4F13 Probabilistic Machine Learning: Coursework 1: Gaussian Process

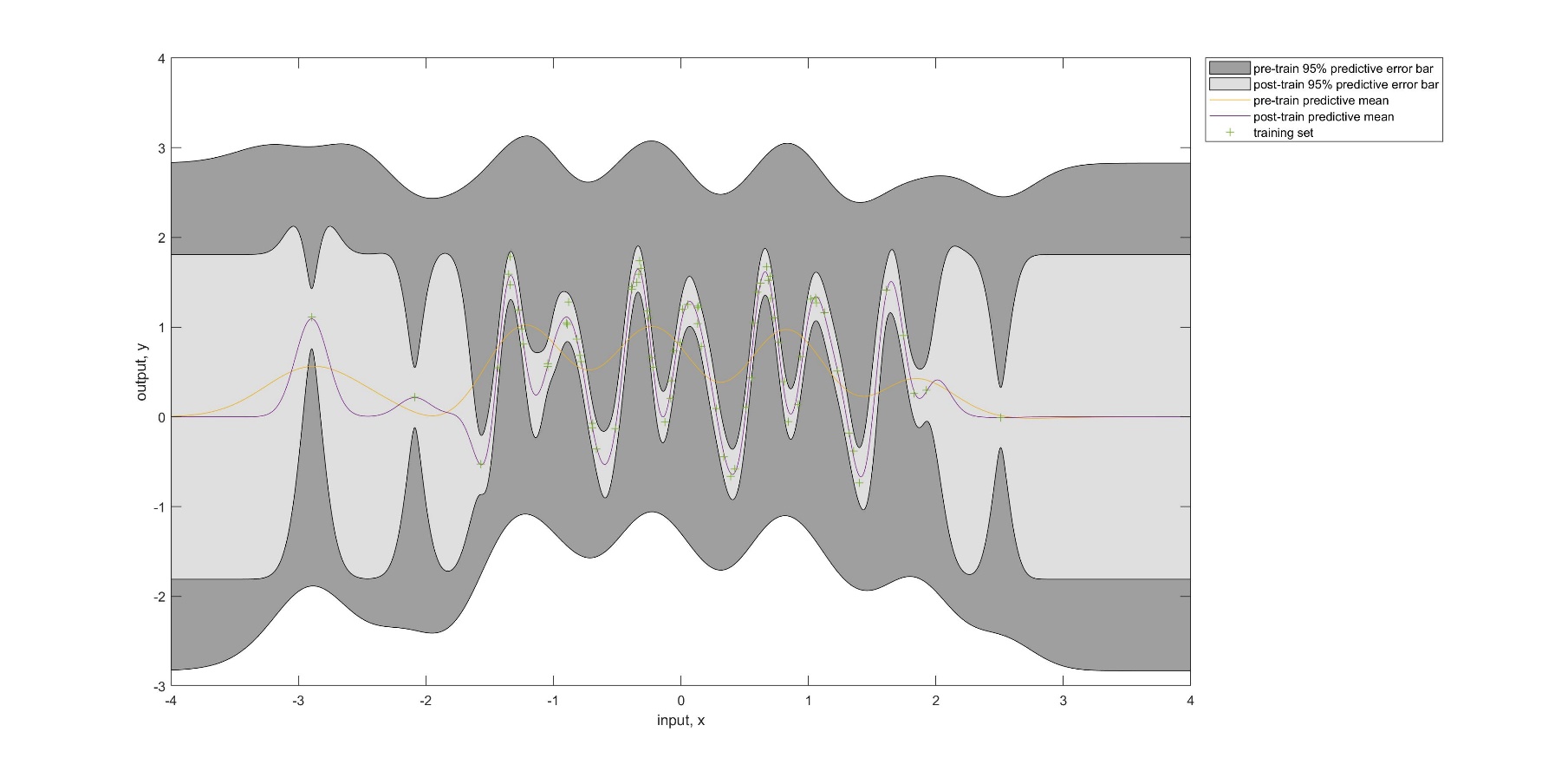


Figure 1 shows the predictive means and the predictive error bars of the squared exponential model before and after training.

