Bayesian deep learning

Deep Learning Practical Work

Aymeric Delefosse & Charles Vin

2023 – 2024

SCIENCES
SORBONNE
UNIVERSITÉ

Contents

1 Bayesian Linear Regression 2

1.1 Linear Basis function model 2

1.1.1 Gaussian basis functions 2

1.2 Non Linear models 5

1.2.1 Polynomial basis functions 5

1.2.2 Gaussian basis functions 6

1.3 Approximate Inference in Classification 7

1.3.1 Bayesian Logistic Regression 7

1.4 Bayesian Neural Networks 7

1.4.1 Variational Inference with Bayesian Neural Networks 7

1.4.2 Monte Carlo Dropout 7

Chapter 1

Bayesian Linear Regression

1.1 Linear Basis function model

1.1.1 Gaussian basis functions

1.2. Recall closed form of the posterior distribution in linear case. Then, code and visualize posterior sampling. What can you observe? Let N denote the number of training examples, K the number of dimensions of the outputs, and p the number of features (or predictors) in the input data. Consider $X \in \mathbb{R}^{N\times p}$, the input matrix, and $Y \in \mathbb{R}^{N\times K}$, the output matrix.

We typically define the posterior distribution as $p(w|X, Y)$. This distribution represents our updated beliefs about the parameters w after observing the data X and Y. According to Bayes' rule, we can express the posterior distribution as the product of two components:

1. $p(w)$, which represents our prior beliefs about the distribution of w before observing the data.
2. $p(Y|X, w)$, indicating the probability of observing the data Y given the parameters w and the data X. This quantifies how likely our data is under different hypotheses represented by w.

Therefore, the formula for the posterior distribution is given by:

$$p(w|X, Y) \propto p(Y|X, w)p(w)$$

In this formula, we start with our initial beliefs (priors) and then update these beliefs based on the new data (likelihood).

In the case of our linear model, we know that $y_i = \Phi_i^T w + \epsilon$, with $\Phi \in \mathbb{R}^{N\times(p+1)}$ representing the design matrix and ϵ denoting the residual. Assuming that the error follows a centered Gaussian distribution with standard deviation $\beta^{-1} = 2\sigma^2$, meaning that $\epsilon \sim \mathcal{N}(0, \beta^{-1})$. Consequently, we can conclude that:

$$p(y_i|x_i, w) \sim \mathcal{N}(\Phi_i^T w, \beta^{-1})$$

Furthermore, we selected a centered Gaussian prior with a variance of $\alpha^{-1}I$ whereα governs the prior distribution over the weights $\mathbf{w}$:

$$p(w|\alpha) \sim \mathcal{N}(0, \alpha^{-1}I)$$

In this specific case, we can demonstrate that the posterior distribution $p(\mathbf{w}|\mathbf{X}, \mathbf{Y})$ follows a Gaussian distribution as follows:

$$p(\mathbf{w}|\mathbf{X}, \mathbf{Y}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

The precision matrix Σ, which is the inverse of the covariance matrix of the distribution parameters, is defined as:

$$\boldsymbol{\Sigma}^{-1} = \alpha\mathbf{I} + \beta\Phi^T\Phi$$

The mean of the distribution parameters is given by:

$$\boldsymbol{\mu} = \beta\boldsymbol{\Sigma}\Phi^T\mathbf{Y}$$

Parameters α and β serve analogous roles, with α governing the prior distribution and β regulating the likelihood.

Now, we can proceed to sample from the updated (with data) posterior distribution $p(w|X,Y)$. This process is demonstrated in Figure 1.1, with $\alpha = 2$ and $\beta = (2 * 0.2^2)^{-1}$. It's worth noting that when $N = 0$, the posterior distribution $p(w|X,Y)$ simplifies to the prior distribution $p(w)$.

