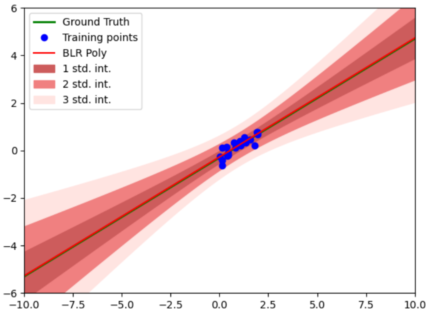
Robust Deep Learning
Experimental Report

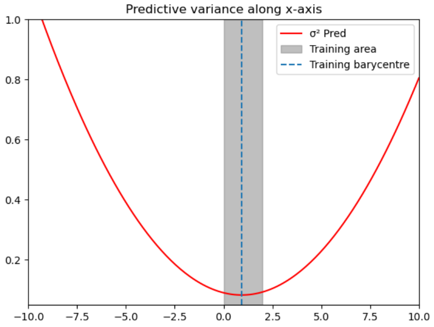
Maxence Gollier

November 2023

Week 1

Bayesian Linear Regression





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

We see that near the data, the variance is low and that it increases as we move away from the data on the left plot. On the right plot, we see that when we are at the training barycenter, the predictive variance is about 0.08 which is equal to $\frac{1}{\beta}$; we only have aleatoric uncertainty on the data's barycenter.

The predictive variance is a sum of an aleatoric uncertainty $(\frac{1}{\beta})$ and an epistemic one $(\phi(x^*)^T \Sigma \phi(x^*))$, the aleatoric variance is independant of the data this is why we have a residual variance near the data but the epistemic one increases as we get farther away of the training set. We prove this claim in the simple case where $\alpha = 0, \beta = 1$:

In that case, we have

$$\Sigma^{-1} = \phi^T \phi = \begin{pmatrix} N & 1^T X \\ 1^T X & X^T X \end{pmatrix} \tag{1}$$

which implies

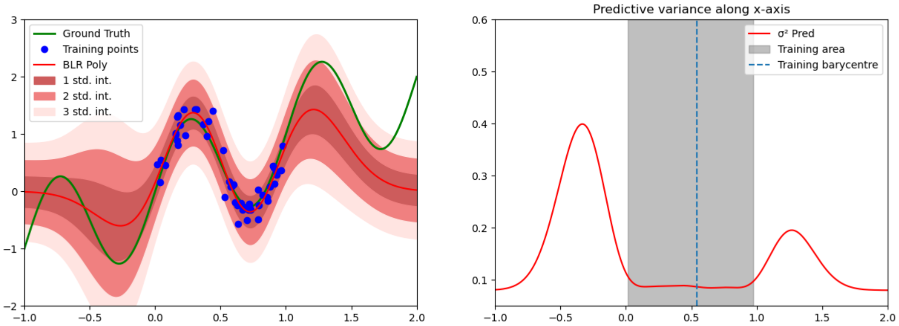
$$\Sigma = \frac{1}{NX^TX - (1^TX)^2} \begin{pmatrix} X^TX & -1^TX \\ -1^TX & N \end{pmatrix} \tag{2}$$

Now, computing the epistemic uncertainty $\phi(x^*)^T\Sigma\phi(x^*)$,

$$\phi(x^*)^T\Sigma\phi(x^*) \propto \kappa(X) - 2x^*\sum_i X_i + N(x^*)^2 \tag{3}$$

Taking the derivative, we find that the minimum of the epistemic uncertainty lies at $x^*_{\text{min}} = \frac{1}{N}\sum_i X_i$. Hence, once x^* goes away from the barycenter of data $\frac{1}{N}\sum_i X_i$, the variance increases.

