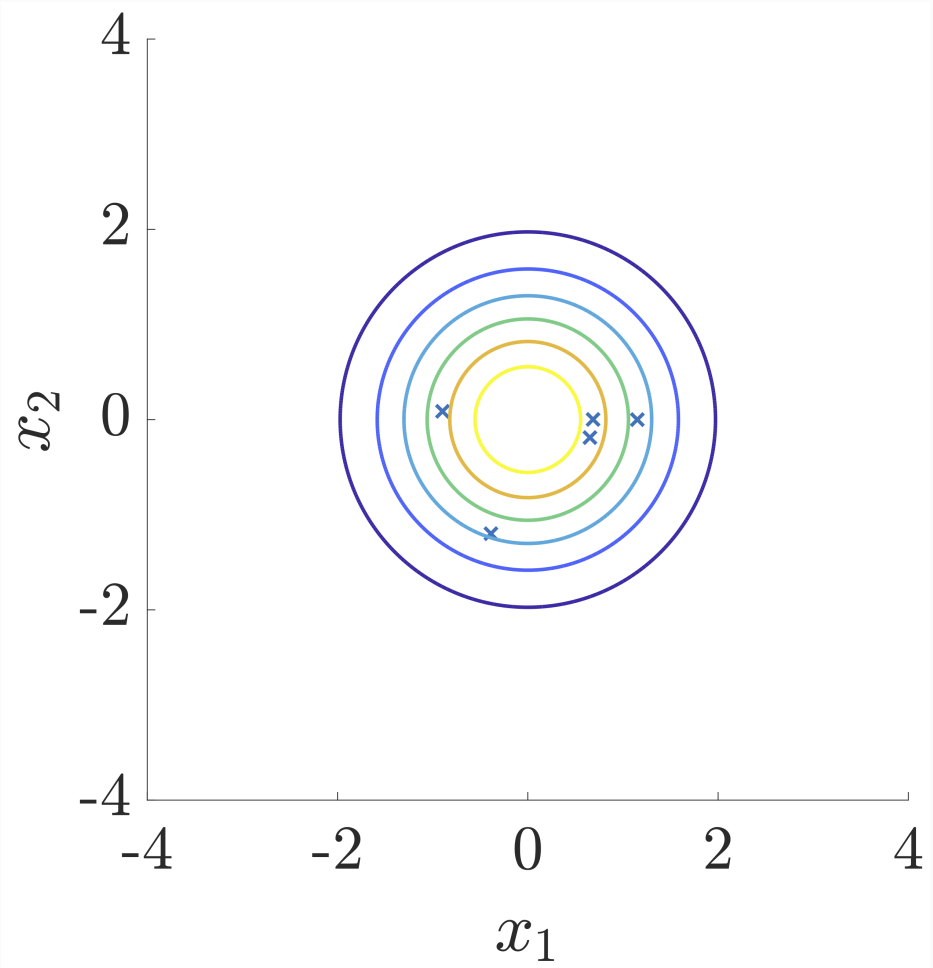
Example: Density Estimation

Hypothesis 1: $\theta = [-2, 0]$

$$p(\mathcal{D}|\theta = [-2, 0]) = 0.00059 \times 10^{-5}$$

Hypothesis 2: $\theta = [0, 0]$

$$p(\mathcal{D}|\theta = [0, 0]) = 0.99 \times 10^{-5}$$



Example: Density Estimation

Hypothesis 1: $\theta = [-2, 0]$

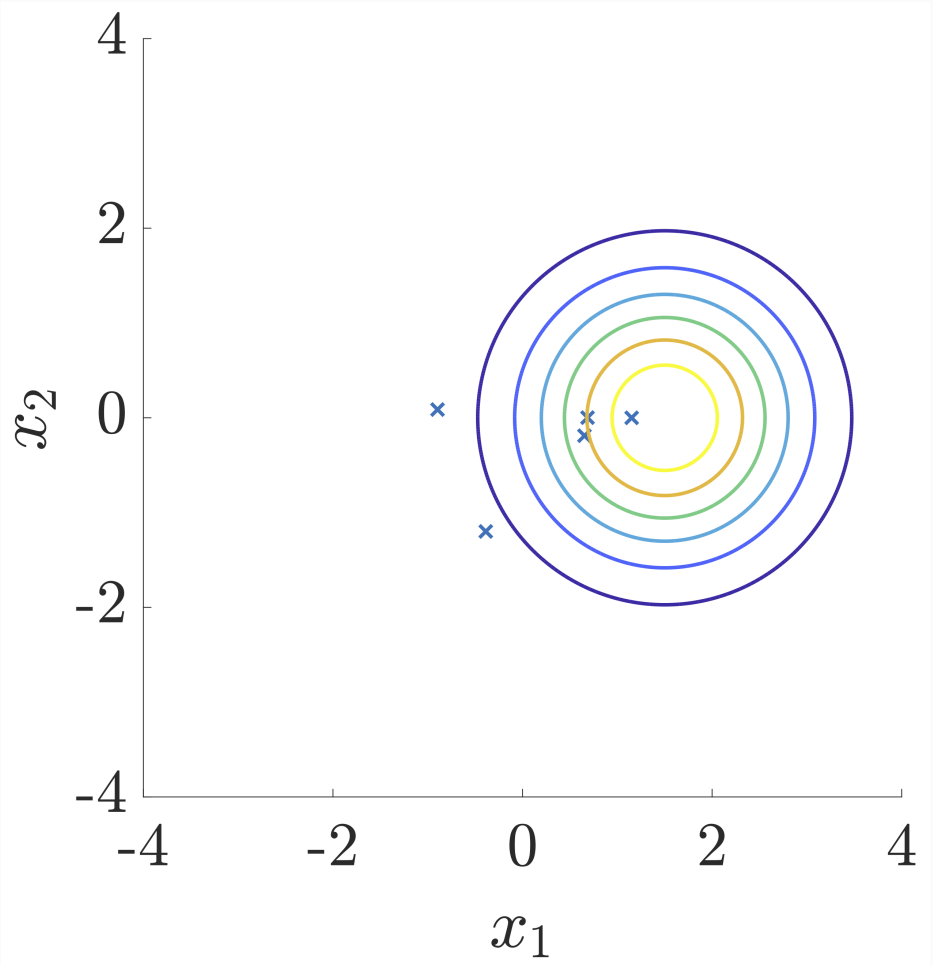
$$p(\mathcal{D}|\theta = [-2, 0]) = 0.00059 \times 10^{-5}$$

Hypothesis 2: $\theta = [0, 0]$

$$p(\mathcal{D}|\theta = [0, 0]) = 0.99 \times 10^{-5}$$

