*Mathematics of Deep Learning Reading Group.*

*April 28th, 2022*

# Developments in Inference with Linearised Neural Networks

Riccardo Barbano, Javier Antorán

UNIVERSITY OF CAMBRIDGE

# Reading group outline

1. **Preliminaries**: probabilistic inference in neural networks and the linearised Laplace method

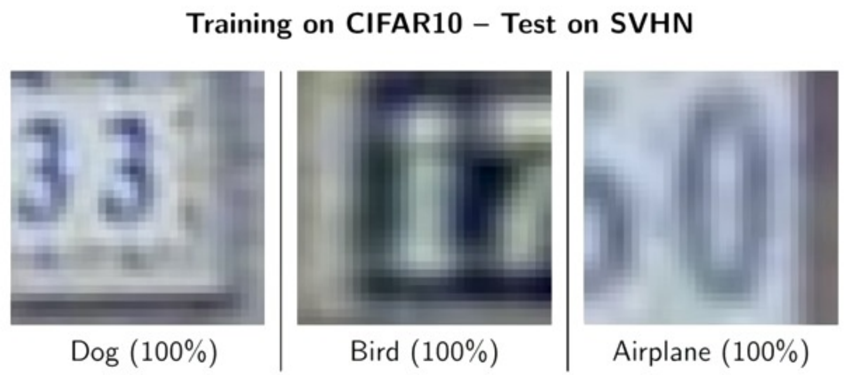
2. **Paper overview**: "Improving predictions of Bayesian neural networks via local linearization"