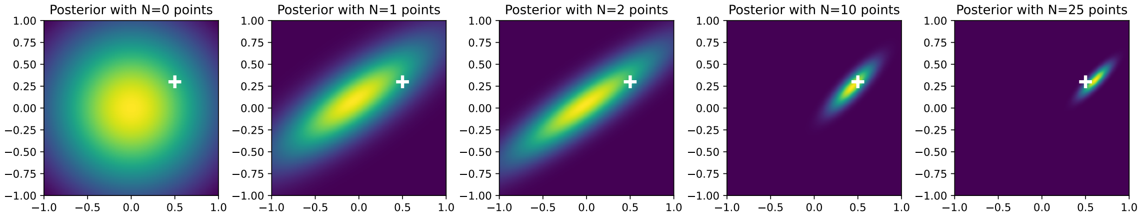

Figure 1.1: Evolution of a Bayesian posterior distribution with increasing data points: A visual representation of Bayesian inference, depicting how the posterior distribution updates as more data is incorporated. From left to right, the figures show the posterior with $N = 0$ (prior distribution), $N = 1$, $N = 2$, $N = 10$, and $N = 25$ data points, respectively. The cross mark represents the ground truth.

1.3. Recall and code closed form of the predictive distribution in linear case. Using the information from the previous question, we can now calculate the predictive distribution for new data point $x^\star$ by marginalizing over the parameter w. This is represented by the following equation where $\mathcal{D} = \{X, Y\}$ denotes the dataset:

$$p(y|x^\star, \mathcal{D}, \alpha, \beta) = \int p(y|x^\star, w, \beta) p(w|\mathcal{D}, \alpha, \beta)$$

By leveraging the property that the convolution of two Gaussian distributions results in another Gaussian distribution, we can demonstrate that the closed-form expression for the predictive distribution in the linear case is as follows:

$$p\left(y|x^\star; \mathcal{D}, \alpha, \beta\right) = \mathcal{N}\left(y; \mu^T \mathbf{\Phi}(x^\star), \frac{1}{\beta} + \mathbf{\Phi}(x^\star)^T \mathbf{\Sigma} \mathbf{\Phi}(x^\star)\right)$$

We can see that the variance in the predictive distribution σ^2_{pred} for a new observation can be divided into two parts:

1. The aleatoric uncertainty, which represents the inherent noise in the data, that we fixed around β^{-1}.
2. The epistemic uncertainty related to the model parameters w, characterized by $\mathbf{\Phi}(x^\star)^T \mathbf{\Sigma} \mathbf{\Phi}(x^\star)$

It's worth noting that as the number of data points N approaches infinity ($\lim_{N\to\infty} \mathbf{\Phi}(x^\star)^T \mathbf{\Sigma} \mathbf{\Phi}(x^\star) = 0$), our understanding of the model parameters becomes nearly perfect. In this scenario, the only remaining source of uncertainty is the aleatoric uncertainty, stemming from the noise in the data.

1.4. Based on previously defined `f_pred()`, predict on the test dataset. Then visualize results using `plot_results()` defined at the beginning of the notebook. Figure 1.2 illustrates a comparison between the predictions made by a Bayesian Linear Regression (BLR) model and the actual ground truth. In the left panel, you can see the model's linear fit to the training data, shown as blue points, alongside the true ground truth represented by the green line. To visualize the model's predictive uncertainty, shaded areas are used, ranging from dark to light, which correspond to one, two, and three standard deviation intervals, respectively.

The right panel of the figure focuses on the predictive variance σ^2_{pred} along the x-axis. This variance is depicted by a curve that widens as it moves away from the center of the training data, marked by the vertical dashed line. These visual elements together provide a comprehensive view of the model's confidence in its predictions across the entire domain.

