Marginal likelihood

Mark van der Wilk

Department of Computing
Imperial College London

@markvanderwilk
m.vdwilk@imperial.ac.uk

February 3, 2023

Learning objectives

Previously we saw how the Bayesian framework tells us how to infer unseen parameters. Here we ask **why** it works.
We seek to answer:

- What happens if we minimise the error to the training data?
- Does uncertainty prevent overfitting? If so, how?
- Why does the marginal likelihood prevent overfitting?
- What does the marginal likelihood measure?

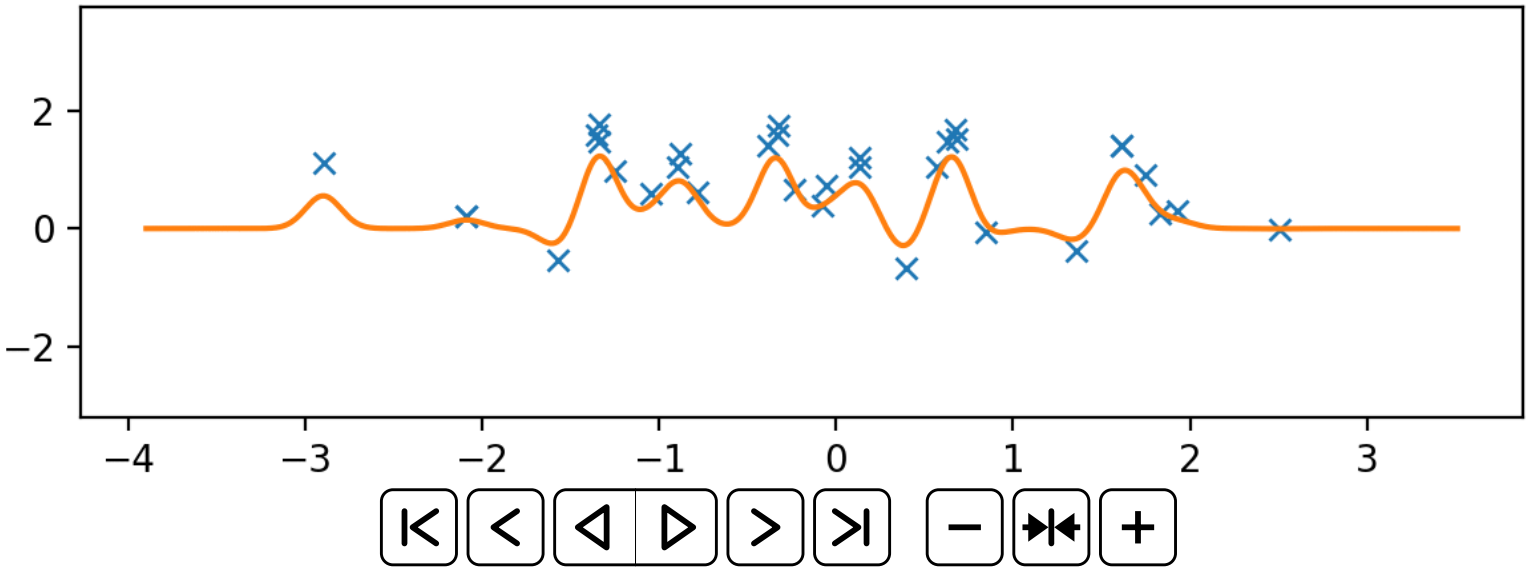
Use e.g. Adobe Acrobat to view animations.

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.