3. **Informal discussion**

# Preliminaries

# Preliminaries: open problems in deep learning

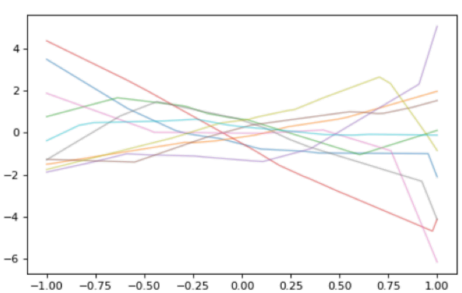
## Overconfidence



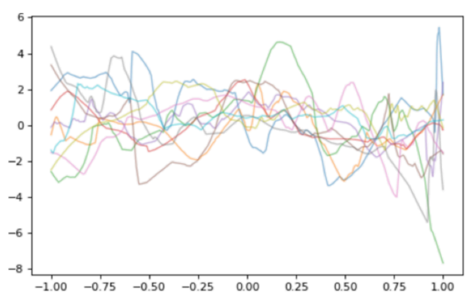

https://vitalab.github.io/article/2019/07/11/overconfident.html

## Model Selection

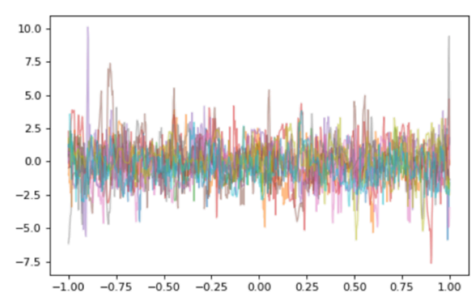
### 1 Hidden Layer




### 5 Hidden Layer

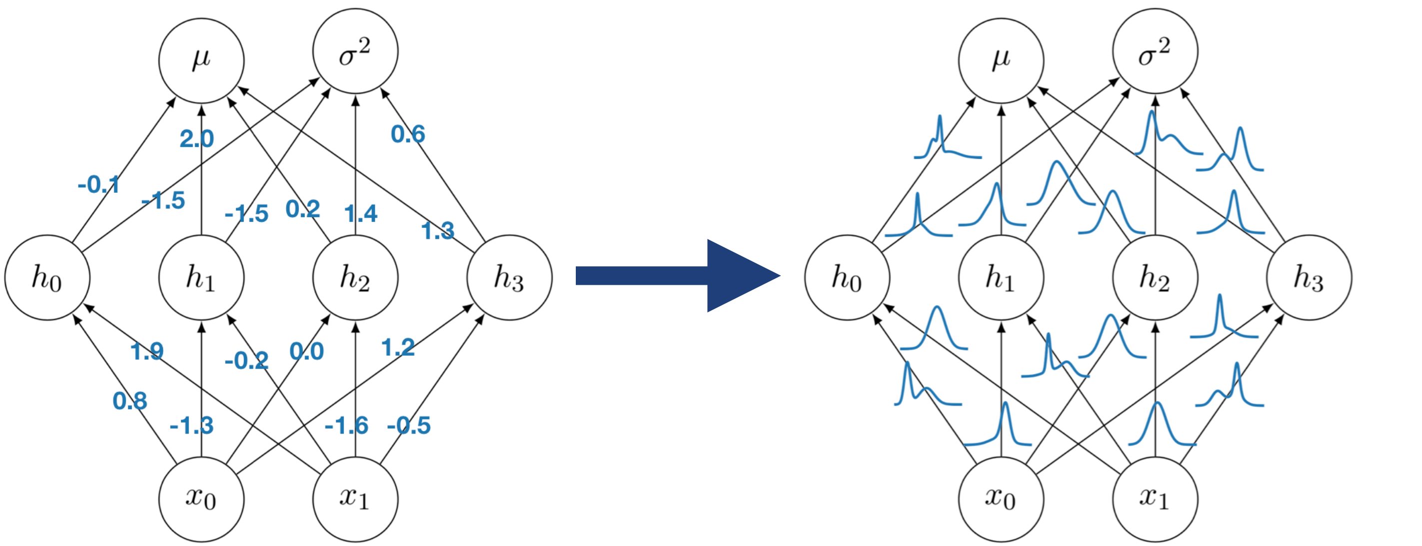



### 20 Hidden Layer



# Preliminaries: probabilistic inference in NNs

1. Place a prior distribution $\pi(\theta)$ over NN parameters.
2. Define some likelihood function $p(y \mid f(\theta, x))$ to characterise the agreement of the NN function $f(\theta, \cdot\,)$ with the observations $(y, x)$
3. Update the weight distribution using Bayes' rule

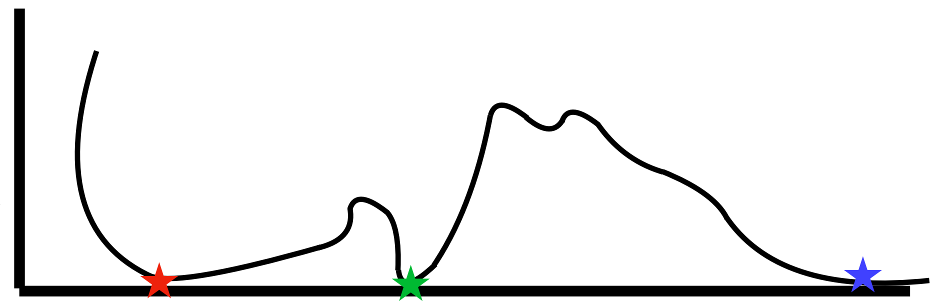


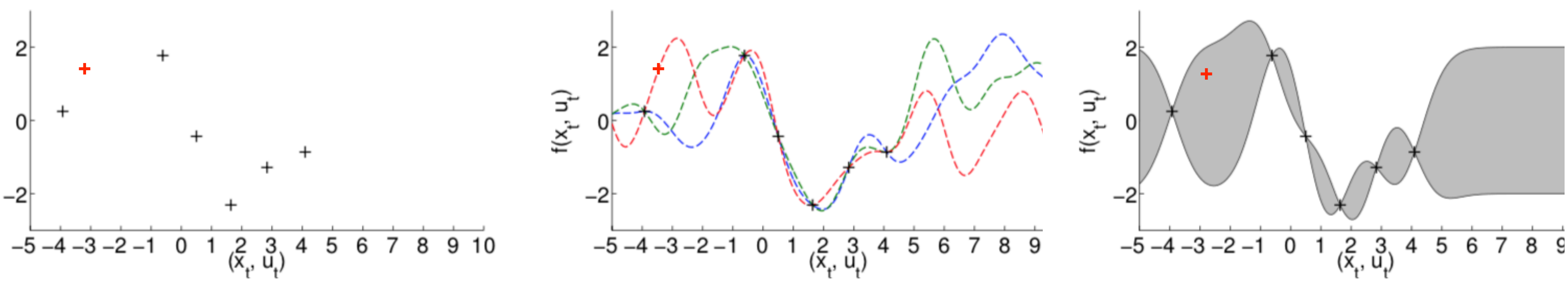

$$\tilde{\theta} \in \operatorname{argmax}_\theta \underbrace{\log p(y \mid f(\theta, x))}_{\text{data fit}} + \underbrace{\log \pi(\theta)}_{\text{regulariser}}$$