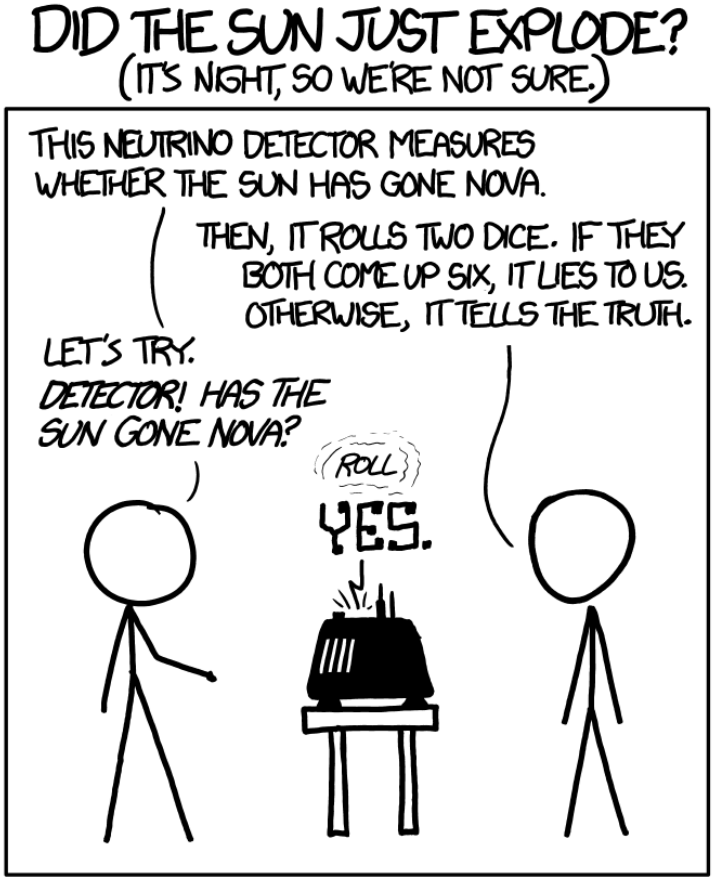
Hypothesis 3: $\theta = [2, 0]$

$$p(\mathcal{D}|\theta = [2, 0]) = 0.021 \times 10^{-5}$$



Bayesian Modelling Allows us To Express our Prior Beliefs

- In the above example we only considered the likelihood of each hypothesis
- We may though have unequal prior beliefs about each hypothesis



https://xkcd.com/1132/

The Posterior Predictive Averages over Hypotheses

The posterior predictive distribution allows us to **average** over each of our hypotheses, weighting each by their posterior probability.

