

TP 3-b – Bayesian Linear Regression

Summary

In this TP, we studied Bayesian Linear Regression, which is a model that estimates the uncertainty of weight parameters instead of crisp weight values. Instead of learning a single optimal weight vector, Bayesian Linear Regression sets a prior distribution over the weights and computes a posterior distribution after observing data, which allows us to quantify uncertainty.

We derived and implemented the closed-form expressions for both the posterior distribution over the weights and the predictive distribution over outputs at new input points. By visualizing posterior samples and predictive uncertainty bands, we analyzed how uncertainty evolves with the number and distribution of training points, the choice of basis functions, and the structure of the dataset. In particular, we observed how predictive variance behaves when extrapolating outside the training region, when the dataset contains a hole, and when using polynomial basis functions on a sinusoidal dataset.

Questions

Question 1.2: Recall closed form of the posterior distribution in linear case. Then, code and visualize posterior sampling. What can you observe?

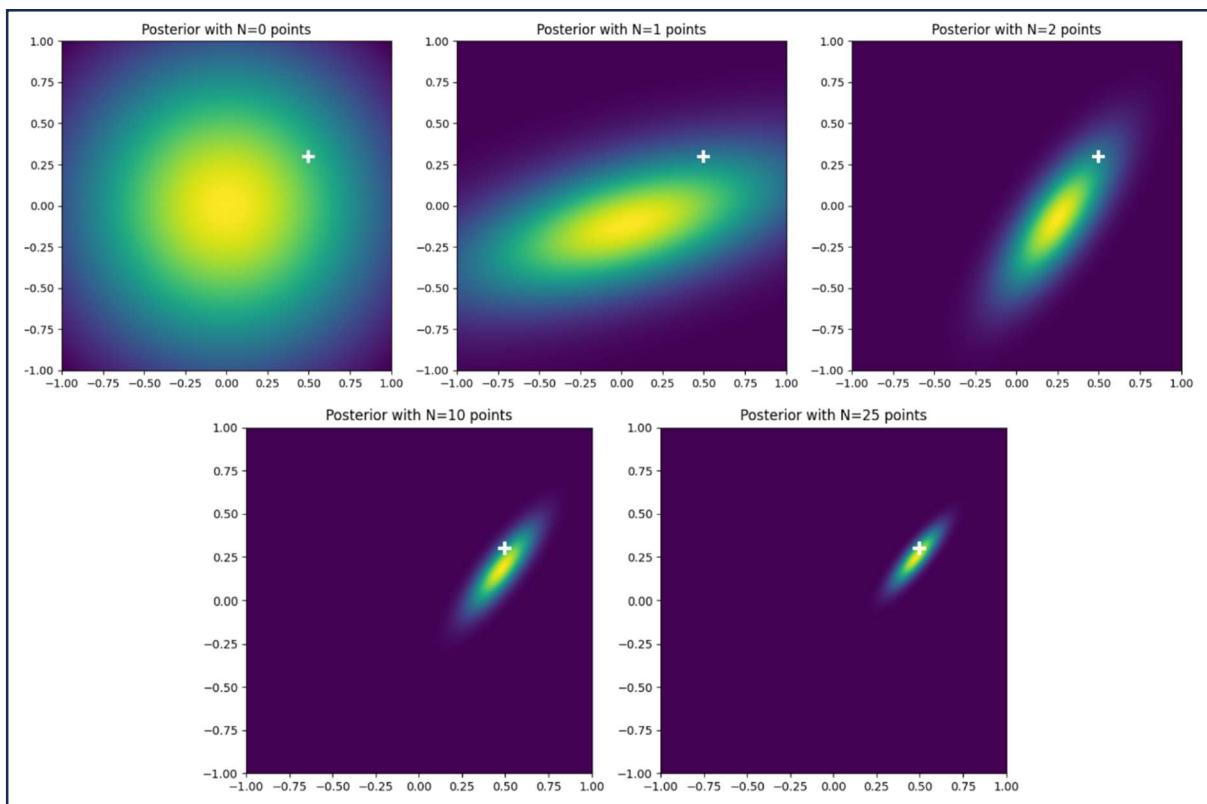


Figure 1 Visualization of the posterior distribution with growing number of training points N

The posterior is the probability distribution of the weights given the data.

So, the posterior distribution corresponds to $p(w|X, Y)$ when the data is 2D, with X and Y corresponding to the data.