$$p(\theta \mid x, y) = \frac{p(y \mid f(\theta, x))\pi(\theta)}{p(y \mid x)}$$

# Preliminaries: uncertainty estimation

$$\mathscr{L}_f(\theta) = \underbrace{-\log p(y\,|\,f(\theta, x))}_{\text{data fit}} \underbrace{-\log \pi(\theta)}_{\text{regulariser}}$$



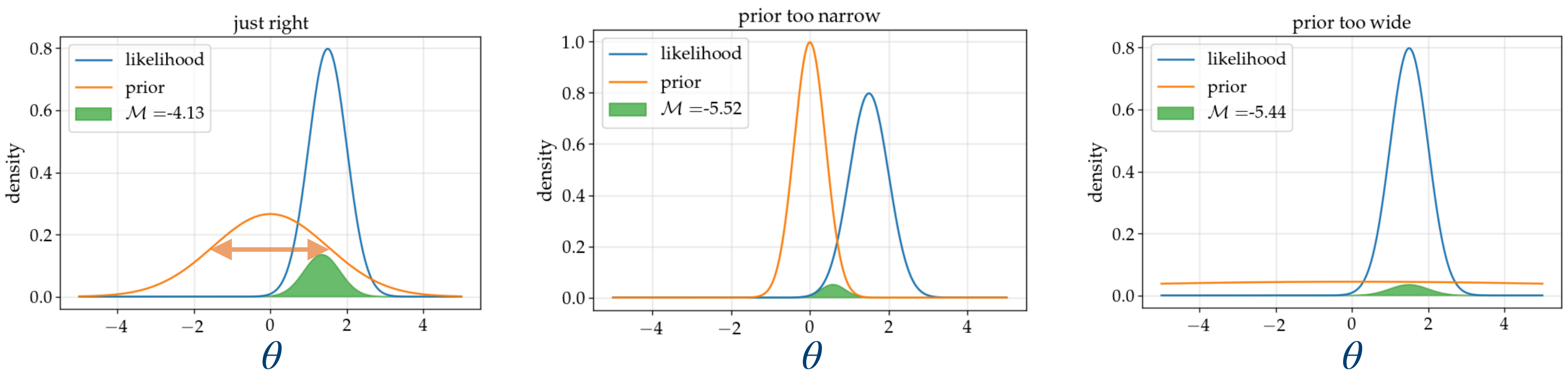

$$p(\theta\,|\,y, x) = \frac{1}{\exp(\mathscr{M})}\exp(-\mathscr{L}_f(\theta)) \qquad f(\theta, \cdot\,),\ \ \theta \sim p(\theta\,|\,y, x)$$

# Preliminaries: model selection

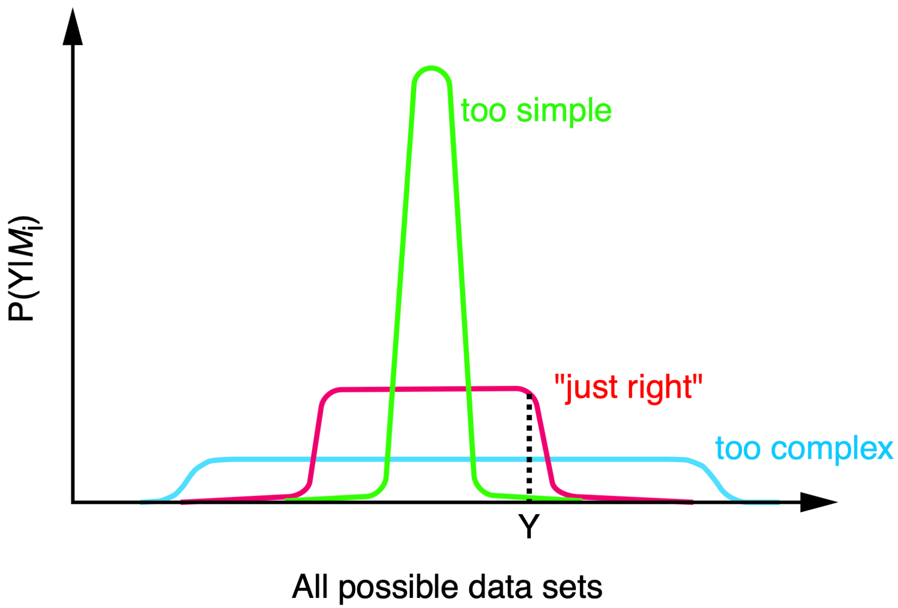
$$\mathscr{L}_f(\theta) = -\log p(y\,|\,f(\theta,x)) - \log \pi(\theta) \quad p(\theta\,|\,y,x) = \frac{1}{\exp(\mathscr{M})}\exp(-\mathscr{L}_f(\theta))$$

