GenU @ University of Copenhagen

Imperial College London

Inductive Biases, Input Densities, and Predictive Uncertainty

Mark van der Wilk

Department of Computing
Imperial College London

@markvanderwilk
m.vdwilk@imperial.ac.uk

Oct 13, 2021

Questions

- How do uncertainty and inductive bias interact?
- What is good behaviour of predictive error bars?
- Should we be uncertain ”far away” from the training data?
- Can we use input density as a metric for predictive uncertainty?

How should we measure uncertainty quality?

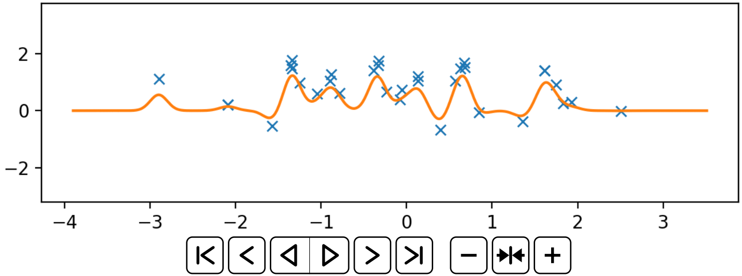
- Toy examples to illustrate what it looks like when it **works**
- Inspiration for new ways to measure and probe behaviour?
- It’s early, let’s look at some pretty pictures (need Acrobat for animations)

Minimising training loss

We're looking for a fit that will **generalise** to new unseen test data. Let's minimise the training loss of the posterior mean.

$$\mathcal{L}(\theta, \sigma) = \sum_{n=1}^{N} \left[k_\theta(\mathbf{x}_n, X)\left(\mathbf{K}_\theta + \sigma^2 \mathbf{I}\right)^{-1} \mathbf{y} - y_n \right]^2 \tag{1}$$

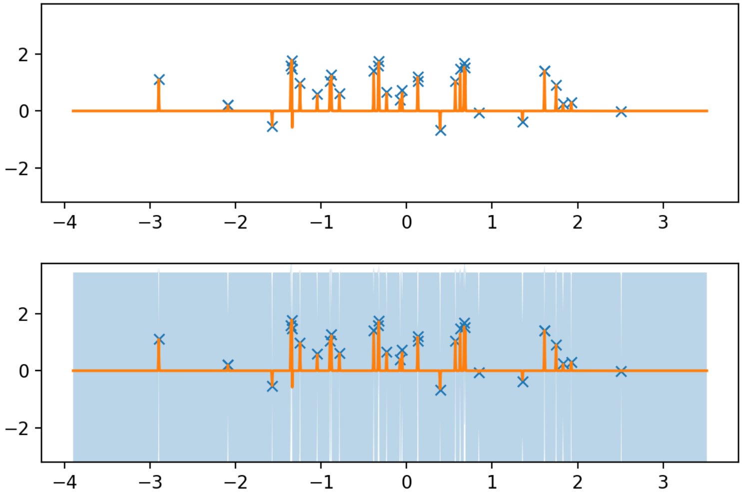
$$\{\theta^*, \sigma^*\} = \operatorname*{argmin}_{\theta, \sigma} \mathcal{L}(\theta, \sigma) \tag{2}$$



We can fit anything with a tiny lengthscale and noise variance!

How does uncertainty help?

Does uncertainty help against the overfitting?



Model Selection according to Bayes

Model selection from a Bayesian point of view:

$$p(f, \boldsymbol{\theta} \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid f) p(f \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathbf{y})}$$

$$= \underbrace{\frac{p(\mathbf{y} \mid f) p(f \mid \boldsymbol{\theta})}{p(\mathbf{y} \mid \boldsymbol{\theta})}}_{p(f \mid \mathbf{y}, \boldsymbol{\theta})} \underbrace{\frac{p(\mathbf{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathbf{y})}}_{p(\boldsymbol{\theta} \mid \mathbf{y})}$$