Non-Linear Regression



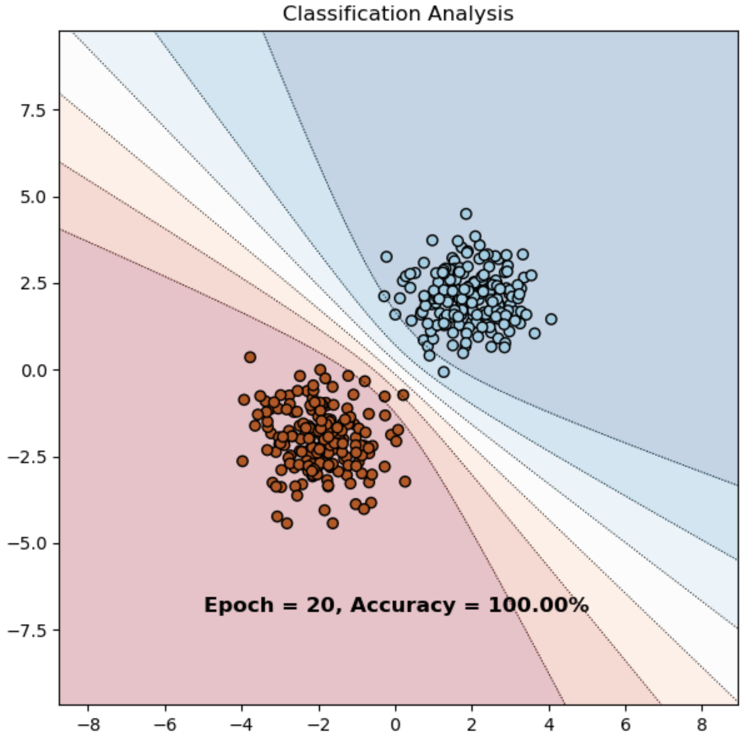
We see that the predictive variance increases as we leave the training area and then decreases as we get farther from it. This is not good because the epistemic uncertainty should increase.

Question 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.

Because $\phi(x^*) \to 0$ when x^* is far away from the different means of the gaussians basis function which are in the range $[0, 1]$ in our case. We can see on the right plot that the epistemic uncertainty starts to increase as we leave the training area, but at some point, It is overshadowed by the fact that $\phi(x^*)$ is too small so the epistemic uncertainty goes to zero and the predictive variance tends to $\frac{1}{\beta}$.

Week 2

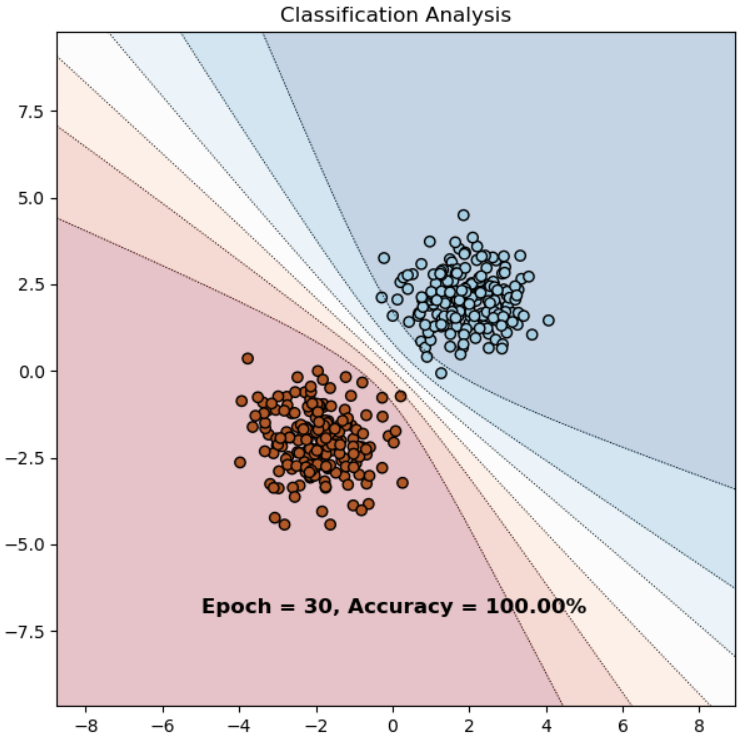
Question 1.2: Analyze the results provided by previous plot. Compared to previous MAP estimate, how does the predictive distribution behave?



Close to the data, we have the same result as the previous MAP estimate. However, for unseen data, the epistemic uncertainty increases. This is what we expect as ouput since the predictive variance is no longer constant and is the sum of an aleatoric part and an epistemic one.

Question 1.3: Analyze the results provided by previous plot. Compared to previous MAP estimate, how does the predictive distribution behave?

Again, like in Question 1.2, the predictive variance is no longer constant and we can repeat what has been said about the MAP estimate for this network. This network however is different from the Laplace Approximation one; we see that the variance is larger everywhere. In a sense, the network is more cautious on its predictions. This might be due to the fact that the variance with the



Laplace approximation is only local on the mode of the posterior distribution.

