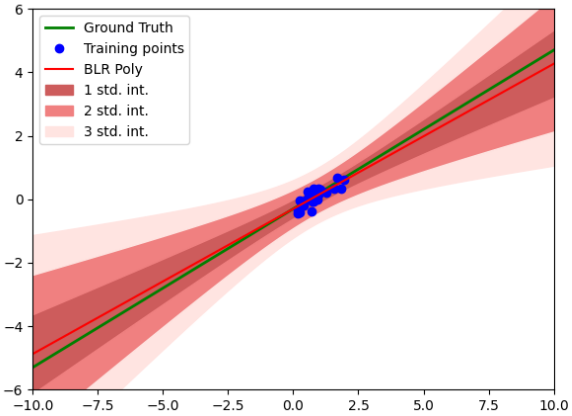
Maxime CORLAY

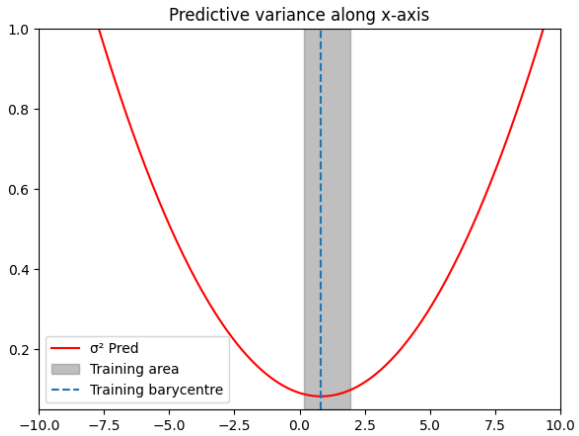
Master MVA 2024-2025

Report of the three practical sessions of Foundations of deep learning

Week 1:

- Bayesian linear regression: results of the predictive distribution on the synthetic dataset [Question 1.4]





- Theoretical analysis to explain the form of the distribution (simplified case α=0, β=1) [Question 1.5]

Predictive variance increases far from the training distribution because it is a quadratic form ! Indeed, we know that Σ is a covariance matrix of a gaussian distribution, so it is symmetric definite positive. Thus, we can write $\Sigma = A^T A$ where A is a symmetric definie positive matrix (all eigenvalues > 0). Hence, $\sigma^2_{pred} = 1/\beta + ||A\Phi(x^*)||^2$ is clearly a quadratic form of x^*.

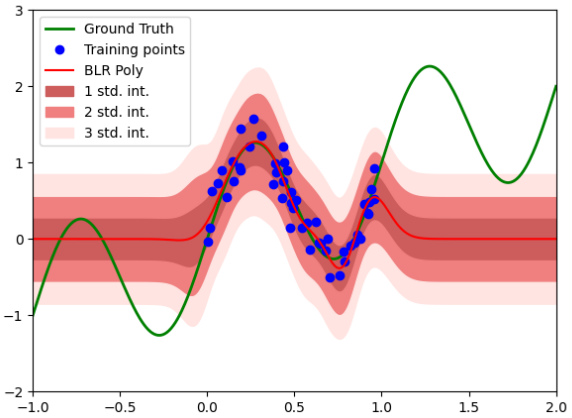
Let's prove that more particularly in the case $\alpha = 0$ and $\beta = 1$. Now we have that $\Sigma = (\Phi^T\Phi)^{-1}$ so

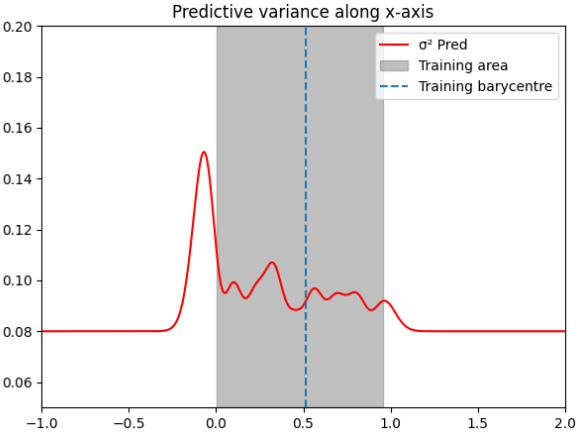
$$\sigma^2_{pred} = 1 + \Phi(x^*)^T(\Phi^T\Phi)^{-1}\Phi(x^*) = 1 + \begin{pmatrix}1 & x^*\end{pmatrix}(\Phi^T\Phi)^{-1}\begin{pmatrix}1\\ x^*\end{pmatrix} = \begin{pmatrix}1 & x^*\end{pmatrix}\begin{pmatrix}(\Phi^T\Phi)^{-1}_{11} + (\Phi^T\Phi)^{-1}_{12}x^*\\ (\Phi^T\Phi)^{-1}_{21} + (\Phi^T\Phi)^{-1}_{22}x^*\end{pmatrix} = 1 + (\Phi^T\Phi)^{-1}_{11}.$$

$$+ (\Phi^T\Phi)^{-1}_{12}x^* + (\Phi^T\Phi)^{-1}_{21}x^* + (\Phi^T\Phi)^{-1}_{22}x^{*2}$$

It is a quadratic function. That explains why the variance grows when x^* is far from the training dataset.