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

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

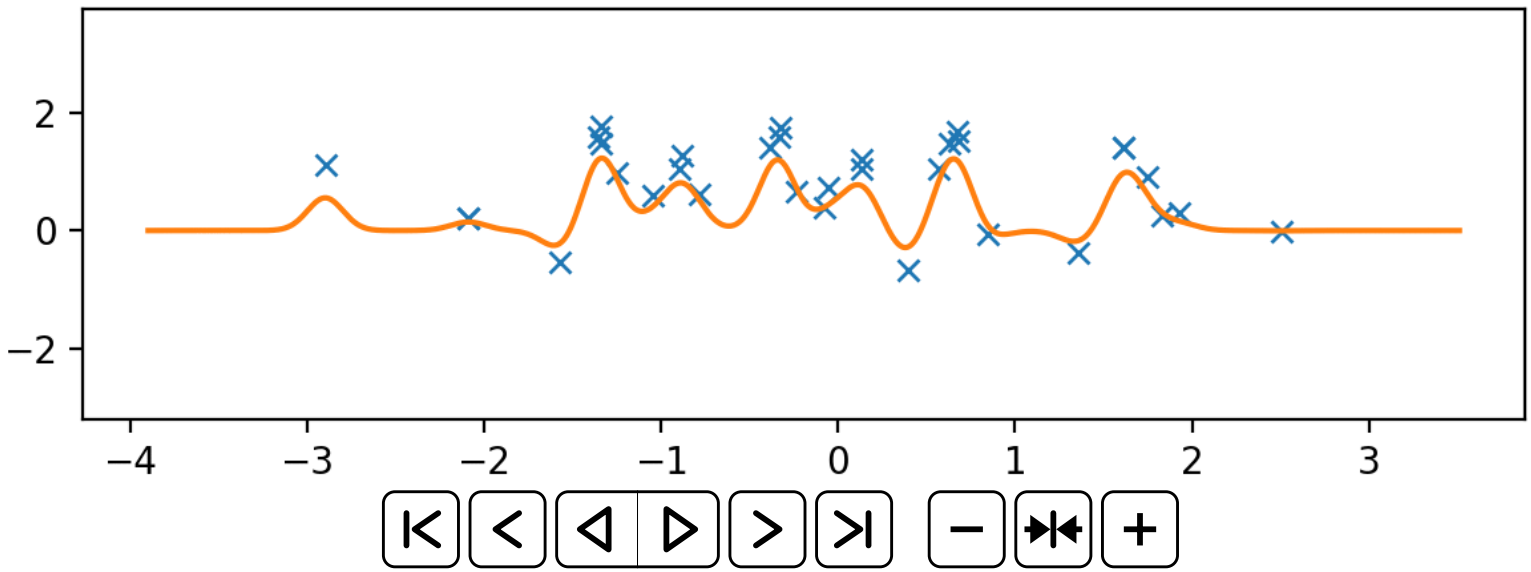


Minimising training loss

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

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

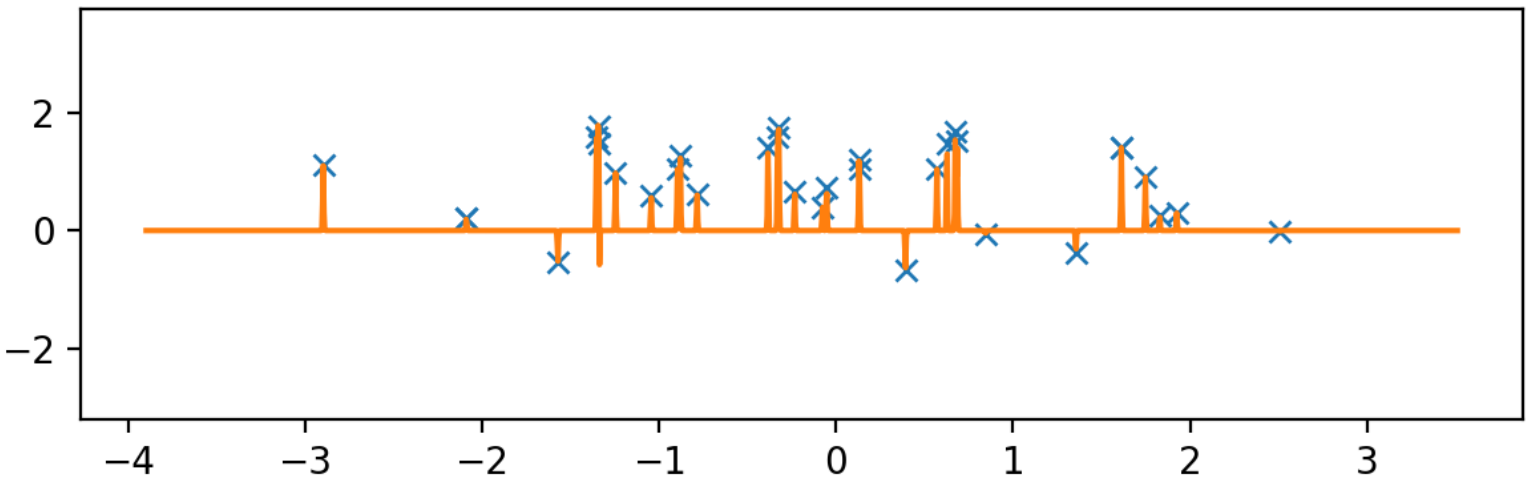
$$\{\theta^*, \sigma^*\} = \underset{\theta, \sigma}{\operatorname{argmin}}\, \mathcal{L}(\theta, \sigma) \tag{2}$$



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

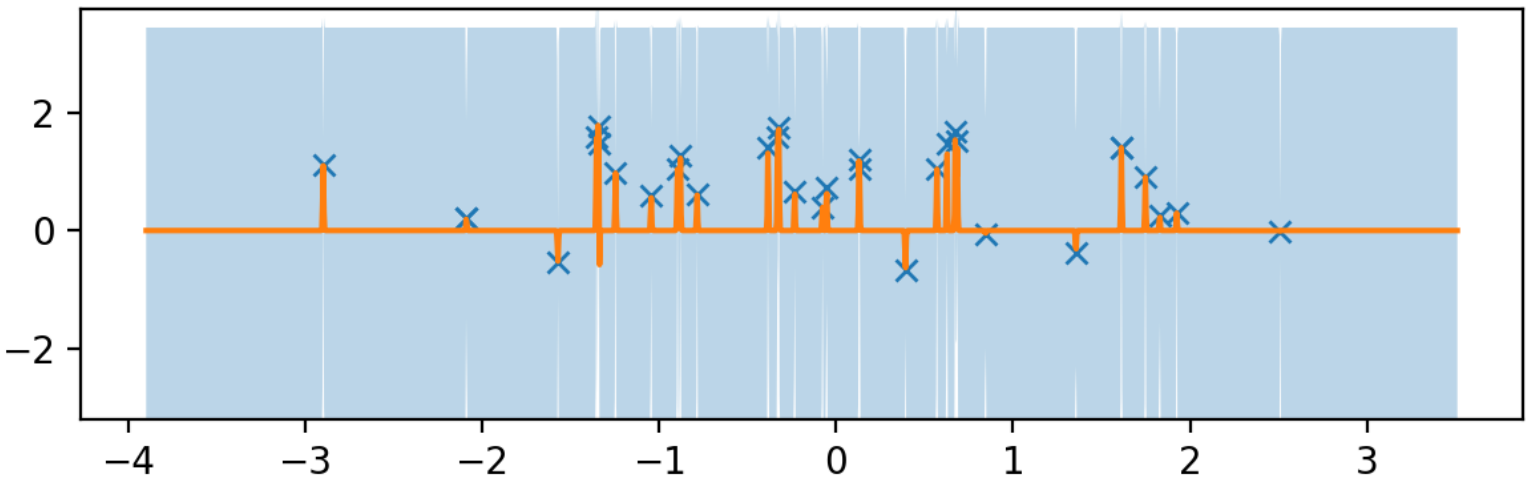
How does uncertainty help?

Does uncertainty help against the overfitting?



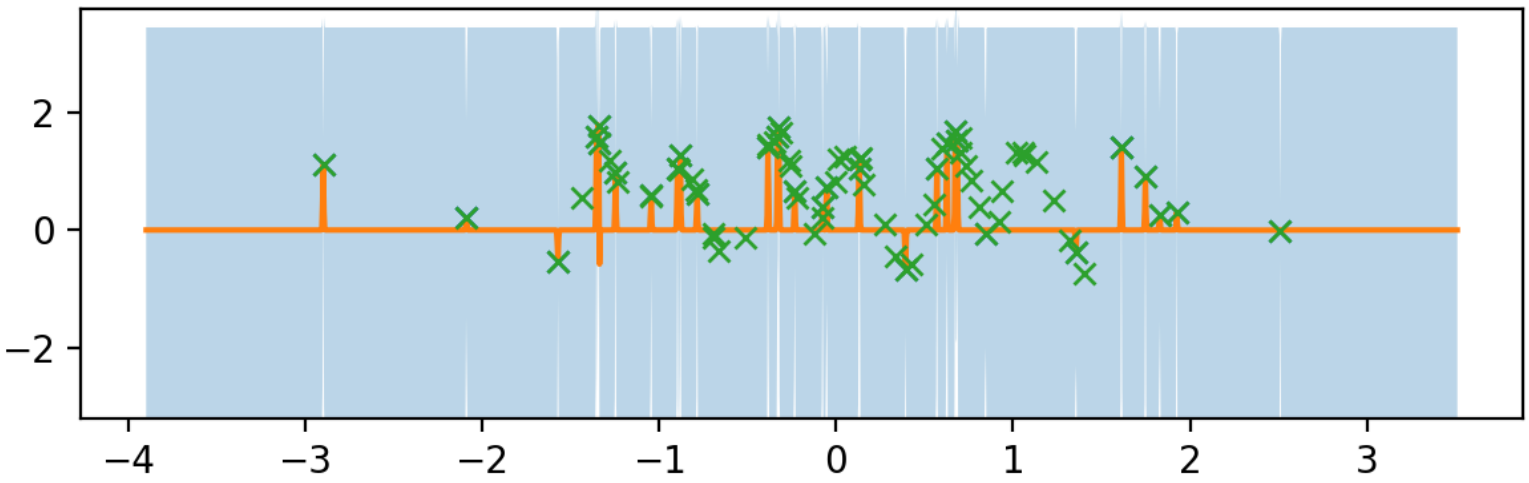
How does uncertainty help?