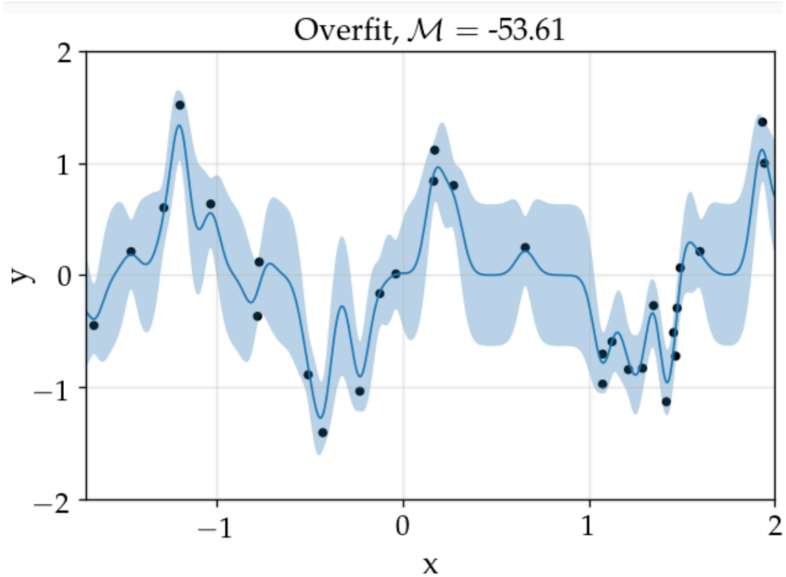
The normalisation constant, $\mathscr{M}$, is the ***marginal likelihood***, or ***model evidence***. It is the probability that our observations where generated by our prior. It provides an objective for hyperparameter selection without the need for validation data.

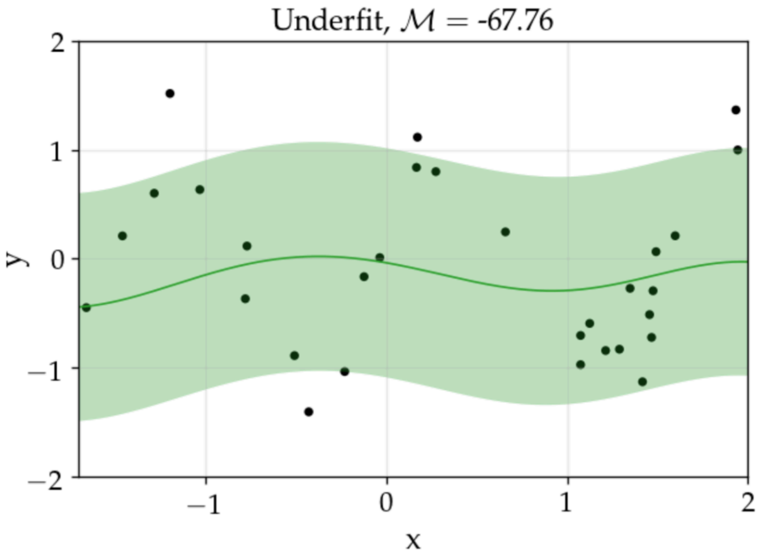
$$\mathscr{M} = \log p(y\,|\,x) \;= \log \int p(y\,|\,f(\theta,x))d\pi \;= \log \int \exp(-\mathscr{L}_f(\theta))d\nu$$