Table 1 shows the values of negative log marginal likelihood and hyperparameters before and after the minimization.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Negative Log Marginal Likelihood () | Length-scale () | Signal Standard  Deviation () | Noise Standard Deviation () |
| Pre-train | 92.91 | 0.37 | 1.00 | 1.00 |
| Post-train | 11.90 | 0.13 | 0.90 | 0.12 |

From Figure 1, the area of the shaded region for the predictive error bar is reduced after the Negative Log Marginal Likelihood is minimized from 92.91 to 11.90. The region (i.e. [-2, 2]) with more training data has a narrower error bar, since the prediction is more certain. The characteristic length-scale, signal standard deviation and noise standard deviation are also optimized as shown in Table 1.



Table 2 compares how different hyperparameters can lead to different trained models.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Initialisation Hyperparameters | | Negative Log Marginal Likelihood () | Length-scale () | Signal Standard  Deviation () | Noise Standard Deviation () |
| hyp.cov | hyp.lik |
| [-1 , 0] | 0 | 11.90 | 0.13 | 0.90 | 0.12 |
| [-0.25, 0] | 0 | 78.22 | 8.04 | 0.70 | 0.66 |
| [-10 , 0] | 0 | 106.35 | 0.00 | 0.71 | 0.71 |
| [ 10 , 0] | 0 | 106.35 | 0.00 | 0.71 | 0.71 |
| [-0.25, -5] | 0 | 78.22 | 8.04 | 0.70 | 0.66 |
| [-10, 5] | 0 | 106.35 | 0.00 | 0.71 | 0.71 |
| [10, -5] | 0 | 106.35 | 0.00 | 0.71 | 0.71 |
| [-0.25, 0] | 5 | 78.22 | 8.04 | 0.70 | 0.66 |
| [-10, 0] | -5 | 106.35 | 0.00 | 0.71 | 0.71 |
| [10, 0] | -5 | 106.35 | 0.00 | 0.71 | 0.71 |