Does uncertainty help against the overfitting?



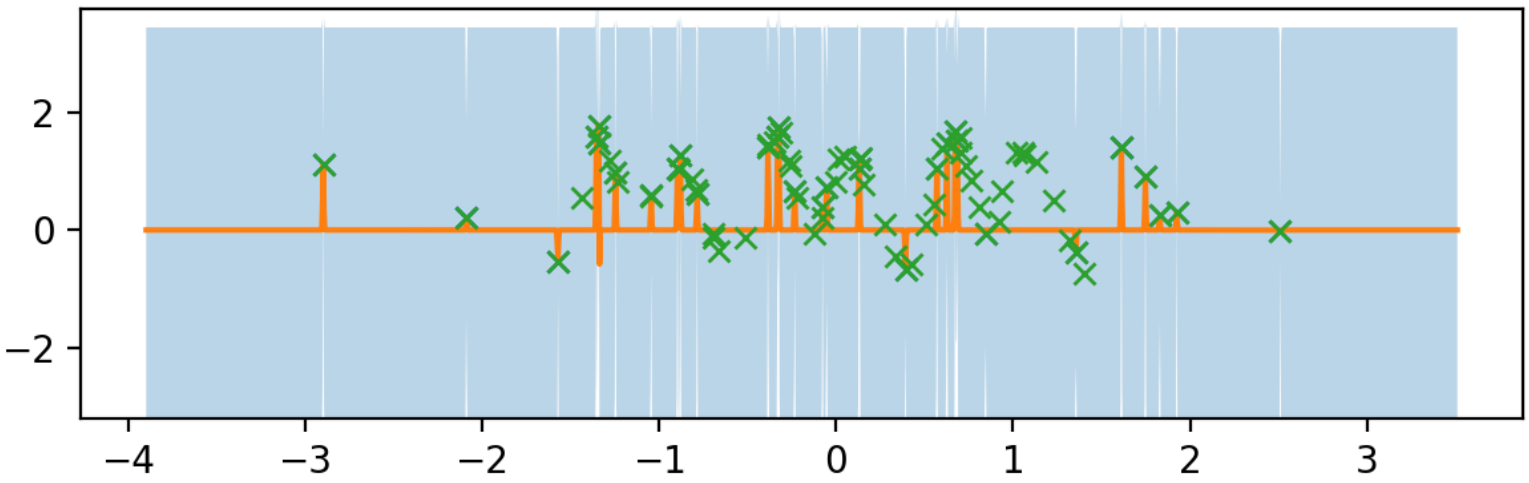
How does uncertainty help?

Does uncertainty help against the overfitting?



How does uncertainty help?

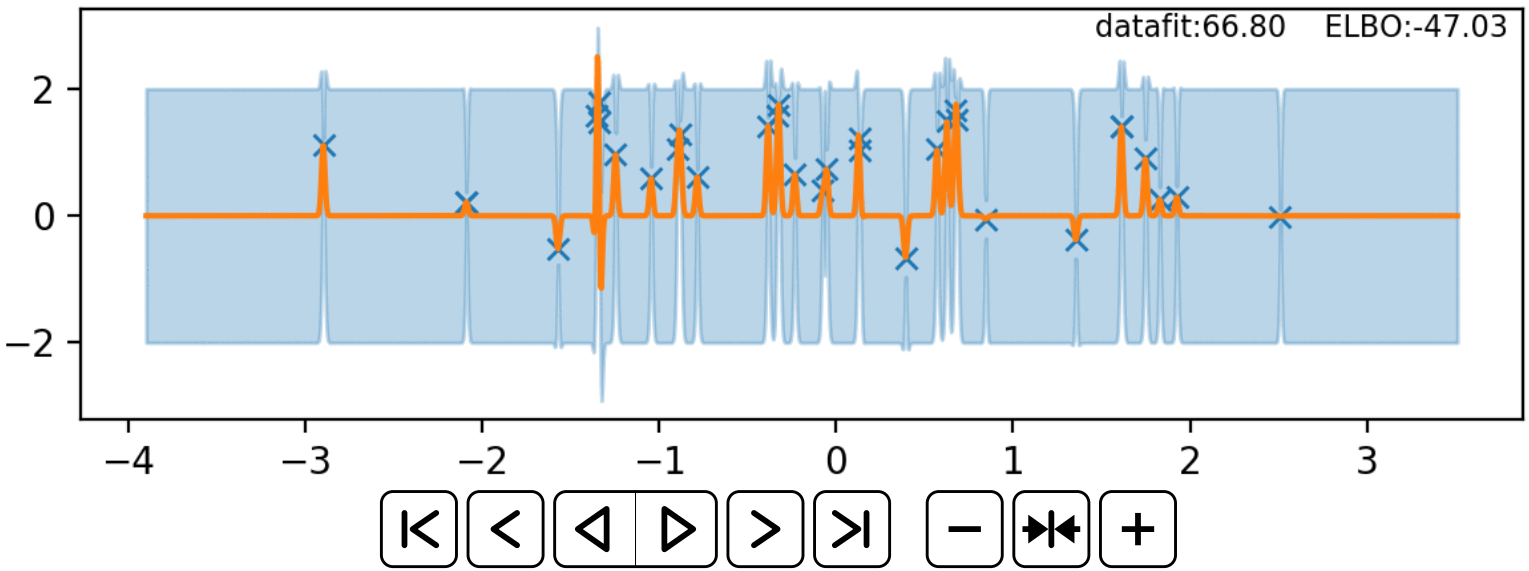
Does uncertainty help against the overfitting?



- Uncertainty by itself does not necessarily make predictions better, if the wrong model is chosen
- Uncertainty does make predictions more cautious, which can be very useful!