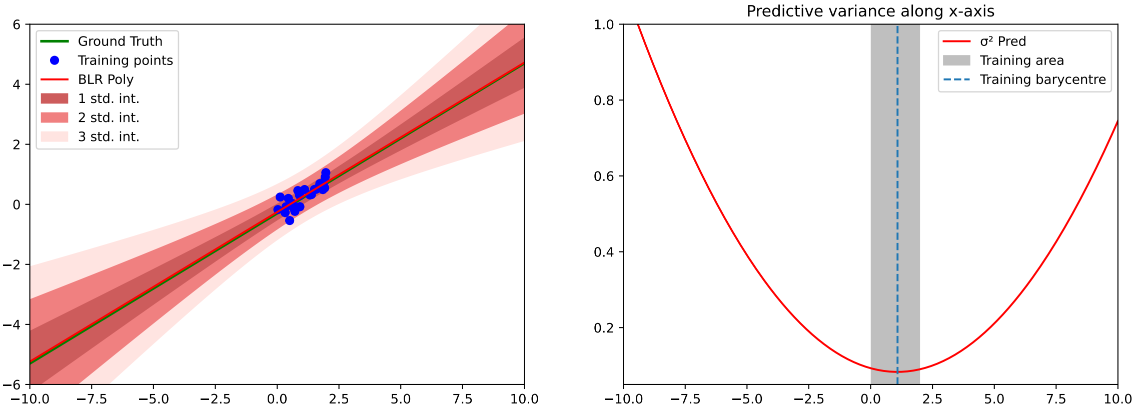

Figure 1.2: The left panel illustrates the Bayesian Linear Regression (BLR) linear fit to the training data (blue points) against the ground truth (green line). The shaded areas represent the predictive uncertainty, with one, two, and three standard deviation intervals shown in progressively lighter shades. The right panel displays the predictive variance σ^2_{pred} across the x-axis. The vertical dashed line indicates the center of the training data.

1.5. Analyse these results. Why predictive variance increases far from training distribution? Prove it analytically in the case where $\alpha = 0$ and $\beta = 1$. In the left panel of Figure 1.2, we can observe that the confidence intervals are narrowest near the cluster of training points. This suggests higher confidence in predictions within this area. As we move away from the center of the training data (towards the extremities of the x-axis), the confidence intervals become wider, indicating increasing uncertainty in the model's predictions.

Looking at the right panel of Figure 1.2, we notice that the variance remains low in the region where the training data is located (the grey shaded area). This correlates with the tight confidence intervals shown in the left panel. As expected, the predictive variance is lowest near the training barycentre, reflecting greater model certainty in this region due to the presence of more training data points.

As we move away from the training area on either side, the predictive variance increases significantly, which is consistent with the expanding confidence intervals in the left panel. This sharp increase in variance indicates a significant decrease in the model's confidence in its predictions outside the range of the training data. This is a common pattern in regression analysis, as the model has less data to rely on for predictions in these regions. Thus, the Bayesian Linear Regression model offers more reliable predictions within the training data range, with confidence diminishing as predictions move away from this range.

With $\alpha = 0$ and $\beta = 1$, the computation of Σ^{-1} simplifies as follows:

$$\begin{aligned}
\Sigma^{-1} &= 0 \cdot Id_3 + 1 \cdot \phi^T \phi \\
&= \phi^T \phi \\
&= \begin{pmatrix} 1 & \dots & 1 \\ x_1 & \dots & x_N \end{pmatrix} \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{pmatrix} \\
&= \begin{pmatrix} N & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}
\end{aligned}$$

To invert Σ, we use the classic formula for a 2×2 matrix:

$$\Sigma = \frac{1}{\det \Sigma^{-1}} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & N \end{pmatrix}$$

Returning to our expression for $\sigma^2_{pred} = \Phi(x^\star)^T\Sigma\Phi(x^\star)$, we have:

$$\begin{aligned}
\sigma^2_{pred} = \Phi(x^\star)^T\Sigma\Phi(x^\star) &= \frac{1}{\det \Sigma^{-1}} \begin{pmatrix}1\\ x^\star\end{pmatrix} \begin{pmatrix}\sum x_i^2 & -\sum x_i \\ -\sum x_i & N\end{pmatrix} \begin{pmatrix}1 & x^\star\end{pmatrix} \\
&= \frac{1}{\det \Sigma^{-1}} \left(\sum_i x_i^2 - x^\star \sum_i x_i \quad x^\star N - \sum_i x_i\right) \begin{pmatrix}1\\ x^\star\end{pmatrix} \\
&= \frac{1}{\det \Sigma^{-1}} \left(\sum_i x_i^2 - x^\star \sum_i x_i + {x^\star}^2 N - x^\star \sum_i x_i\right) \\
&= \frac{1}{\det \Sigma^{-1}} \left(\sum_i x_i^2 - 2x^\star \sum_i x_i + {x^\star}^2 N\right) \\
&= \frac{1}{\det \Sigma^{-1}} \sum_i (x_i - x^\star)^2
\end{aligned}$$

This formula reveals that the predictive variance is directly proportional to the squared differences between each training data point x_i and the prediction point $x^\star$. As $x^\star$ moves away from the region where the training data is concentrated, these squared differences increase, consequently leading to an increase in the predictive variance.