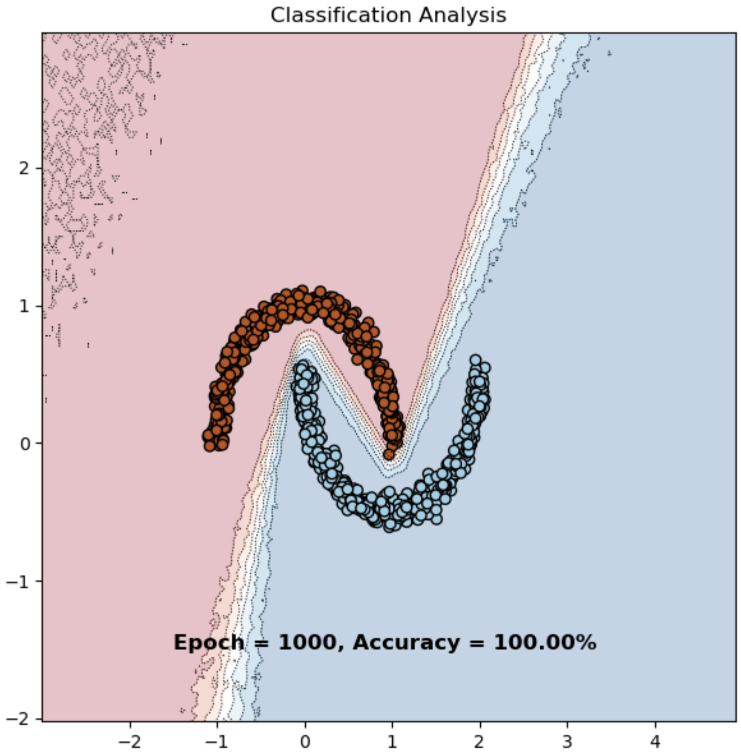
Comment on the class LinearVariational : This class samples its parameters using the reparametrization trick and computes

$$out = Wx + b \tag{4}$$

Note that the sigmoid is not applied in this class. It also computes the KL-divergence at each forward pass.

Question 2.1: Again, analyze the results showed on plot. What is the benefit of MC Dropout variational inference over Bayesian Logistic Regression with variational inference?

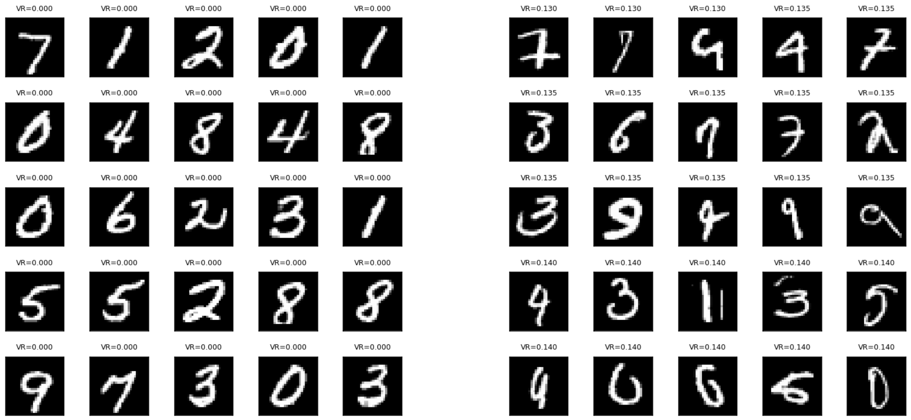
We see that, up to some noise for the MC Dropout variational inference, the results are almost identical and the separating curve has the same shape on both outputs. The main benefit is the simplicity as well as the execution time of the MC Dropout, it only took 11.58 seconds for the whole training process with dropout against 35.81 seconds for the Bayesian Logistic Regression. Furthermore, the implementation of the Dropout is easier as we just need to add a



dropout layer in the network and we don't need to directly sample from distributions.

Week 3

Explore image with low and high confidence. Look at images and comment the confidence metric.