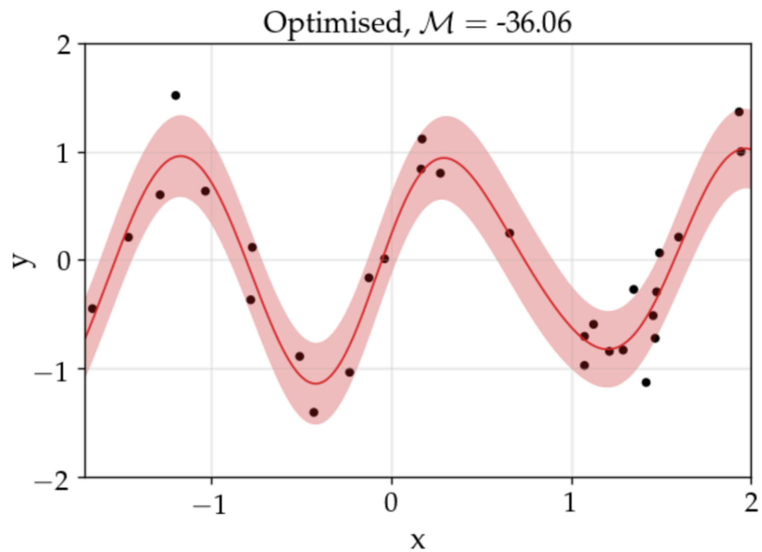


# Preliminaries: automatic Occam's razor



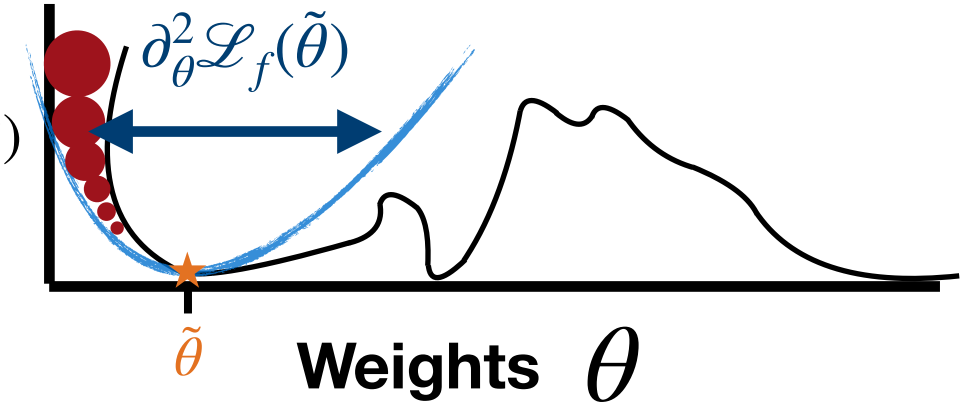



# Preliminaries: the Laplace approximation

$$\mathscr{L}_f(\theta) = -\log p(y \mid f(\theta, x)) - \log \pi(\theta)$$

$$\mathscr{M} = \log \int \exp(-\mathscr{L}_f(\theta)) d\nu$$

For NNs this integral is intractable




**Idea**: Find a mode of $\mathscr{L}_f$: $\tilde{\theta}$ and perform 2-order Taylor expansion

$$\mathscr{G}_f(\theta) = \mathscr{L}_f(\tilde{\theta}) + ||\theta - \tilde{\theta}||^2_{\partial^2_\theta \mathscr{L}_f(\tilde{\theta})}$$

By inspection, $\exp(-\mathscr{G}_{f,\tilde{\theta}}(\theta))$ is proportional to $\mathscr{N}(\tilde{\theta}, (\partial^2_\theta \mathscr{L}_f(\tilde{\theta}))^{-1})$ where

$$\partial^2_\theta \mathscr{L}_f(\tilde{\theta})) = \partial^2_\theta \log p(y \mid f(\tilde{\theta}, x)) + \partial^2_\theta \log \pi(\tilde{\theta}) \qquad \pi(\theta) \rightarrow \mathscr{N}(\theta;\ 0, \Lambda^{-1})$$

**Issue**: A lot of mass falls in low density region, leading to bad predictions

PAPER DISCUSSION:

IMPROVING PREDICTIONS OF BAYESIAN NEURAL NETWORKS VIA LOCAL LINEARIZATION