And the closed form of the posterior distribution is:

$$p(w|X, Y) = N(w|\mu, \Sigma)$$

$$\Sigma_N^{-1} = \alpha I + \beta \Phi^T \Phi$$

$$\mu_N = \beta \Sigma_N \Phi^T Y$$

With α being a scalar that weights the prior and β being a scalar that weights the noise, and ϕ being the matrix resulting from the base function.

In the visualization (Figure 1), we can observe the posterior distribution variance getting smaller and smaller as the number of data points we have increases. With 0 training points, the posterior is a gaussian around the origin, and then as the data points increase, the distribution mean moves towards the ground truth (white cross) and the variance gets smaller. This shows the epistemic uncertainty of the model getting smaller as the number of data points available to train on grows.

Question 1.3: Recall and code closed form of the predictive distribution in linear case.

The predictive distribution is the probability distribution of the output at a new input point, after accounting for uncertainty in the model parameters, so after computing the posterior distribution.

We designate it as $P(y^*|x^*, D)$ where x^* is a data point and y^* is the predicted value of y given that data point, and D is the training dataset (a sample from X, Y).

The closed form of the predictive distribution is:

$$P(y^*|x^*, D) = N(y^*|\mu(x^*), \sigma^2(x^*))$$

$$\mu(x^*) = \Phi(x^*)^T \mu_N$$

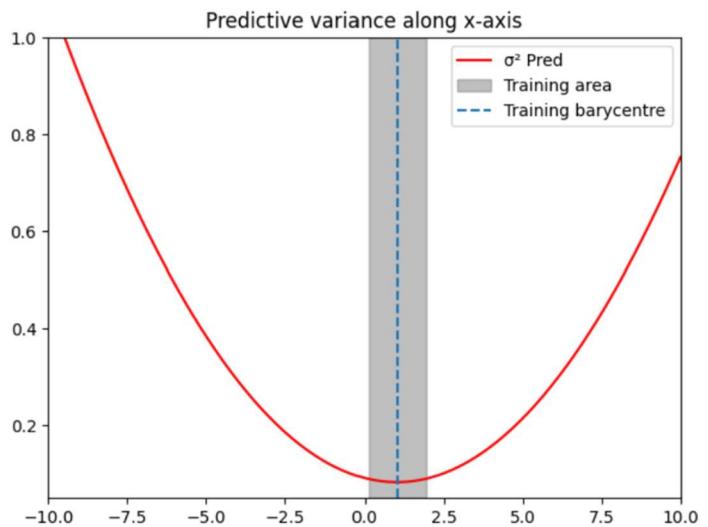
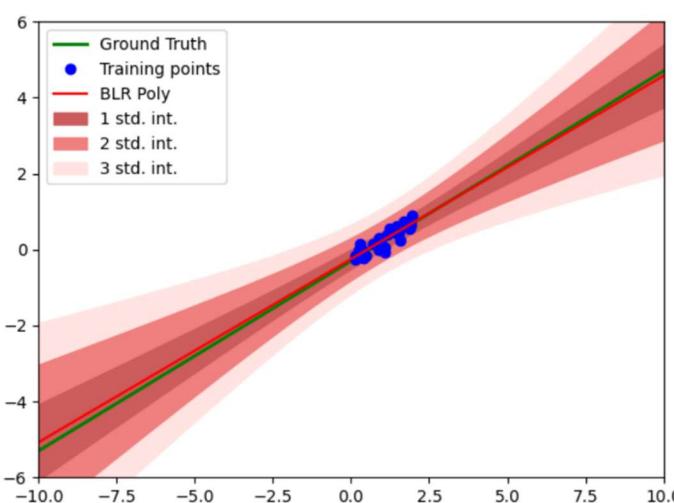
$$\sigma^2(x^*) = \beta^{-1} + \Phi(x^*)^T \Sigma_N \Phi(x^*)$$

With $\Phi(x^*)$ is the basis function evaluated at x^* .

Question 1.5: Analyze these results. Why predictive variance increases far from training distribution? Prove it analytically in the case where $\alpha = 0$ and $\beta = 1$.

Predictive variance increases when far from the training distribution because the model parameters are constrained by the training data, so predictions are confident only around the training data distribution. The model has not been trained to predict instances that do not resemble the training data, so the predictive distribution variance increases when calculated on parts of the distribution that has less or no training points. The model training only reduces model uncertainty when dealing with data from the training distribution, otherwise, it doesn't have any constraints to guide its predictions and the variance increases.