In our density estimation example, we might have the prior,

$$p(\theta) = \begin{cases} 0.05 & \text{if} \quad \theta = [-2, 0] \\ 0.05 & \text{if} \quad \theta = [0, 0] \\ 0.9 & \text{if} \quad \theta = [2, 0] \end{cases}$$

The Posterior Predictive Averages over Hypotheses (2)

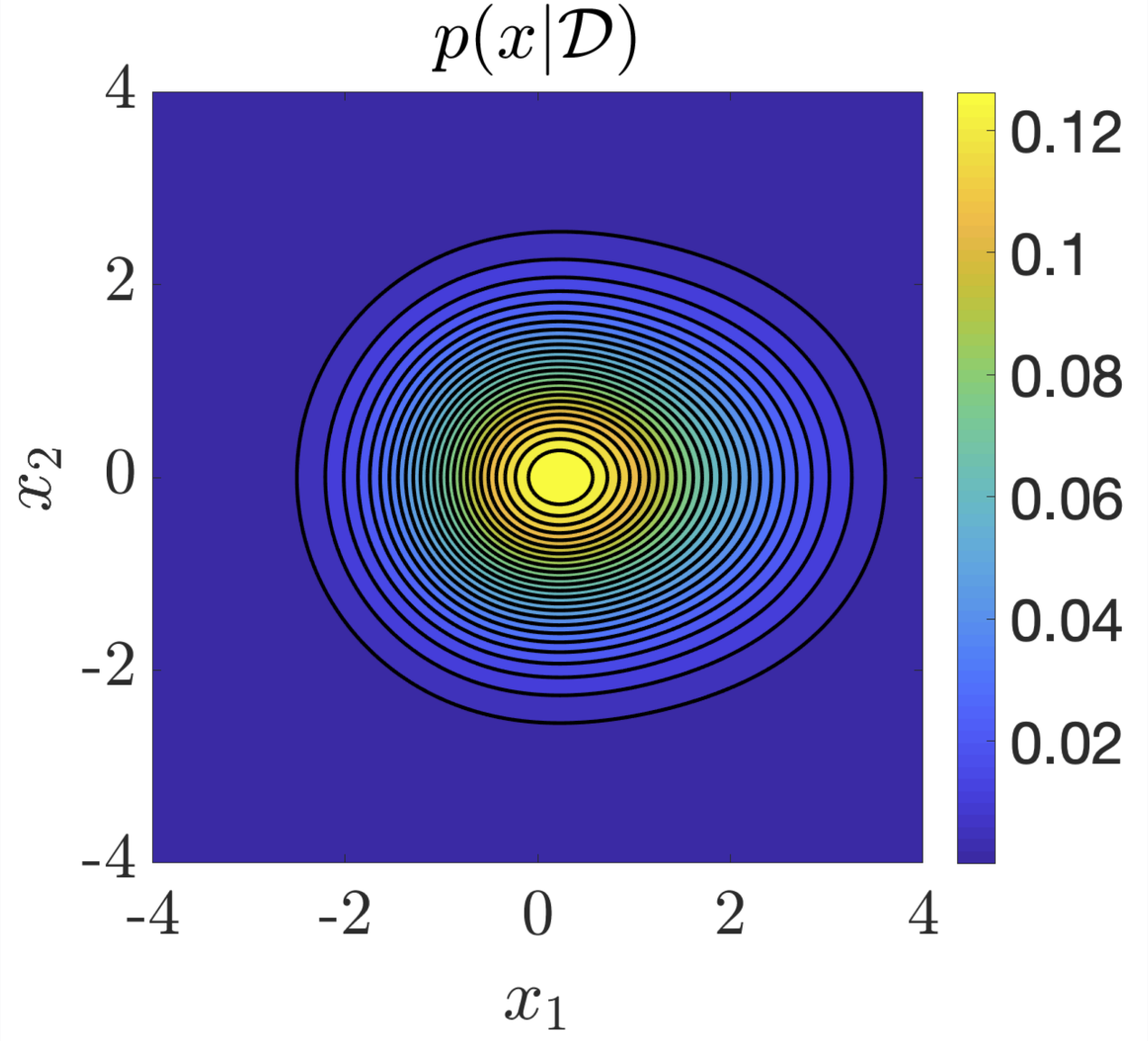
Apply Bayes rule with our previous likelihood leads to the posterior