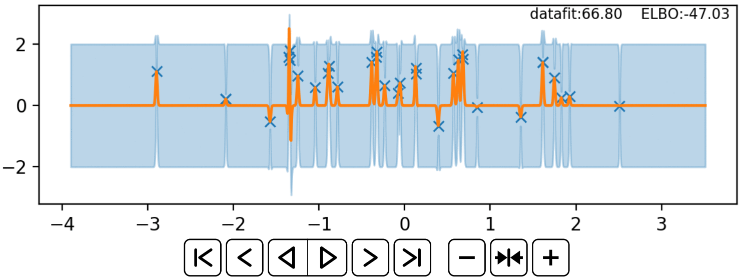
Key quantity for model selection is the **marginal likelihood**

$$p(\mathbf{y} \mid \boldsymbol{\theta}) = \int p(\mathbf{y} \mid f) p(f \mid \boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$$

By handing our uncertainty on $f(\cdot)$ in a Bayesian way, we also get the marginal likelihood for model selection.

Marginal likelihood fixes things

Instead, choose hyperparameters by maximising marginal likelihood:



In above $\mathcal{L}$ is indicated by 'datafit', while 'ELBO' indicates the marginal likelihood.

- More sensible fit as the marginal likelihood rises
- Datafit gets worse!

Marginal likelihood trades off
data fit and **model complexity**.

Why does marginal likelihood work?

We have seen

- Minimising training error doesn't work
- Uncertainty doesn't necessarily help, but does make us more cautious
- Marginal likelihood seems to trade-off complexity and data fit

But **why** does the marginal likelihood lead to models that generalise well?

Marginal likelihood as incremental prediction

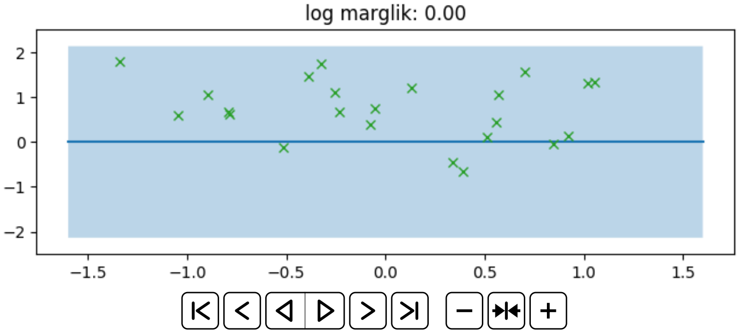
We can split the marginal likelihood up using the **product rule**:

$$p(\mathbf{y}) = p(y_1)p(y_2|y_1)p(y_3|\{y_i\}_{i=1}^2)\dots \tag{3}$$

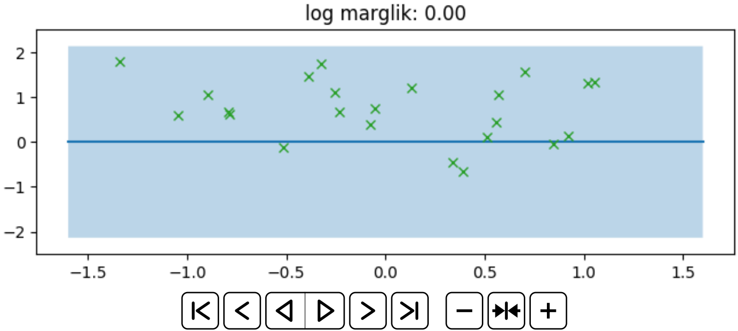
$$= \prod_{n=1}^{N} p(y_n|\{\mathbf{x}_i, y_i\}_{i=1}^{n-1}) \tag{4}$$

- The marginal likelihood measures how well previous training points predict the next one
- If it continuously predicted well on all N points previously, it probably will do well next time

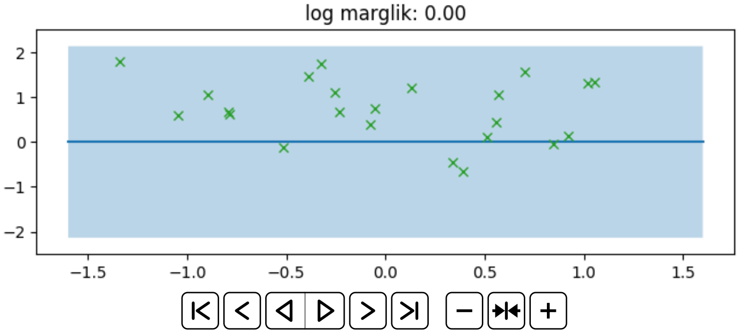
Marginal likelihood computation in action