- Non-linear regression: analysis of the Gaussian basis feature maps results [Question 2.4/2.5]





Outside the training area, we obtain a predictive variance of 0.08.

To explain that, let us recall the formula:

$\sigma^2_{pred} = 1/\beta + ||A\Phi(x^*)||^2$ (I use the same notations as before)

All centers of gaussians are in $[0, 1]$ so when x^* is far away from that area, $\Phi_i(x^*) = 0$ for all i. As a result, $\Phi(x^*) = 0$ and $\sigma^2_{pred} = 1/\beta = 2\sigma^2 = 2 \times 0.2^2 = 0.08$

Week 2:

- Commente Laplace's approximation results [Question 1.2]

It is a bit better than before because there is a big area where the model isn't confident in its prediction, and that is very good. But there is still a problem because "behind" each blob there is still high confidence in a kind of infinite "cone", and this high confidence is only due to the model.

- Part I.3 « Variational inference » : comment the class LinearVariational. What is the main difference between Laplace's and VI's approximations?

I commented many things directly in the cells. I could add here that the structure is very similar to most neural networks, with init and forward.
The main difference is the adding of kl_divergence.

- MC dropout results: analyse predictive distribution on the 2-moons dataset [Question 2.1]. What is the main difference between MCdropout and the VI approximation in part I.3?

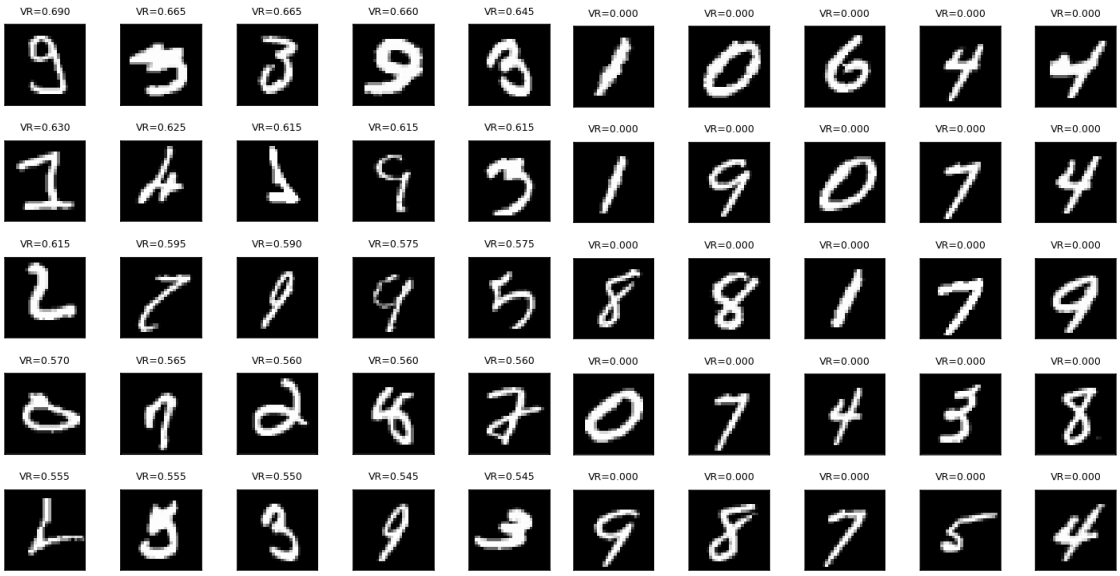
The convergence is very rapid and then the process seems to be essentially aleatoric but the first plot shows that the accuracy is very high.

MC Dropout variational inference costs less memory than Bayesian Logistic Regression with VI. But of course, the result obtained with MC Dropout is more noisy because of the aleatoric property of dropout.

Week 3:

- Comment results for investigating most uncertain vs confident samples [I.1]

From one run of the notebook to another, the order of the "worse" pictures can vary a bit, but I obtain the same picture in general. Visually, I can confirm that these picture are a bit ambiguous, when compared to the "best" pictures in term of confidence.