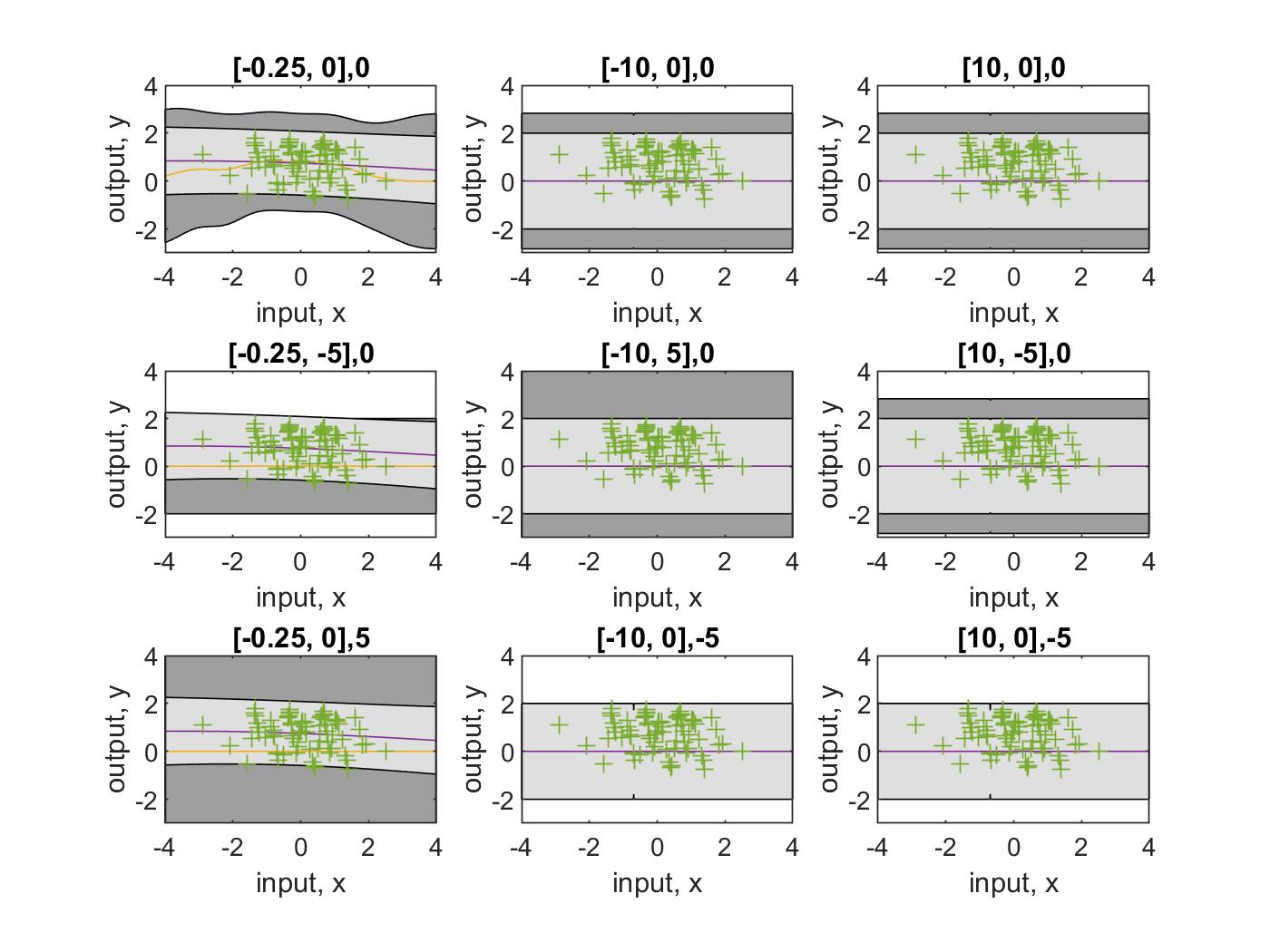


Figure 2 shows the predictive means and the predictive error bars of the squared exponential model with different covariance settings before and after training. Different covariance settings are shown in the subtitles.

By initialising the hyperparameters differently, we can see how they affect the minimization. In short, standard deviations of both noise and signal do not affect the result a lot, in comparison of length-scale. The ones initialized with the same length-scale value will always be minimized to the same nlml value. Therefore, it is sensible to only plot a graph of length-scale against negative log marginal likelihood, and to compare which initial setting of length scale gives the lowest nlml.

A close up of a white background

Description generated with high confidence

Figure 3 illustrates how nlml value changes as a function of ell (i.e. the log of the characteristic length-scale.)

The different local optimum of nlml is 78.22 at ell = -0.25. The first fit with length-scale of -1 is better since it has the least nlml and highest marginal likelihood.



Table 3 shows the values of negative log marginal likelihood and hyperparameters before and after the minimization.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Negative Log Marginal Likelihood () | Length-scale () | Signal Standard  Deviation () | Noise Standard Deviation () | Period () |
| Pre-train | 92.73 | 0.37 | 1.00 | 1.00 | 7.39 |
| Post-train | -1.60 | 0.27 | 0.85 | 0.12 | 3.00 |