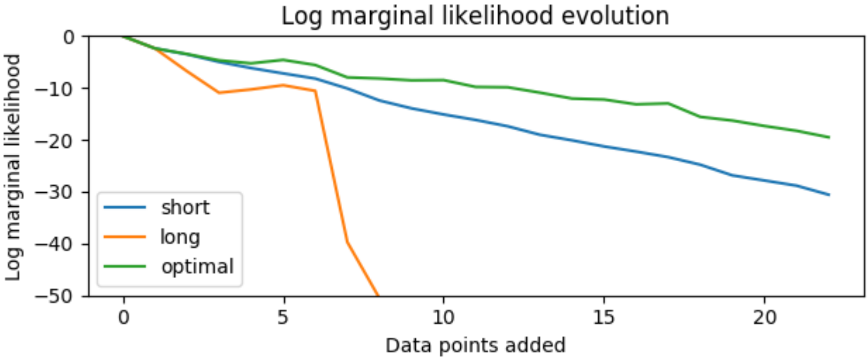
Marginal likelihood computation in action



Marginal likelihood computation in action

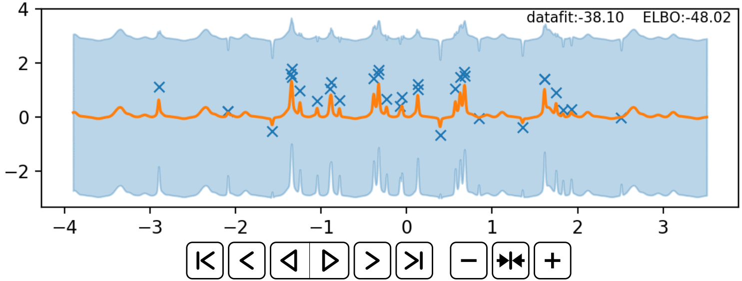


Marginal likelihood evolution



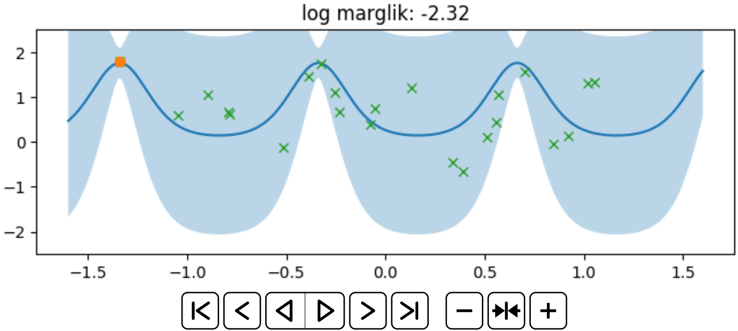
- **Short** lengthscale consistently **over-estimates variance**, so **can't get a high density** even with the observation in the error bars
- **Long lengthscale** consistently **under-estimates variance**, so gets a low density because the **observations are outside error bars**
- Optimal lengthscale **trades off** these behaviours... well.

Marginal likelihood in action



- We chose the prior: $f(\mathbf{x}) = \theta_s f_{\text{smooth}}(\mathbf{x}) + \theta_p f_{\text{periodic}}(\mathbf{x})$, with smooth and periodic GP priors respectively.
- Marginal likelihood learns **how** to generalise not just to fit the data.
- Amount of periodicity vs smoothness is automatically chosen by selecting hyperparameters θ_s, θ_p.