This is clearly seen in Figure 2 where the std gets broader when we get further away from the training points cluster, and in Figure 3 as we can see that the minimum predictive variance occurs at the barycenter of the training area and then it increases the further we are from this barycenter.



Marilyn CHAHINE, 21517286
Jean MACQUET, 3532408

Analytical proof for $\alpha = 0$ and $\beta = 1$:

We have that predictive variance is:

$$\sigma^2(x^*) = \beta^{-1} + \Phi(x^*)^T \Sigma_N \Phi(x^*)$$

With:

$$\Sigma_N^{-1} = \alpha I + \beta \Phi^T \Phi$$

So with $\alpha = 0$ and $\beta = 1$:

$$\Sigma_N = (\Phi^T \Phi)^{-1}$$

And thus:

$$\sigma^2(x^*) = 1 + \Phi(x^*)^T (\Phi^T \Phi)^{-1} \Phi(x^*)$$

Now, we have that $\Phi(x) = \begin{pmatrix} 1 \\ x \end{pmatrix}$ so $\Phi(x^*) = \begin{pmatrix} 1 \\ x^* \end{pmatrix}$

We also know that $(\Phi^T \Phi)^{-1}$ is a 2x2 symmetric matrix so we assume $(\Phi^T \Phi)^{-1} = \begin{pmatrix} A & B \\ B & C \end{pmatrix}$

So we now have

$$\sigma^2(x^*) = 1 + (1 \quad x^*) \begin{pmatrix} A & B \\ B & C \end{pmatrix} \begin{pmatrix} 1 \\ x^* \end{pmatrix}$$

$$\sigma^2(x^*) = 1 + (1 \quad x^*) \begin{pmatrix} A + Bx^* \\ B + Cx^* \end{pmatrix}$$

$$\sigma^2(x^*) = 1 + 1(A + Bx^*) + x^*(B + Cx^*)$$

$$\sigma^2(x^*) = 1 + A + 2Bx^* + Cx^{*2}$$

Taking $a = 1 + A$, $b = 2B$, and $c = C$, we get a quadratic equation of the form:

$$\sigma^2(x^*) = a + bx^* + cx^{*2}$$

And a quadratic formula with a positive 2nd order coefficient is a parabola with a minimum, which aligns with a dip in variance for a certain value of x^* (here, the training barycenter) and increasing values as we extrapolate.

Bonus Question: What happens when applying Bayesian Linear Regression on the linear dataset with a hole?

In this case, we have the same number of training points as in the linear dataset but they are now divided into two cluster of points that are separated by a hole. When we take this new dataset, the predictive variance increases more slowly (Figure 4), which could be because the training data is composed of two separate clusters which cover a wider part of the ground truth. But the overall predictive variance is also bigger, its minima is at 0.1 (Figure 5) while it was around 0.05 for the linear dataset with no hole (Figure 3), which is most likely due to the smaller number of data points concentrated on the same part of the ground truth distribution. The two clusters of training points allow the model to learn a wider part of the training distribution but with less certainty.

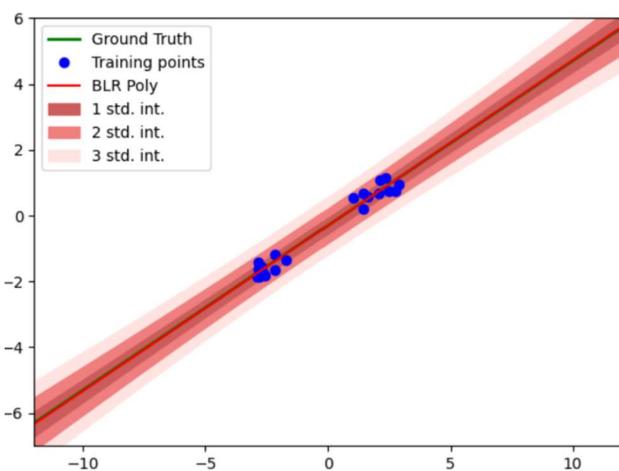


Figure 4 Visualization of the training points, ground truth, and standard deviation bands of the predictive distribution for a linear dataset with a hole learned with a Bayesian linear regression model with a linear base function