Bonus: What happens when applying Bayesian Linear Regression on the following dataset? ...

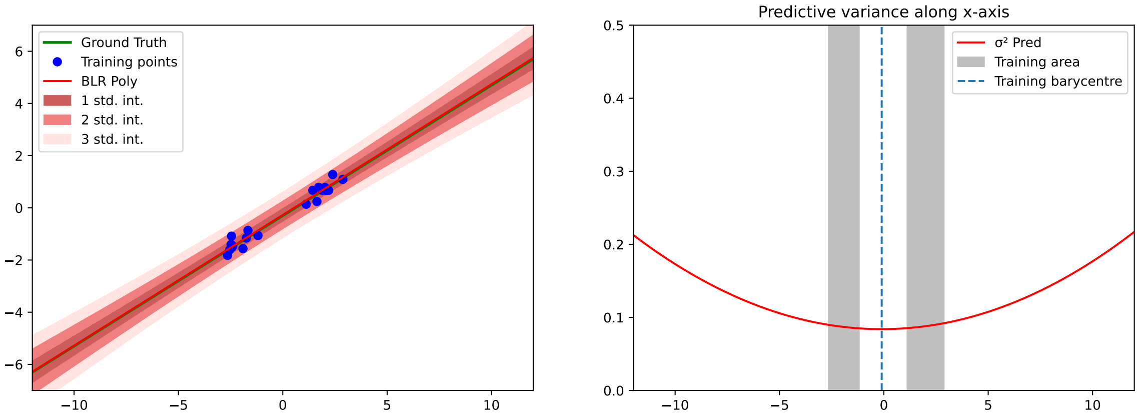

Figure 1.3: Visualization of predictive distribution of a linear dataset featuring a "hole": The left panel illustrates the Bayesian Linear Regression (BLR) linear fit to the training data (blue points) against the ground truth (green line). The shaded areas represent the predictive uncertainty, with one, two, and three standard deviation intervals shown in progressively lighter shades. The right panel displays the predictive variance σ^2 across the x-axis. The vertical dashed line indicates the center of the training data.

1.2 Non Linear models

1.2.1 Polynomial basis functions

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

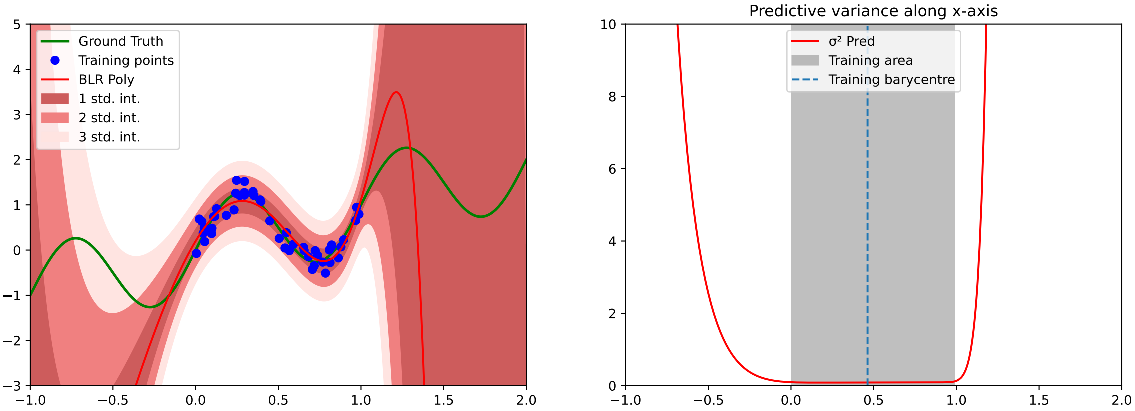

Figure 1.4: <caption>

1.2.2 Gaussian basis functions

2.4. Code and visualize results on sinusoidal dataset using Gaussian basis functions. What can you say this time about the predictive variance? ...