Marginal likelihood in action



Marginal likelihood as a prior probability

A complementary view

- Marginal likelihood is the probability of the data under the prior.

$$p(\mathbf{y}|\theta, X) = \int p(\mathbf{y} \mid f(X), \theta) p(f \mid \theta) \mathrm{d}f \tag{5}$$

- For zero-mean GP regression models it has the explicit form:

$$\begin{aligned} \log p(\mathbf{y}|\theta, X) &= \log \mathcal{N}\left(\mathbf{y}; 0, \mathbf{K} + \sigma^2 \mathbf{I}\right) \\ &= -\frac{N}{2} \log 2\pi - \underbrace{\frac{1}{2} \log\left|\mathbf{K} + \sigma^2 \mathbf{I}\right|}_{\text{Complexity penalty}} - \underbrace{\frac{1}{2} \mathbf{y}^\top \left(\mathbf{K} + \sigma^2 \mathbf{I}\right)^{-1} \mathbf{y}}_{\text{Data fit}} \end{aligned} \tag{6}$$

- Laplace approximations in Neural Networks look similar
- Pretty amazing that you can estimate updating behaviour from the shape of the loss function (ELBOs give lower bound!)

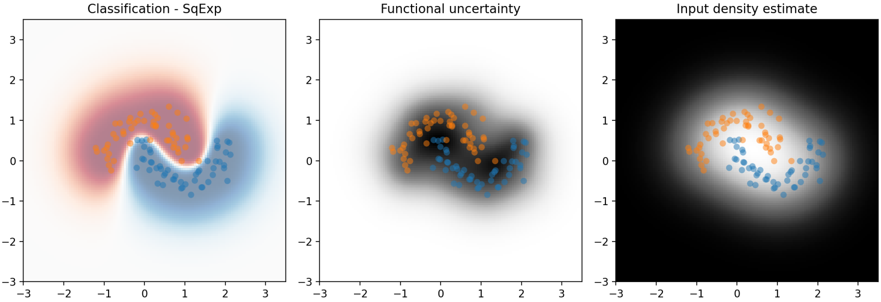
Intermediate take-homes

- Uncertainty and inductive bias interact! Prior is super important to getting the right behaviour in uncertainty
- Can't get strong generalisation without low uncertainty
- Marginal likelihood measures incremental predictive performance
- No need for hyperpriors to get good model selection!
- Is the marginal likelihood safe from overfitting?
  $\implies$ It's safe from the kind of overfitting that the normal likelihood exhibits

Should we be uncertain far from the data?

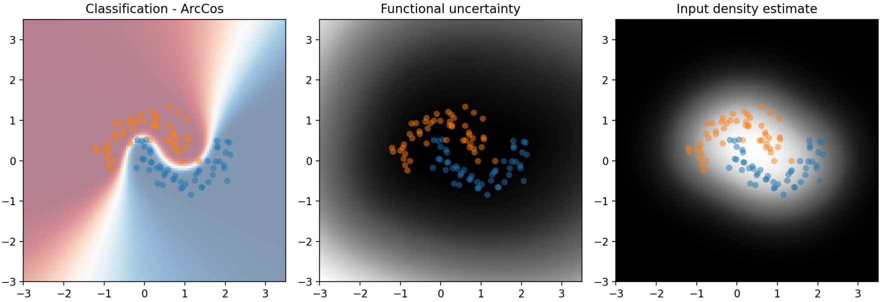
Can we use input density
as a metric for predictive uncertainty?

GPs as a Gold Standard for BNNs



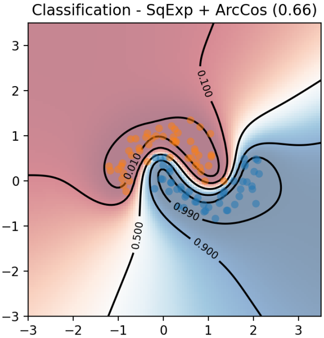
- GPs considered the "gold standard" model for uncertainty estimation.
- Often in Bayesian Deep Learning, aim is to replicate GP properties in DNNs.
- Though implicitly, a GP with a *Squared Exponential* kernel.