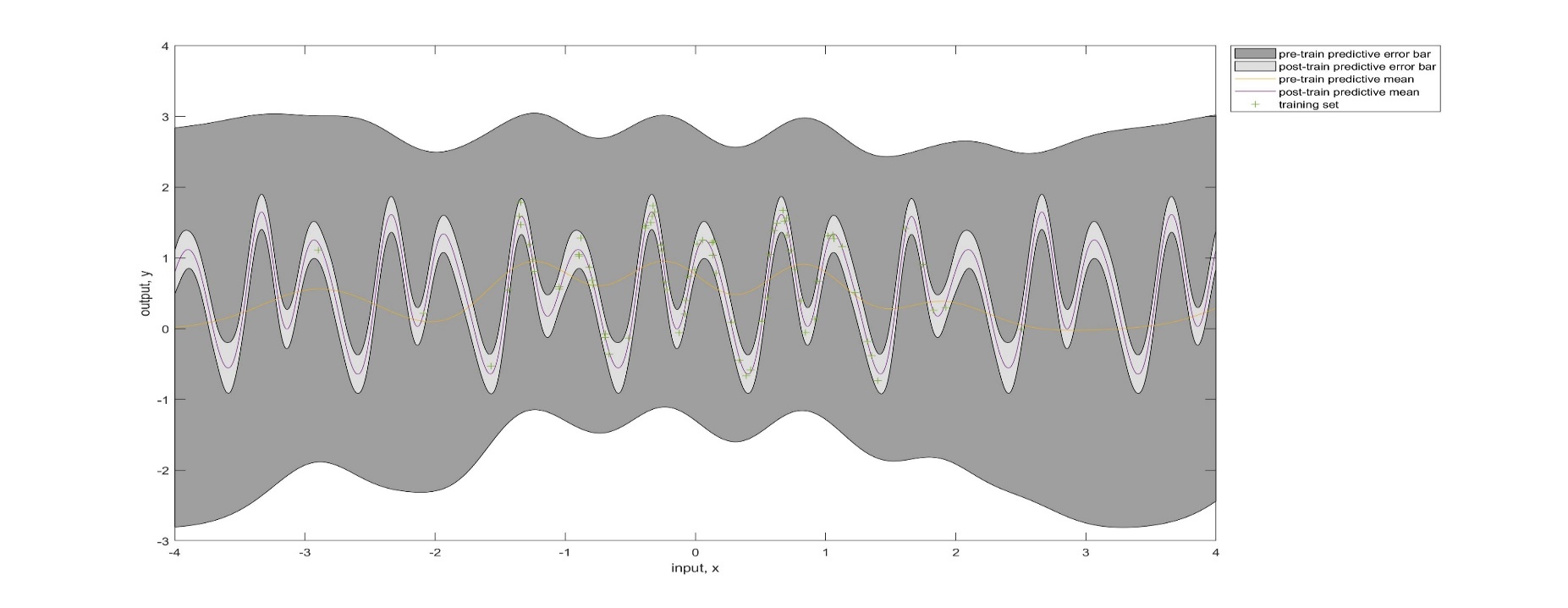


Figure 4 shows the predictive means and the predictive error bars of the periodic model before and after training.

In comparison to the fit from part a, the error bar has a narrower width even for the region with less training data. When comparing the marginal likelihood, this periodic model has better fit to the data since it has a higher marginal likelihood and has an extra hyperparameter which increases the capacity to fit the data.

I do think the data generating mechanism was not strictly periodic even though both plots showed periodic structures, especially for the interval [-1.5, 2]. It is possible that the data was generated with a covariance function that has several changepoints. Also, there is not much data existing outside the interval [-1.5, 2] so it does not have enough evidence to prove ‘the mechanism was strictly periodic’ outside the region mentioned.



A picture containing text

Description generated with high confidence

Figure 5 shows the predictive means and the predictive error bars of the product of periodic and squared exponential kernels model before and after training.

To apply the Cholesky decomposition, the covariance matrix must be symmetric and positive definite. Therefore, adding the diagonal matrix helps to retain the symmetry and change it to be positive definite.

Since the covariance function is a product of periodic and squared-exponential kernels, both the predictive mean and bars oscillate in a periodic exponential pattern. I can also see there are many datapoints positioning within the predictive error bars but further away from the predictive mean, which is sensible since the data is generated with noise.



A picture containing map, text

Description generated with high confidence A close up of text on a white background

Description generated with high confidence

Figure 6 (Left) shows the predictive means and the predictive error bars of model 1 after training.

Figure 7 (Right) shows the predictive means and the predictive error bars of model 2 after training.

Table 4 shows the nlml of both model and the nlml ratio between them.

|  |  |  |
| --- | --- | --- |
| Model | Negative Log Marginal Likelihood | Marginal Likelihood Ratio (1:2) |
| 1 | -19.22 | 3.27x10-10 |
| 2 | -66.39 |

In the 3d plot of both models, they look similarly. The predictive means of both models show the same peaks as also shown in the pure data. But their nlmls are different.

The ml ratio between model 1 and model 2 is 3.27x10-10. It shows the model is a better fit to the data since model 2 is built upon a sum of 2 squared-exponential kernels which give a higher degree of freedom to fit the data.

Notes:

B Length long, amplitude big, amplitude small, noise big, plot of noise and signal

Maybe caught by the local minima, no guarantee, convex if few parameters, concave if more parameters. Signal to noise ratio

Additive kernel, flexibility

Cheskley decomposition positive definite 🡪 positive eigenvalue. If not positive definite, the variance cannot correspond to. The eigenvalue corresponds to the uncertainty. More data, the eigenvalue tends to be smaller. Zero eigenvalue at the two same datapoints, it is not the real degree of freedom. All eigenvalue goes up a bit by adding the identity. ~ adding the noise to the covariance matrix.

Marginal likelihood is hard to compute, but still a valid comparison.

Noise is often heavy tailed.

Kernel ~ radial exp ~ SER squared exponential, istropic

Uncertainty about the hyperparameters … marginal likelihood of hyperparameters difficult to compute, but by Monte Carle method it is possible,

Sparse Gaussian process

E compares the difference of the predictive means and error bars. 0 in independent, how close they are. 1 in dependent, close to each other.