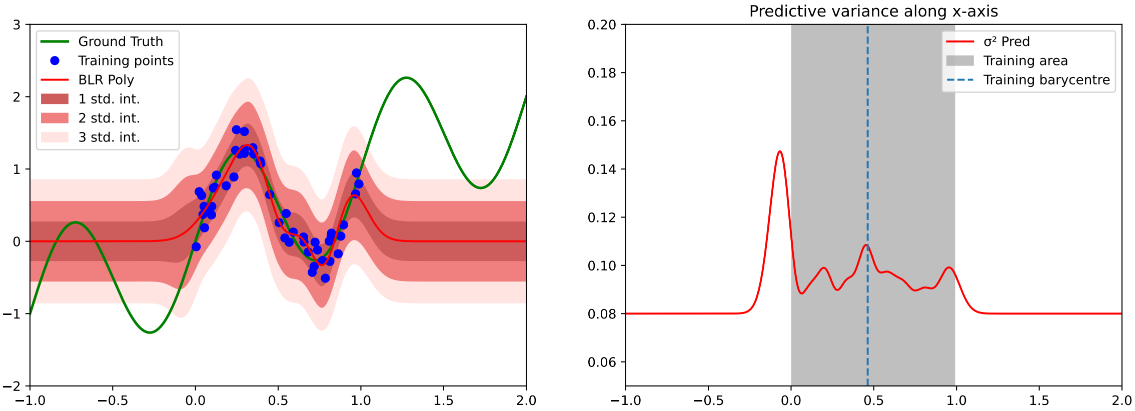

Figure 1.5: <caption>

2.5. Explain why in regions far from training distribution, the predictive variance converges to this value when using localized basis functions such as Gaussians.

1.3 Approximate Inference in Classification

1.3.1 Bayesian Logistic Regression

Maximum-A-Posteriori Estimate

1.1. Analyze the results provided by previous plot. Looking at $p(\mathbf{y} = 1|\boldsymbol{x}, \boldsymbol{w}_{\mathbf{MAP}})$, what can you say about points far from train distribution? In Figure 1.6, which illustrates Bayesian Logistic Regression applied to a binary classification dataset, we can see two clusters of data points, one represented in blue and the other in red, corresponding to different classes. The shaded regions, in blue and red, depict the predictive uncertainty associated with each class. The lighter shade indicates lower confidence in the model's predictions.

Approximating $p(\mathbf{w}|\mathbf{X}, \mathbf{Y})$ with a Dirac delta function is essentially akin to approximating the predictive distribution using $\mathbf{w}_{\mathrm{MAP}}$, meaning $p(y = 1|\mathbf{x}, \mathbf{w}_{\mathrm{MAP}}) \approx p(y = 1|\mathbf{x}, \mathbf{Y})$. This approximation is quite rudimentary.

As depicted in Figure 1.6, the model's uncertainty doesn't significantly increase far from the training data. This limitation indicates that the point-wise estimate of the parameters can only confidently assign points to their respective classes but lacks the capacity to provide nuanced uncertainty measures for points that deviate far from the training data distribution.

Figure 1.6: Illustration of a Bayesian Logistic Regression model applied to a binary classification task with uncertainty display. Two distinct data point clusters — blue and red — represent separate classes, while the surrounding shaded areas reflect the model's predictive uncertainty, with lighter shades indicating lower confidence. The model exhibits a limitation in uncertainty estimation, remaining overconfident far from the training data, hence lacking in providing nuanced uncertainty measures for outlying points.

Laplace Approximation

1.2. Analyze the results provided by previous plot. Compared to previous MAP estimate, how does the predictive distribution behave?

1.3. Comment the effect of the regularisation hyper-parameter WEIGHT_DECAY.

Variational Inference

1.4. Comment the code of the VariationalLogisticRegression and LinearVariational classes. = answer to see if we understand what the code does

1.5. Comment the code of the training loop, especially the loss computation. Analyze the results provided by previous plot. Compared to previous MAP estimate, how does the predictive distribution behave? What is the main difference between the Variational approximation and the Laplace approximation?

1.4 Bayesian Neural Networks

1.4.1 Variational Inference with Bayesian Neural Networks

2.1. Analyze the results showed on plot.

1.4.2 Monte Carlo Dropout

2.2. Again, analyze the results showed on plot. What is the benefit of MC Dropout variational inference over Bayesian Logistic Regression with variational inference?