GPs as a Gold Standard for BNNs



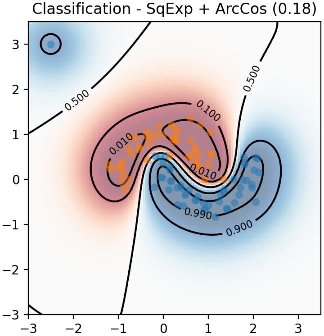
- ArcCos kernel is obtained from infinite limit of ReLU NN.
- Still exact inference in a GP. Different inductive bias!
- So what is the right one? What behaviour should BNNs copy?
- Both extrapolations are reasonable.

"Correct" extrapolation with model selection



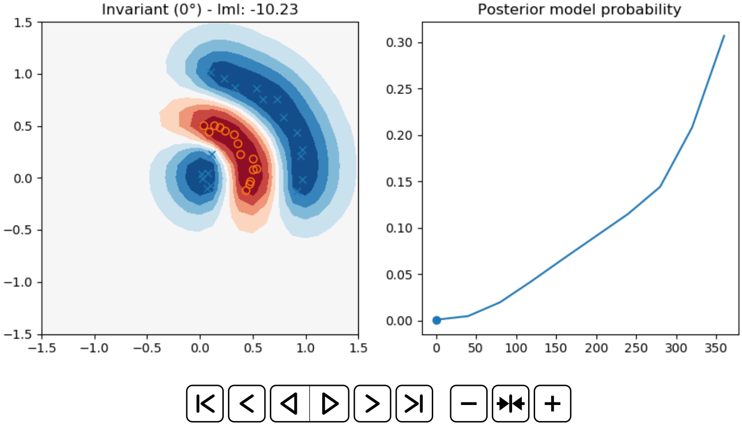
- Marginal likelihood uses appreciable ArcCos component
- What if it's wrong?
- Terrible predictive log likelihood if we're wrong about extrapolation!

Telling the model it's wrong



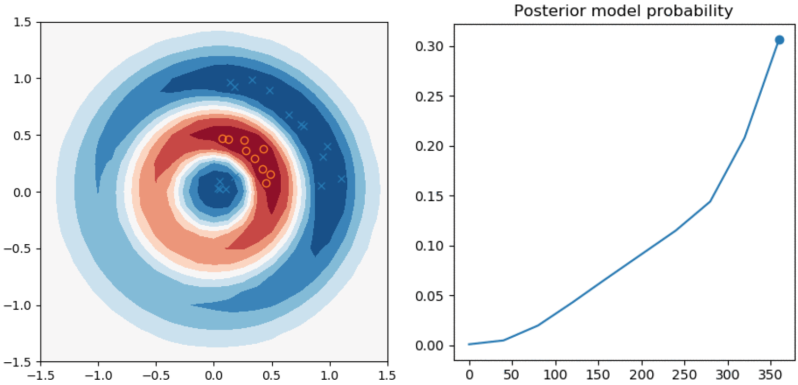
- Single datapoint is enough to change inductive bias.
- How realistic is the train/test split assumption?
- Should we give models a chance to learn under distribution shift?
- We could measure how quickly they adapt?
- Little data can be very informative for OOD / causality

Invariance and Uncertainty



- Another example of strong extrapolation.
- Marginal likelihood prefers really strong predictions

Invariance and Uncertainty: Another solution



- Average over hyperparameters as well!
- More cautious predictions.

$$p(y^*|\mathcal{D}) = \int p(y^*|f)p(f|\theta,\mathcal{D})p(\theta|\mathcal{D})\mathrm{d}f\mathrm{d}\theta \tag{7}$$

Intermediate take-homes

- Extrapolation behaviour can be very desirable
- This is at odds with being uncertain "far from the data"
- Opinion: We should not rely on input density for uncertainty
- Overconfidence can be fixed with additional observations
- More Bayes also helps :-)

Discussion points

- Can we use input density for uncertainty estimation?
- Should we be assessing uncertainty as part of a continual learning process? Is it fair to force our models not to learn on the job?
- Causality is often hard because of a lack of data (coloured MNIST). Single example can break a hypothesis used for generalisation!
- How should we implement this behaviour? Bayes? Neural Processes? Meta-learning? Is Bayesian reasoning helpful with this?