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**:

$$p(\mathbf{y} \mid \theta, X) = p(y_1|\theta, \mathbf{x}_1)p(y_2|\theta, \mathbf{x}_1, y_1, \mathbf{x}_2)p(y_3|\theta, \{\mathbf{x}_i, y_i\}_{i=1}^{2}, \mathbf{x}_3)\ldots \tag{3}$$

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

Marginal likelihood as incremental prediction

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

$$p(\mathbf{y} \mid \theta, X) = p(y_1|\theta, \mathbf{x}_1)p(y_2|\theta, \mathbf{x}_1, y_1, \mathbf{x}_2)p(y_3|\theta, \{\mathbf{x}_i, y_i\}_{i=1}^{2}, \mathbf{x}_3)\ldots \quad (3)$$

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

Remember

$$p(y_n|\theta, \{\mathbf{x}_i, y_i\}_{i=1}^{n-1}, \mathbf{x}_n) = \int p(y_n|f(\mathbf{x}_n))p(f(\mathbf{x}_n) \mid \{\mathbf{x}_i, y_i\}_{i=1}^{n-1}, \mathbf{x}_n)\mathrm{d}f(\mathbf{x}_n)$$

i.e. the predictive distribution of y_n based on the posterior given all points up to $n-1$.

Marginal likelihood as incremental prediction

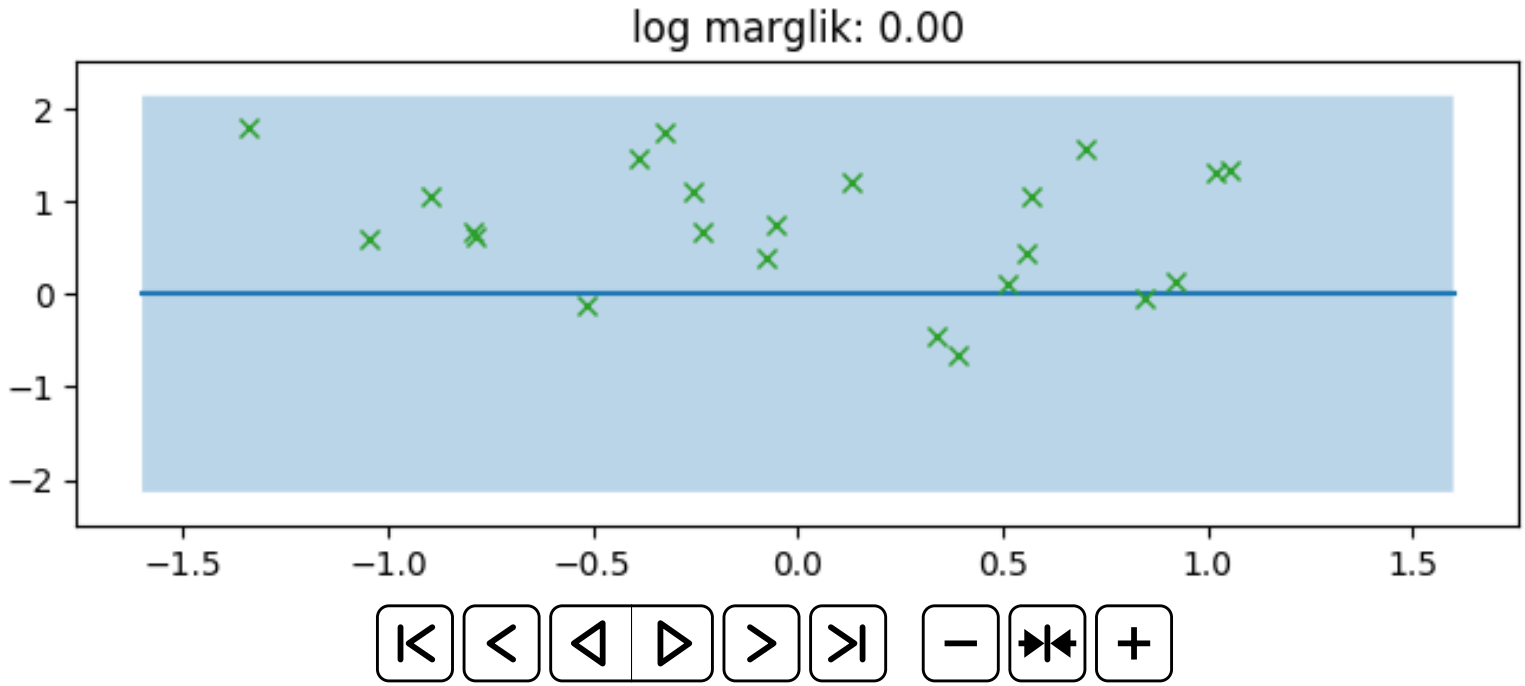
We can split the marginal likelihood up using the **product rule**:

$$p(\mathbf{y} \mid \theta, X) = p(y_1|\theta, \mathbf{x}_1)p(y_2|\theta, \mathbf{x}_1, y_1, \mathbf{x}_2)p(y_3|\theta, \{\mathbf{x}_i, y_i\}_{i=1}^{2}, \mathbf{x}_3)\ldots \quad (5)$$