$$p(\theta|\mathcal{D}) = \begin{cases} 0.0004 & \text{if} \quad \theta = [-2, 0] \\ 0.716 & \text{if} \quad \theta = [0, 0] \\ 0.283 & \text{if} \quad \theta = [2, 0] \end{cases}$$

and thus the following posterior predictive that is a weighted sum of the three possible predictive distributions

$$\begin{aligned} p(x|\mathcal{D}) = & \, 0.0004 \times \mathcal{N}(x; [-2, 0], I) \\ & + 0.716 \times \mathcal{N}(x; [0, 0], I) \\ & + 0.283 \times \mathcal{N}(x; [2, 0], I) \end{aligned}$$

The Posterior Predictive Averages over Hypotheses (3)



An Important Subtlety

- Even though we average over θ, a Bayesian model is still implicitly assuming that there is still a single true θ
  - The averaging over hypotheses is from our own uncertainty as to which one is correct
  - This can be problematic with lots of data given our model is an approximation
- In the limit of large data, the posterior is guaranteed to collapse to a point estimate (given some weak assumptions):

$$p(\theta|x_{1:n}) \to \delta(\theta = \hat{\theta}) \quad \text{as} \quad n \to \infty \tag{1}$$

- The value of $\hat{\theta}$ and the exact nature of this convergence is dictated by the Bernstein–von Mises Theorem (see the notes)
- Note that, subject to mild assumptions, $\hat{\theta}$ is independent of the prior
  - With enough data, the likelihood always dominates the prior

Model Comparison: What Makes a Good Model?

Model Comparison: What Makes a Good Model?