We see a clear difference between numbers on the left and numbers on the right.

- Failure precition:

- Explain the goal of failure prediction

The goal of failure prediction is simple : we have a classifier. We want to predict the uncertainty of each prediction in order to see if it is relevant to ask a human operator a complementary advice. If the predicted uncertainty is higher than a threshold, we reject automatic prediction and call human.

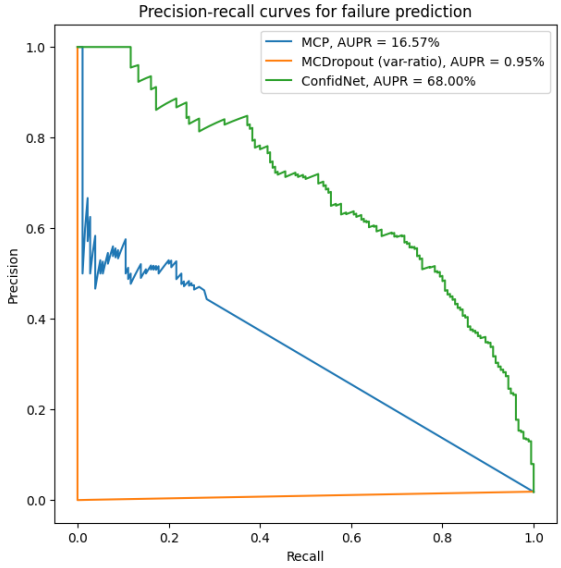
The fun thing is that we used a model to make the prediction, and we use an auxiliary model to predict the uncertainty of the predictions of the first model.

- Comment the code of the LeNetConfidNet class [II.1]

We compute the feature map, we flatten it, do the 2 dropouts and the first linear layer. But instead of going to the second (and last) linear layer, we go in the “uncertainty MLP”.

- Analyze results between MCP, MCDropout and ConfidNet [II.2]

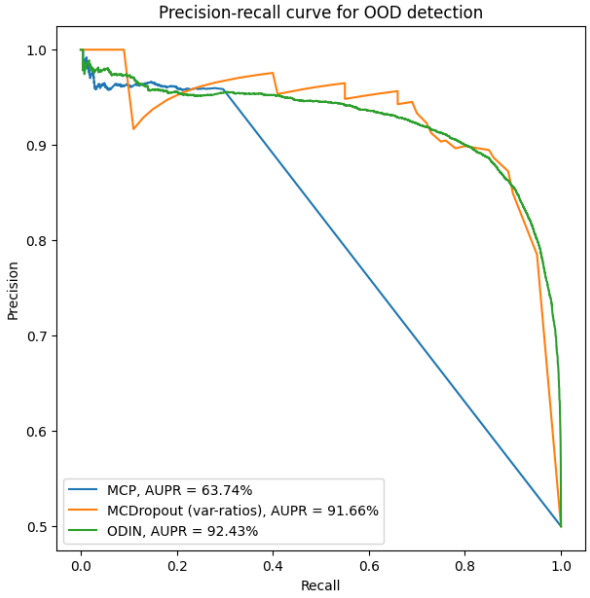
This is what I obtained:



The best method is the one for which the corresponding curve is the nearest to the upper right hand corner (precision=1,recall=1).

I obtained that the best method here is ConfidNet. It is also the method with the highest AUPR.

- OOD detection: analyse results and explain the difference between the 3 methods [III.1]



The best method is the one for which the corresponding curve is the nearest to the upper right hand corner (precision=1,recall=1). I obtained that the best method here is ODIN (because the best situation is the upper right hand corner), as in the slide. But note that depending on the execution, there can be sometimes a difference with the slide result.

MCP and ODIN both produce for each data a row of probabilities using softmax function and the prediction is the class associated to the maximum probability, and **the uncertainty is that maximum probability.** The difference between them is that MCP does not modify the model or the data, while ODIN modifies the DATA (like in an adversary attack).

MC Dropout is very different. **The uncertainty is computed by activating dropout and collecting predictions many times (each time, the weights activated change, they are not the same). We use s samples. We create a histogram for each data (in which each bar represents a class). The "global" prediction for the data is the class with the highest bar, and the corresponding uncertainty is then the frequency of that "global prediction".**

	Data modified ?	Model modified ?	Uncertainty	
MCP	**No**	**No**	Max probability	
ODIN	**Yes**, Attack	**No**	Max probability	Often the best
MC Dropout	**No**	**Yes**, dropout	Histogram of s predictions (and we use entropy or frequency of max…)	