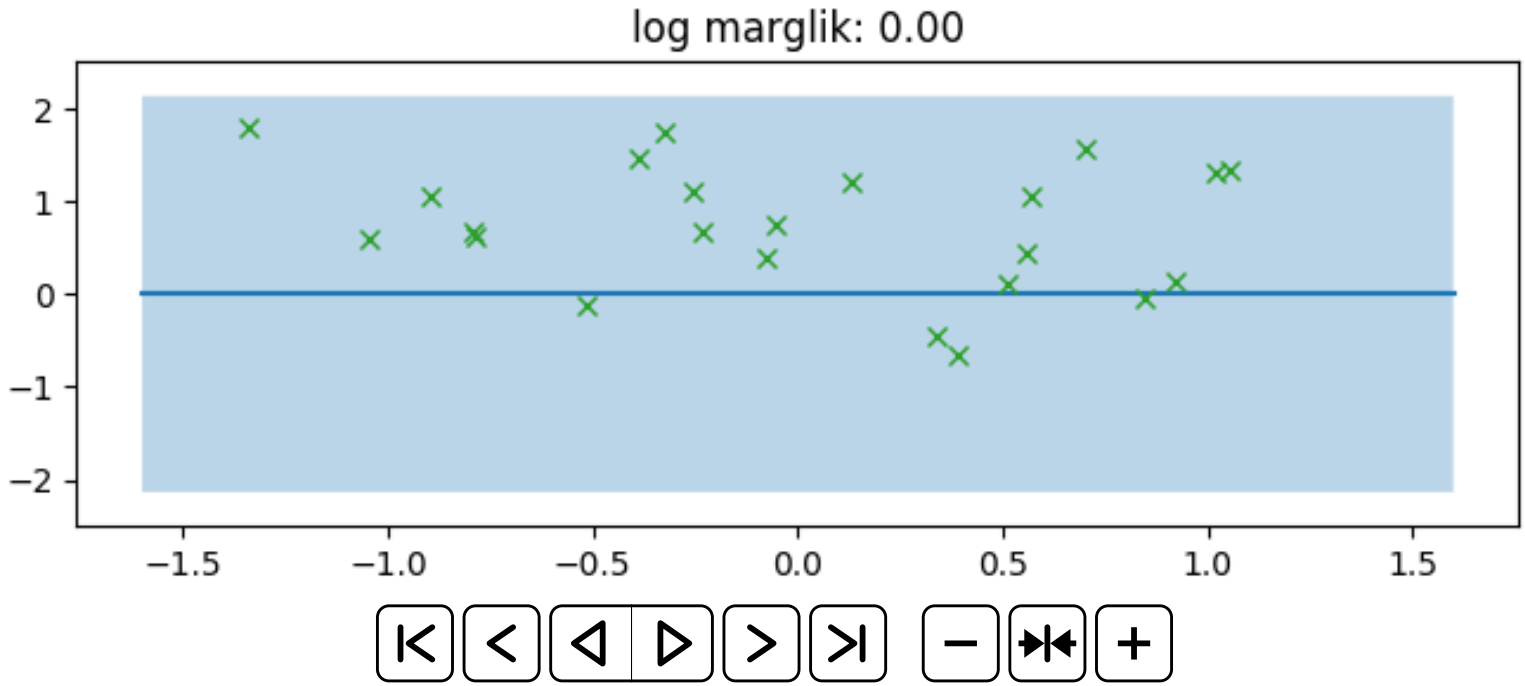
$$= \prod_{n=1}^{N} p(y_n|\theta, \{\mathbf{x}_i, y_i\}_{i=1}^{n-1}, \mathbf{x}_n) \quad (6)$$

- 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

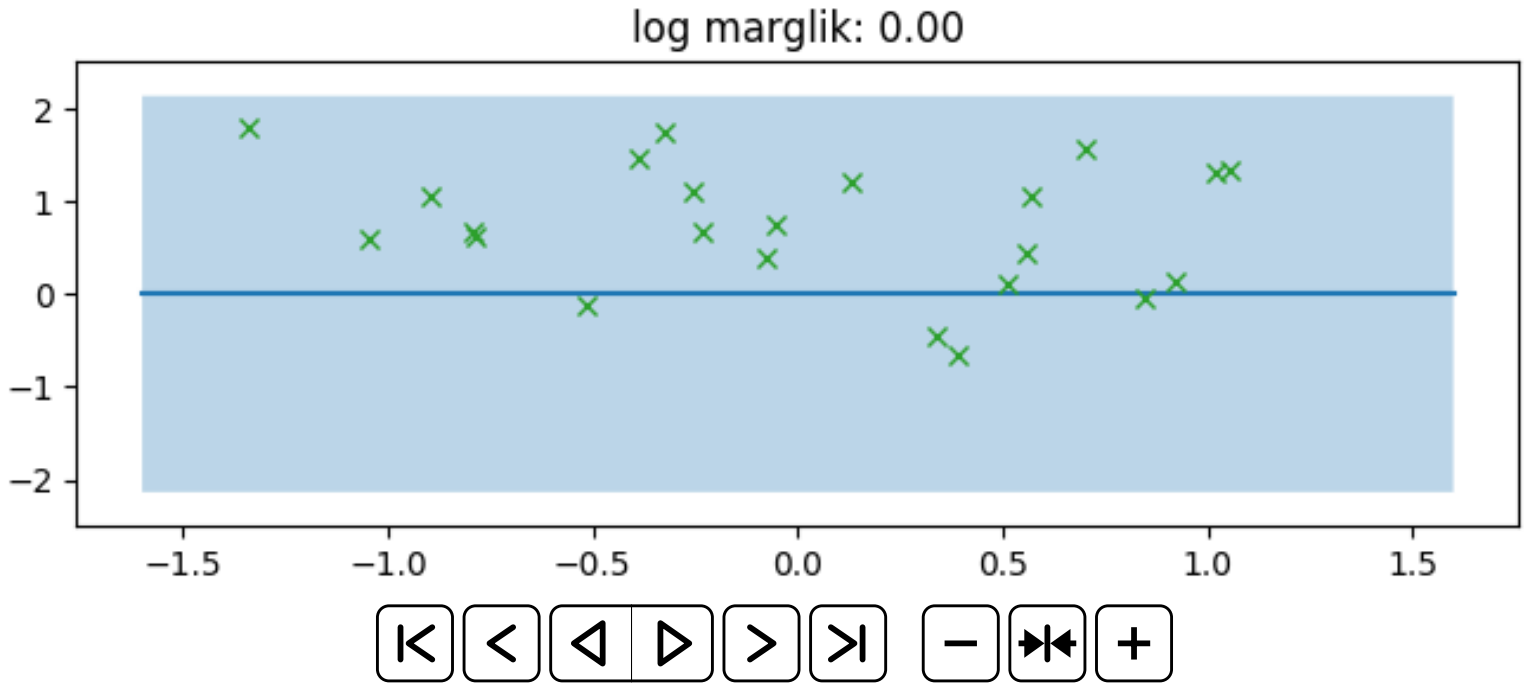
Marginal likelihood computation in action