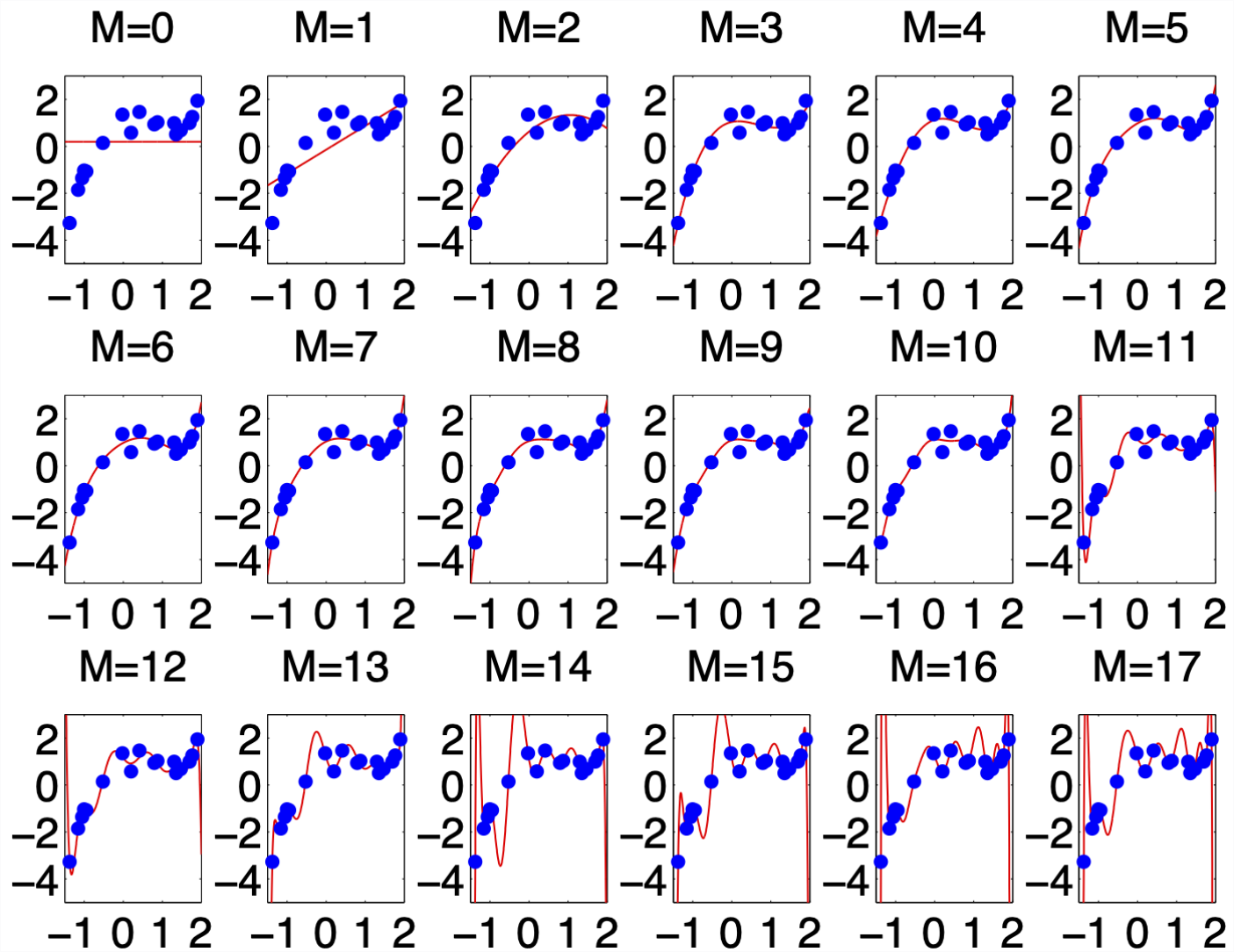


Image credit: Carl Rasmussen

Model Selection: Using the Evidence

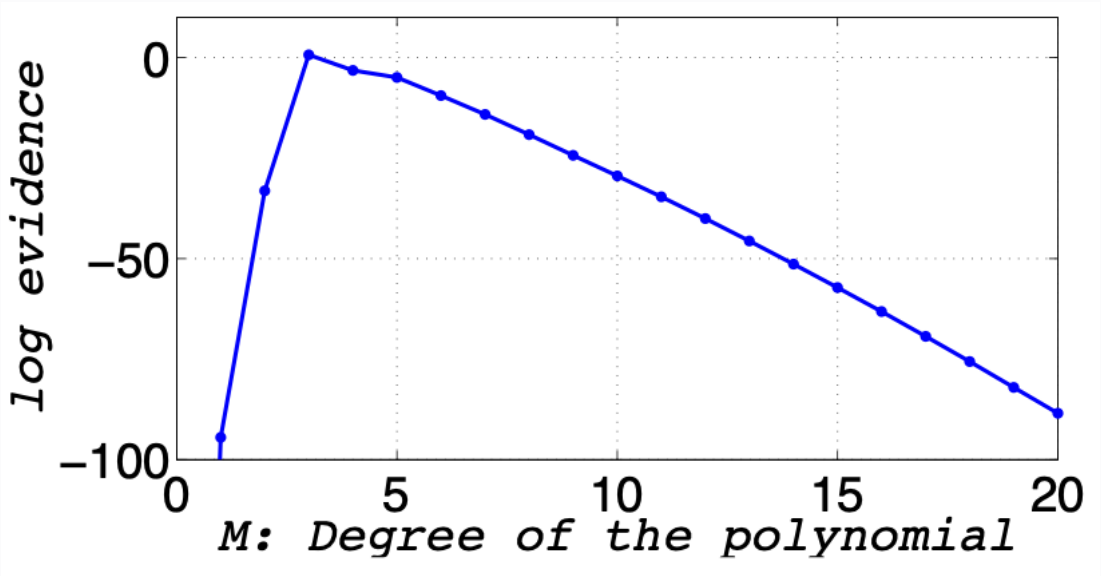
- Let's revisit Bayes' rule and now condition on the model m:

$$p(\theta|\mathcal{D}, m) = \frac{p(\mathcal{D}|\theta, m)p(\theta|m)}{p(\mathcal{D}|m)} \tag{2}$$

- The marginal likelihood of a model $p(\mathcal{D}|m)$ represents the probability of the data under the model, averaging over all possible parameter values.
- We can use this **model evidence** to choose between models: a high marginal likelihood generally indicates a good model
- The ratio of two marginal likelihoods, e.g. $p(\mathcal{D}|m_1)/p(\mathcal{D}|m_2)$, is known as a **Bayes factor**

Marginal Likelihoods for Polynomial Regression

Returning to our polynomial regression problem and now taking a Bayesian approach with a Gaussian prior on the weights, we see that the model evidence prefers a degree of $M = 3$



This is a "sweet-spot": complex enough to accurately match the data, simple enough to retain strong predictive power

Image credit: Carl Rasmussen

Why Marginal Likelihoods?

Why should we use the model evidence to compare models?