(a) Confident Sample

(b) Uncertain Sample

We see that the variation-ratio metric does give an idea on how good the image can be predicted. The confident sample only shows images which should be well predicted by any MINST classification network. On the other hand, the uncertain sample shows some images that are ambiguous even for the human eye. We can see for example on the bottom left that the network is unsure wether this is a 4 or a 9

Explain the goal of failure prediction

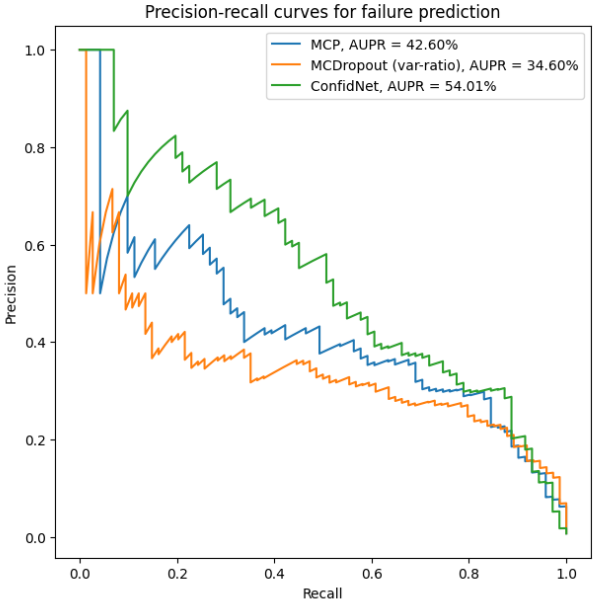
The goal of failure prediction is to estimate the quality of a prediction and to be able to accept or reject a prediction based on this estimation. This is needed in a lot of contexts where for instance safety is needed like medical imaging because errors can have serious implications.

Comment the code of the LeNetConfidNet class

LeNetConfidNet reimplements the LeNet Network from the last section. In a forward pass, we compute the result from the LeNet Network and then compute the uncertainty on this input by using 5 fully connected layers with the output from LeNet as input and ReLU activation function.

From that, the ConfidNet network trains at the same time the output and the confidence estimation. In practice, we first train the output estimation and then the confidence by freezing the LeNet convolutional layers.

Analyze results between MCP, MCDropout and ConfidNet

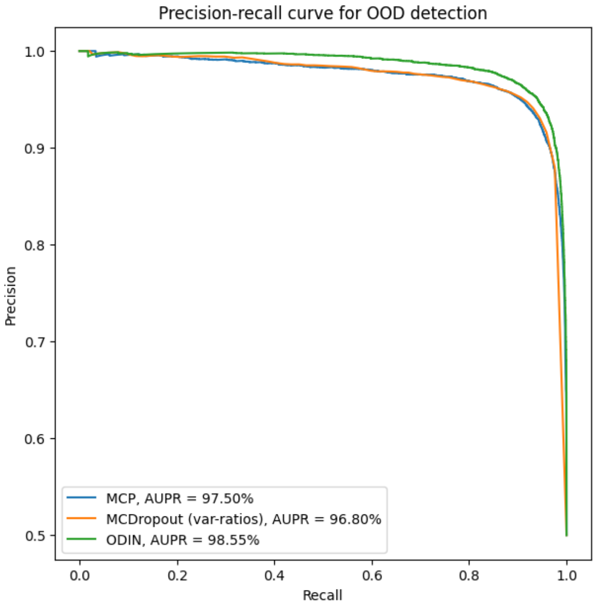


We see that ConfidNet outperforms MCP which outperforms MCDropout. ConfidNet tries to learn the True Class Probability. This implies that when it is well trained, we get better results than simply taking the Maximum Class Probability.

For the MCP, the confidence is the probability of the softmax of the output of LeNet Network while we use samples and dropout for the MCDropout with entropy as confidence measure.

Question 3.1: Compare the precision-recall curves of each OOD method along with their AUPR values. Which method perform best and why?

We see that ODIN uniformly outperforms MCP and MCDropout with variational ratios. This is due to the fact that ODIN tries to maximize the difference between inliers and outliers and then makes a forward pass on this data again. Since the data is preprocessed, it performs better than MCP which does not make this preprocessing.


Precision-recall curve for OOD detection
1.0
0.9
0.8
0.7
0.6
0.5
Precision
0.0
0.2
0.4
0.6
0.8
1.0
Recall
MCP, AUPR = 97.50%
MCDropout (var-ratios), AUPR = 96.80%
ODIN, AUPR = 98.55%