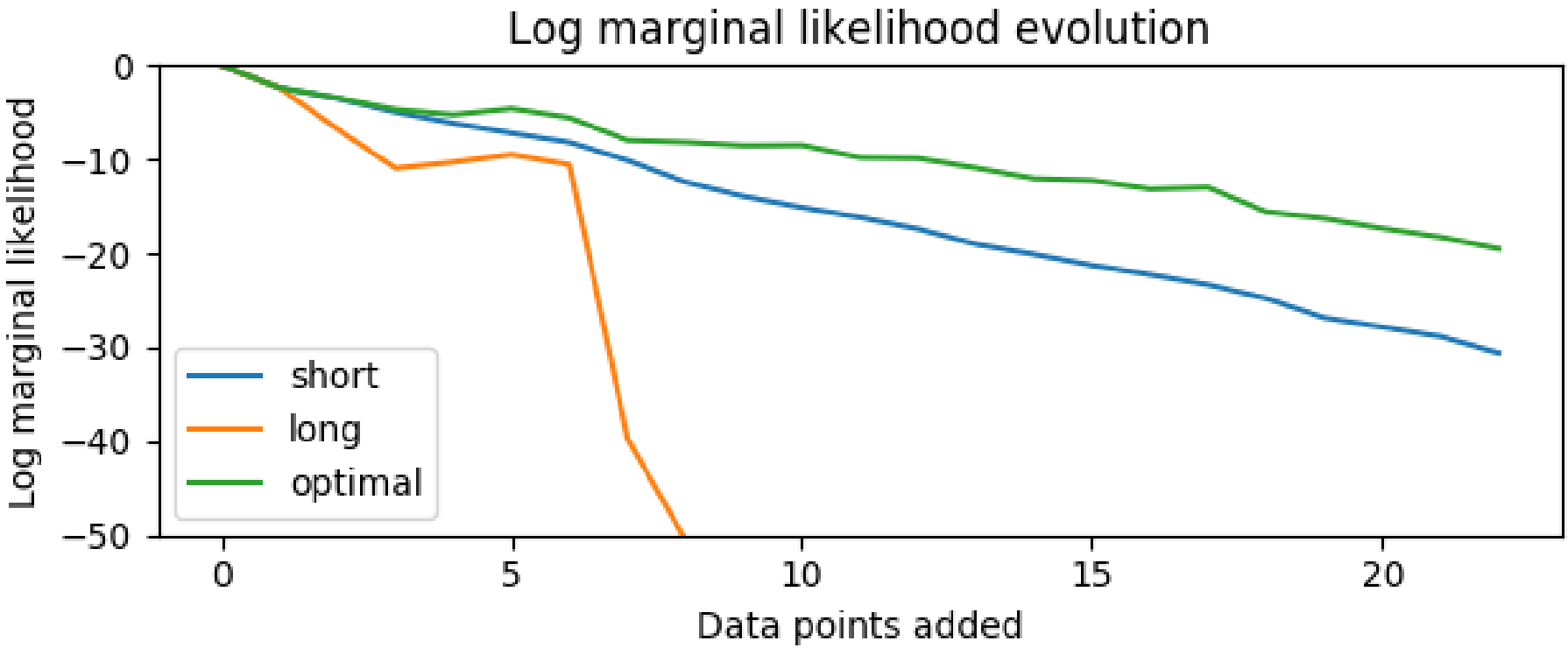
Marginal likelihood computation in action



Marginal likelihood computation in action

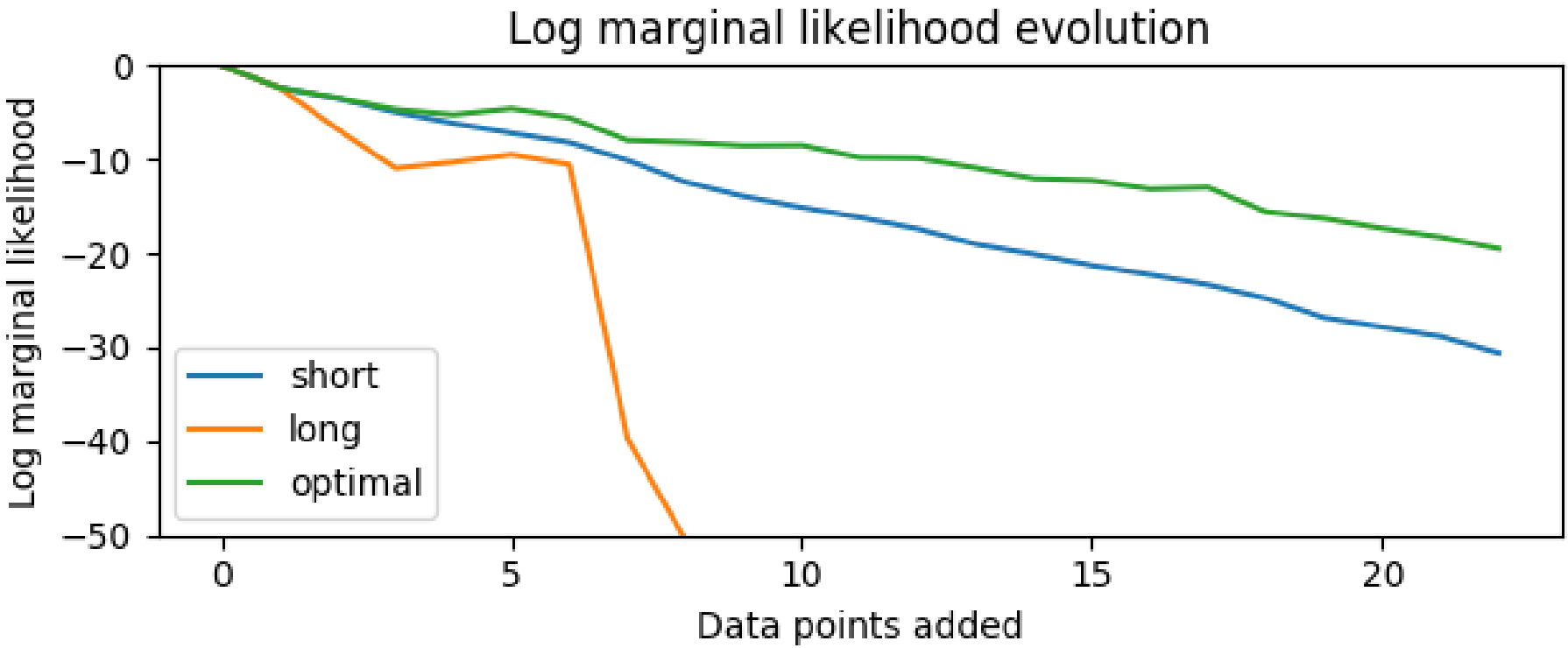


Marginal likelihood evolution



- **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...

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.

Generalisation

- A model with a high marginal likelihood is likely to **generalise** well.
- Its inductive bias has correctly predicted the next training point throughout the entire training set.
- Marginal likelihoods are also related to **generalisation error bounds**.

Generalisation

- A model with a high marginal likelihood is likely to **generalise** well.
- Its inductive bias has correctly predicted the next training point throughout the entire training set.
- Marginal likelihoods are also related to **generalisation error bounds**.