Apply Bayes rule to the models themselves:

$$p(m|\mathcal{D}) = \frac{p(\mathcal{D}|m)p(m)}{p(\mathcal{D})} \tag{3}$$

This give us a direct relationship between the model evidence and the posterior probability of that model

If we a priori have no preference between models such that $p(m)$ is uniform, we even get that $p(m|\mathcal{D}) \propto p(\mathcal{D}|m)$

Bayesian Occam's Razor

- Occam's Razor states that if two explanations are able to explain a set of observations, the simpler one should be preferred.
- We can apply this in a Bayesian context by noting that the marginal likelihood is the probability that randomly selected parameters from the prior would generate $\mathcal{D}$
- Models that are too simple are unlikely to generate the observed dataset.
- Models that are too complex can generate many possible datasets, so again, they are unlikely to generate that particular dataset at random.

Bayesian Occam's Razor (2)

Imagine a hypothetical order on datasets where they get more complicated as we move away from the origin.

The model with highest evidence is the one that is powerful enough to explain that data but not anything more complicated.

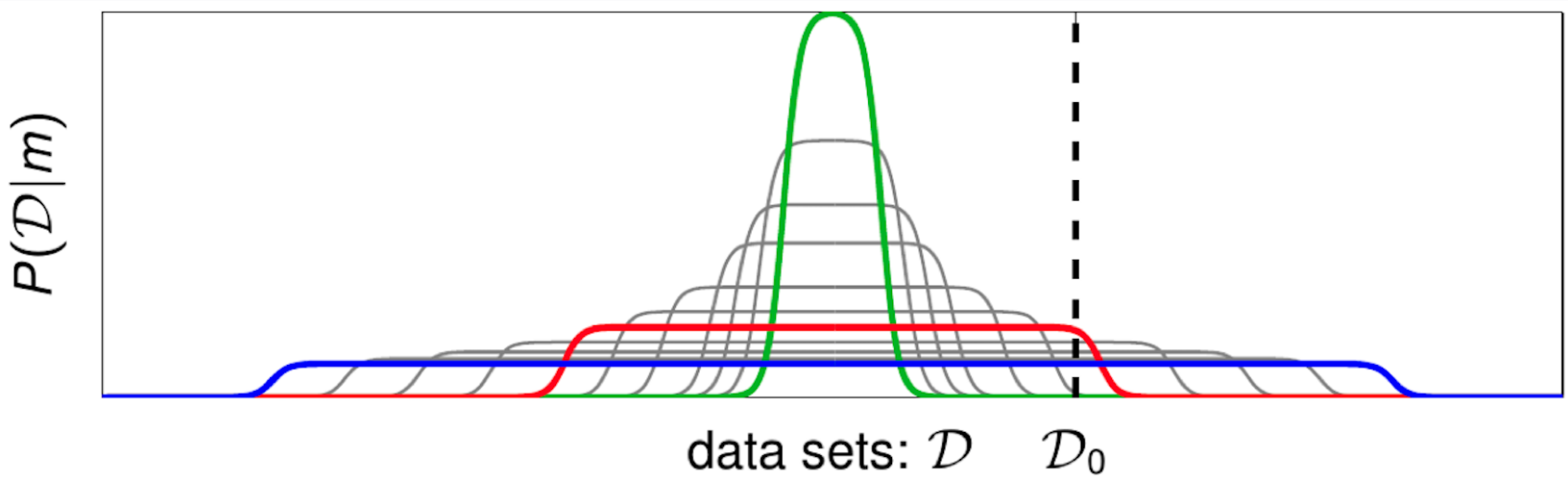


Image credit: Maneesh Sahani

Recap

- Bayesian models provide a probabilistic way of reasoning about the world
- All models are **wrong**, but they can still be **useful** by allowing us to make predictions and incorporate prior information
- The posterior predictive can be thought of as averaging over model hypotheses: for **finite data** it can produce more complex predictive models than using any single θ
- We can compare models using their **model evidence**
- Bayesian Occam's Razor shows that the model evidence incorporates both the accuracy to which data is explained and the predictive power of the model

Further Reading

- C M Bishop. **Pattern recognition and machine learning**. 2006, Chapters 1-3
- K P Murphy. **Machine learning: a probabilistic perspective**. 2012, Chapter 5
- D Barber. **Bayesian reasoning and machine learning**. 2012, Chapter 12
- Zoubin Ghahramani on Bayesian machine learning (there are various alternative variations of this talk): `https://www.youtube.com/watch?v=y0FgHOQhG4w`
- Iain Murray on Probabilistic Modeling `https://www.youtube.com/watch?v=pOtvyVYAuW4`