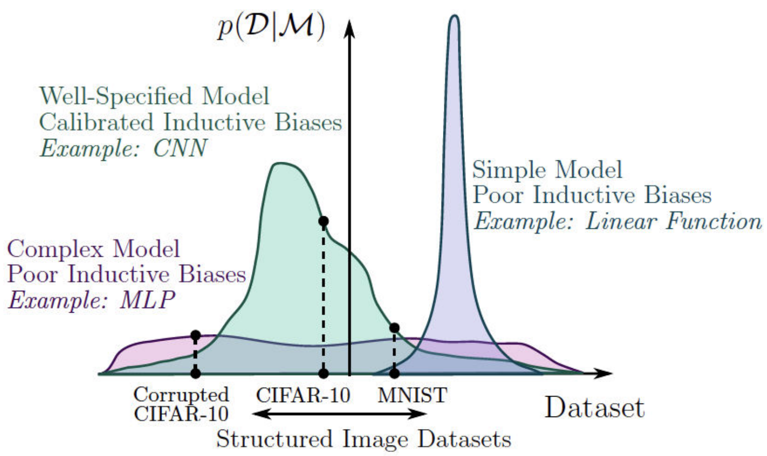
https://jorisbaan.nl/2021/03/02/introduction-to-bayesian-deep-learning.html

Bayesian Probability - the **belief that something will happen**, which is a belief we update given new information

- **Prior -** Our old belief of seeing something
- **Likelihood -** evidence, something that we have observed and its associated probability distribution
- **Posterior -** out new belief, which is the old belief updated with the new evidence that we have gained

Support - range of dataset classes that a model can support

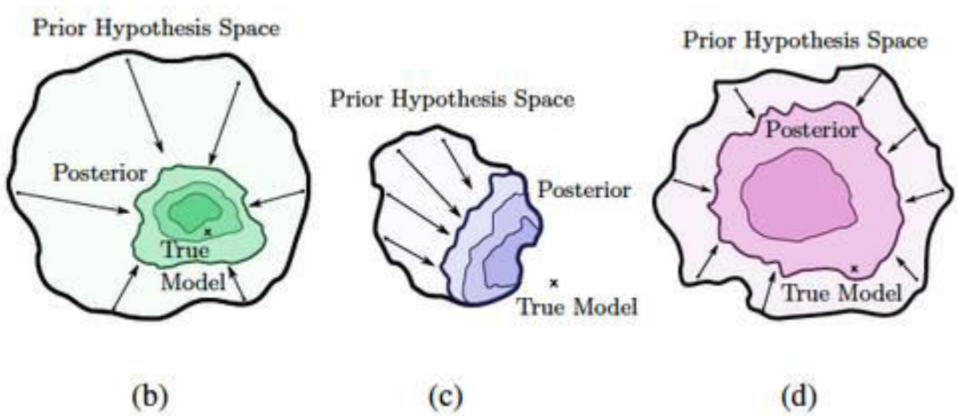
- Range of functions a model can represent
- **Inductive Bias -** how **good** a model class is at fitting a specific dataset
- **Generalisation of a model -** depends on two properties
    - **Support - how many** functions a model can support
    - **Inductive Bias - how good** is the performance



The Marginal Likelihood (Bayesian Evidence) - performance of our model, how well it fits a dataset

The Bayesian Posterior - Our model, which should get the right solution with the right inductive bias

- **Prior hypothesis space** should have a **broad support**



Key Idea - We use **marginalization** instead of **optimisation**

Frequentist Approach - it is to **optimise** a loss function to obtain optimal parameters

- We try to get a parameter **point estimate**
- e.g. using SGD with cross entropy and backpropagation
- We try to **maximise** the **likelihood** p(D|w,M)
    - **w -** parameters (weights)
    - **D -** Dataset
    - **M -** Model (often left out)
    - We try to pick **w** such that we maximise the **p**robability of observed dataset **D** given our model **M**
    - Called **Maximum Likelihood Estimation**, a special case of **Maximum a posteriori (MAP)** estimation with a uniform prior

Bayesian Approach - It is to quantify uncertainty

- We get a **full probability distribution** over parameters, called a **posterior distribution**
- Represents how much uncertainty we have about each parameter
- We use **Bayes' Theorem** to compute the **posterior distribution**

$$posterior = \frac{likelihood * prior}{evidence}, \; or \; p(w|\mathcal{D}) = \frac{p(\mathcal{D}|w)p(w)}{p(\mathcal{D})}$$

$$evidence = p(\mathcal{D}) = \int p(\mathcal{D}|w)p(w)dw$$

- **Posterior p(w|D) -** represents our **belief/hypothesis/uncertainty** about parameters **after** getting data
- **Prior Distribution p(w) -** Belief about what our model parameters are **before (prior)** to getting data
- **Likelihood p(D|w) -** How well our data is explained by our parameters
  - This is the same as thing as in the frequentist approach
  - **Loss function** of our parameters
- **Marginalisation -** use **marginal likelihood p(D) (Bayesian (model) evidence)** to get **probability distribution**
  - This normalises the data
  - Marginalisation = **summing and integrating** over all parameter settings
  - Tells us **how likely the data is** over **all possible parameter settings**
  - Provides **evidence** for how good our model is
  - *Helps us get new probabilities given existing probabilities*

Bayesian approach is about **marginalization** rather than **optimization**

- It is hard to find p(D) exactly
- Therefore instead of optimising **w**, we find the best **distribution** for it
- We do this by using:
  - **Sampling methods** (Markov Chain Monte Carlo MCMC)
  - **Variational Inference -**
  - **Normalizing flows -**
- For **generative** models, we use:
  - **VAE** (similar to variational inference) **-**