Generalisation error bounds state things like: "With high probability, the error for method X on a test set will not be larger than Y"

PAC-Bayesian Theory Meets Bayesian Inference [1]

Marginal likelihood as a prior probability

A complementary view

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

$$p(\mathbf{y}|\theta, X) = \int p(\mathbf{y} \mid f(X), \theta) p(f(X) \mid \theta) \mathrm{d}f(X) \tag{7}$$

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

$$\begin{aligned} \log p(\mathbf{y}|\theta, X) &= \log \mathcal{N}\left(\mathbf{y}; 0, \mathbf{K} + \sigma^2 \mathbf{I}\right) \\ &= -\frac{N}{2} \log 2\pi - \underbrace{\frac{1}{2} \log \left|\mathbf{K} + \sigma^2 \mathbf{I}\right|}_{\text{Complexity penalty}} - \underbrace{\frac{1}{2} \mathbf{y}^\intercal \left(\mathbf{K} + \sigma^2 \mathbf{I}\right)^{-1} \mathbf{y}}_{\text{Data fit}} \end{aligned} \tag{8}$$

Complexity penalty and data fit

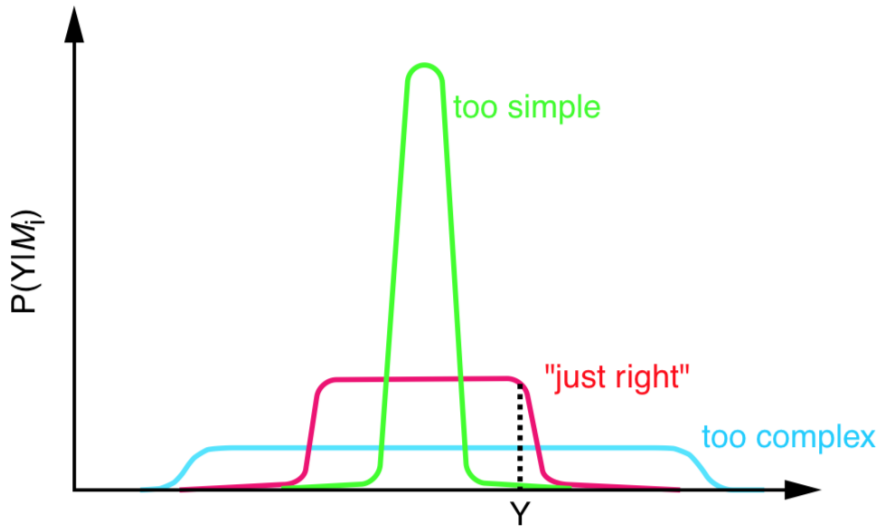
$$\log p(\mathbf{y}|\theta, X) = -\frac{N}{2}\log 2\pi - \underbrace{\frac{1}{2}\log\left|\mathbf{K} + \sigma^2\mathbf{I}\right|}_{\text{Complexity penalty}} - \underbrace{\frac{1}{2}\mathbf{y}^\top\left(\mathbf{K} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{y}}_{\text{Data fit}} \quad (9)$$

- Determinant is product of eigenvalues (variances) of the covariance matrix – the volume of the prior
- Quadratic term measures whether the observation $\mathbf{y}$ is within the variation allowed by the prior – by lining $\mathbf{y}$ up with the eigenvectors of the covariance

Simple and complex models

Probabilities have to normalise to 1, so a model **cannot** both

- be flexible enough to fit many datasets, and
- make specific predictions after only a small amount of data.



From "Occam's Razor" – Rasmussen & Ghahramani (2000)

- Complex / flexible models spread their probability over many possible explanations of the data

Occam's razor

"Entities are not to be multiplied without necessity"

or

The simplest solution is most likely the right one

Occam's razor

"Entities are not to be multiplied without necessity"

or

The simplest solution is most likely the right one

The marginal likelihood prefers the simplest model that still fits the data.

Occam's razor

"Entities are not to be multiplied without necessity"

or

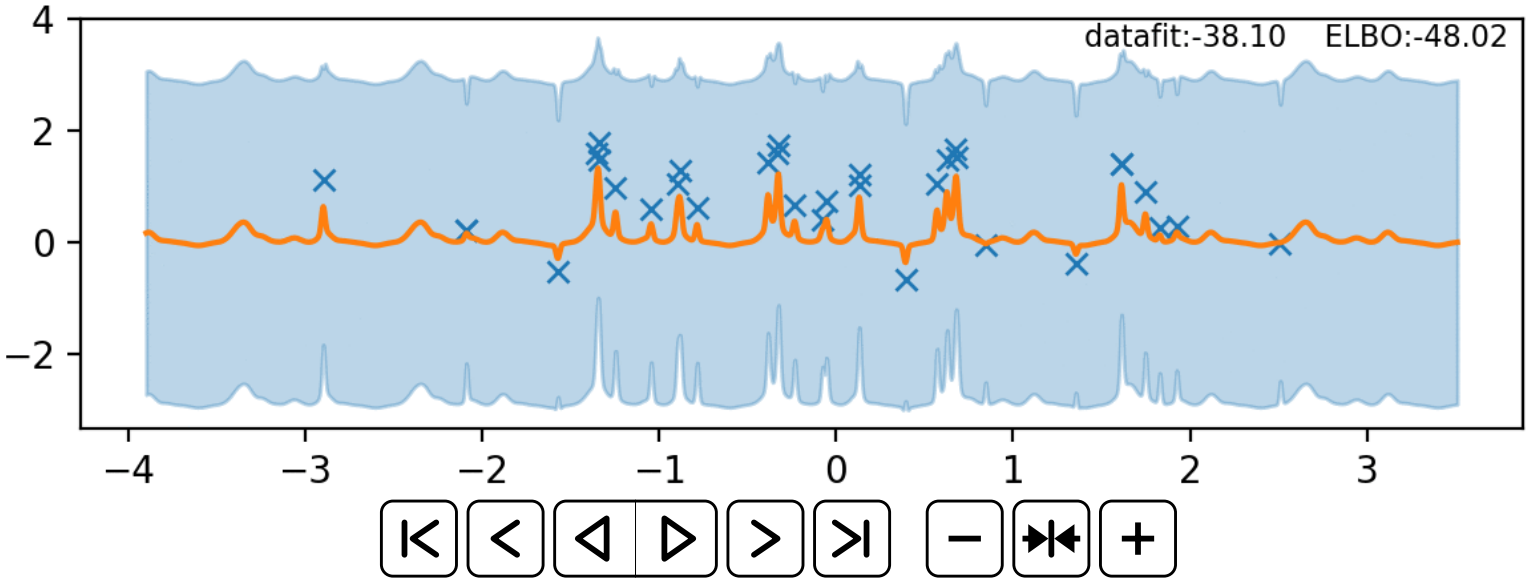
The simplest solution is most likely the right one

The marginal likelihood prefers the simplest model that still fits the data.