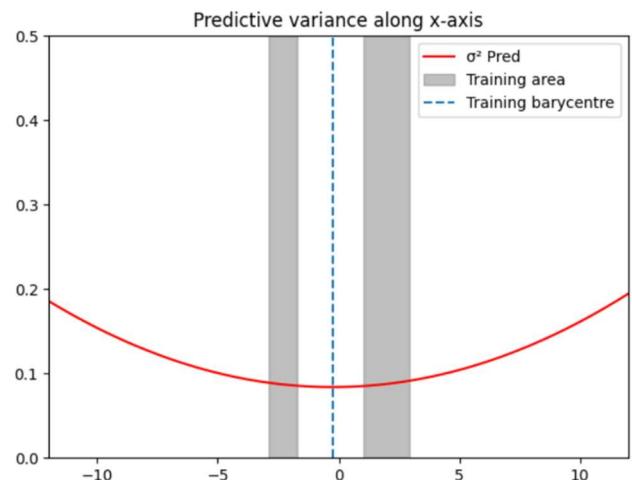


Figure 5 Predictive variance along the x-axis w.r.t. training data for a linear dataset with a hole learned with a Bayesian linear regression model with a linear base function

Question 2.2 : Code and visualize results on sinusoidal dataset using polynomial basis functions. What can you say about the predictive variance?

With the polynomial basis and sinusoidal training dataset, we can see that the predictive variance increases way faster when going further away from the training dataset. This behavior is expected based on the proof we developed in question 1.5, since in this case, Φ is a polynomial function, and thus the resulting formula for the predictive variance is not quadratic anymore, it will now also be a polynomial of degree at least D where D is the degree of Φ .

We can see this polynomial increase in Figures 6 as the std bands grow beyond the scope of the figure by $x=-1$ and $x=1.25$, and in Figure 7 where the predictive variance grows from around 0 at $x=0$ to to 3.5 at $x=-0.5$ and beyond 10 by $x=-1$, and from around 0 at $x=1$ to beyond 10 at $x=1.25$.

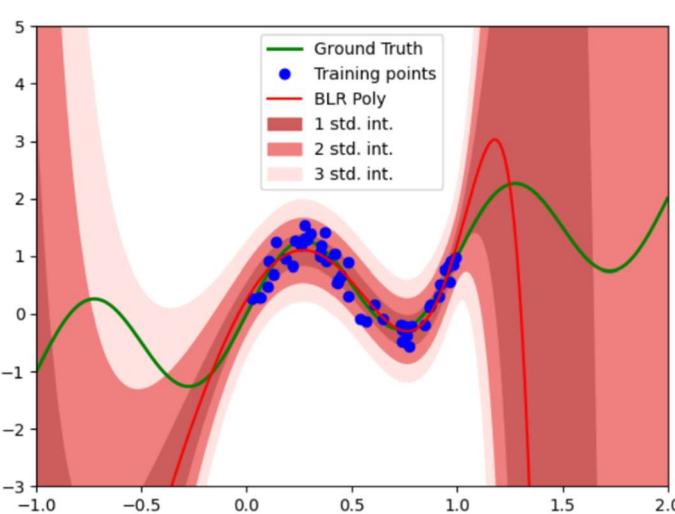


Figure 6 Visualization of the training points, ground truth, and standard deviation bands of the predictive distribution for a sinusoidal dataset learned with a Bayesian linear regression model with a polynomial base function

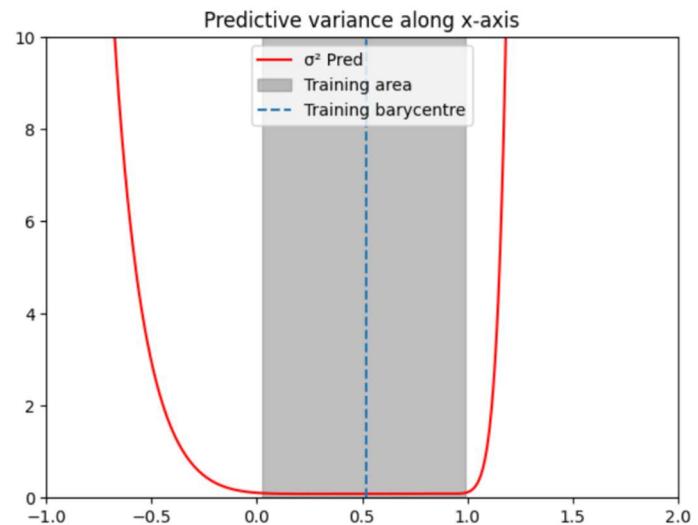


Figure 7 Predictive variance along the x-axis w.r.t. training data for a sinusoidal dataset learned with a Bayesian linear regression model with a polynomial base function