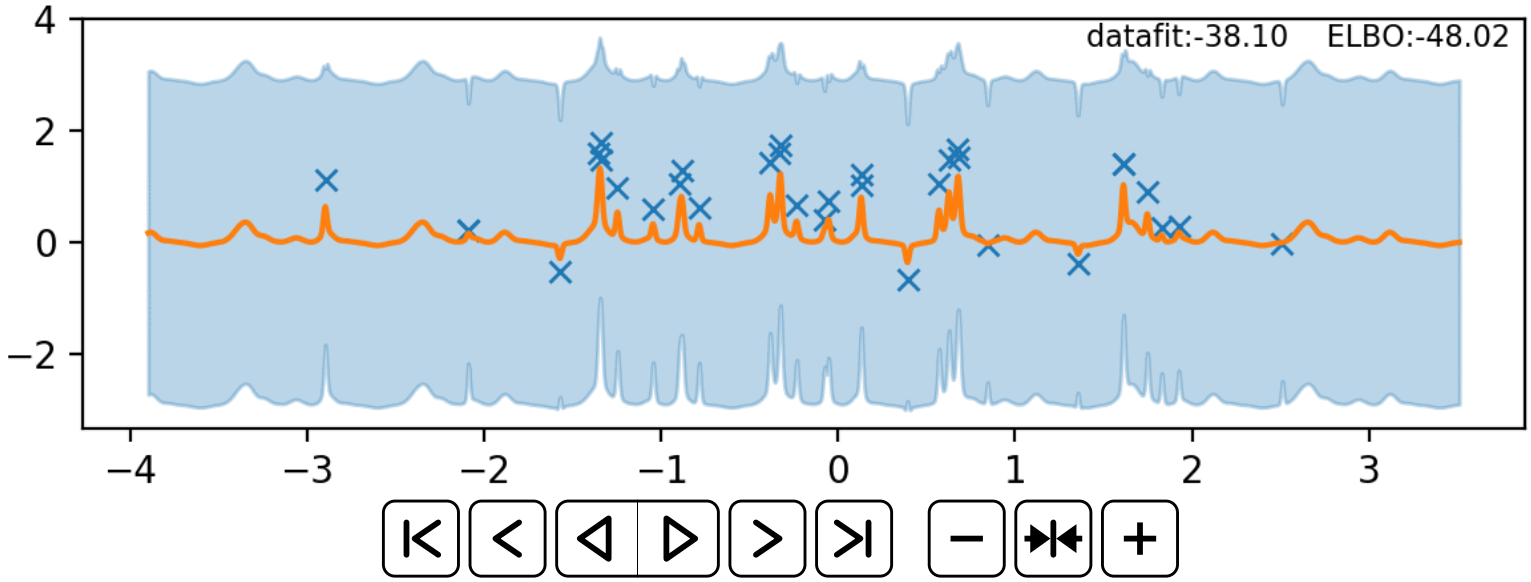
The marginal likelihood

- automatically penalises complex models, as the old adage states
- comes from a principle as simple as representing your belief using probability
- is automatically applied if you use Bayes' rule properly

Marginal likelihood in action

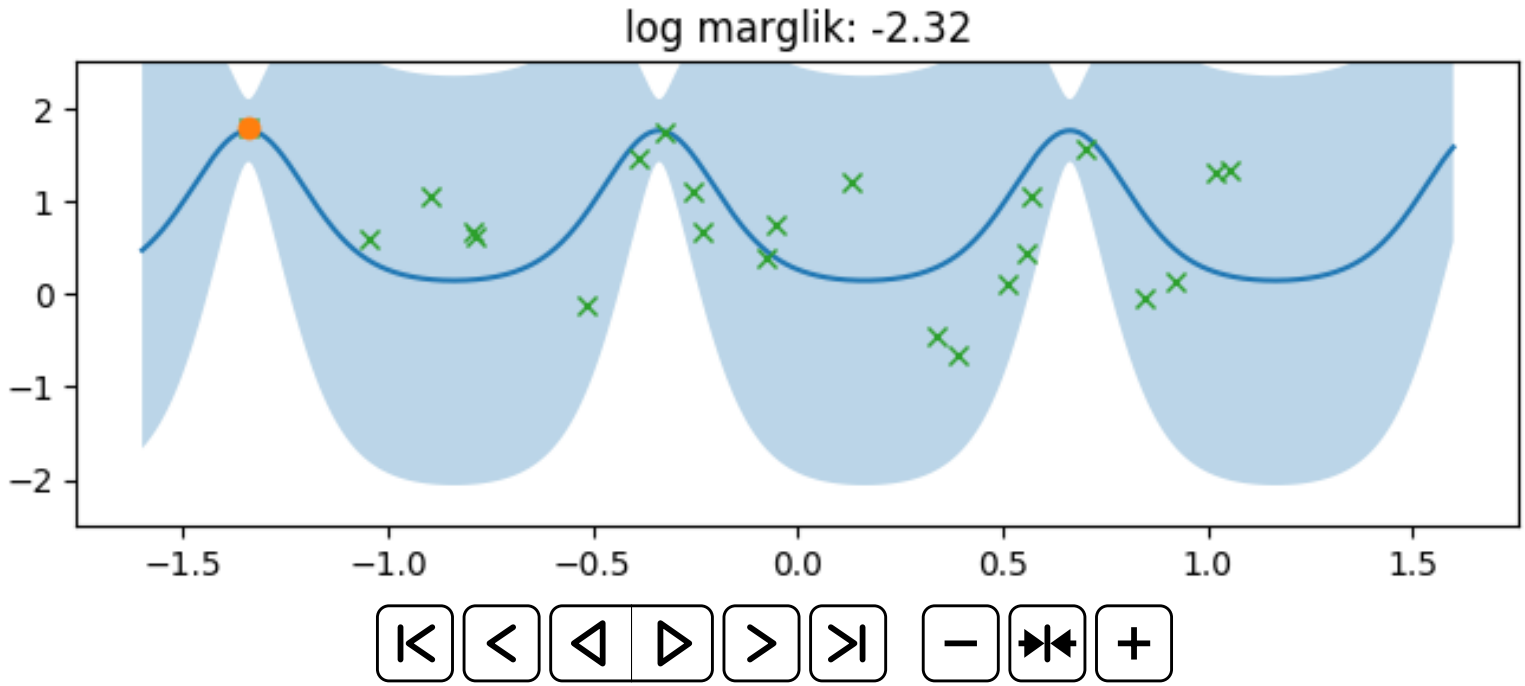


Marginal likelihood in action



- Marginal likelihood learns **how** to generalise not just to fit the data.
- 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.
- Amount of periodicity vs smoothness is automatically chosen by selecting hyperparameters θ_s, θ_p.

Marginal likelihood in action



Further reading

- David J.C. MacKay. *Information Theory, Inference, and Learning Algorithms*, chapter 28.

References I

[1] P. Germain, F. Bach, A. Lacoste, and S. Lacoste-Julien. Pac-bayesian theory meets bayesian inference. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 1884–1892. Curran Associates, Inc., 2016.

[2] D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.