Robust Deep Learning

Simon Hugo
Master MVA
hugo.simon@telecom-paris.fr

1 Week 1

1.1 Bayesian linear regression: results of the predictive distribution on the synthetic dataset.

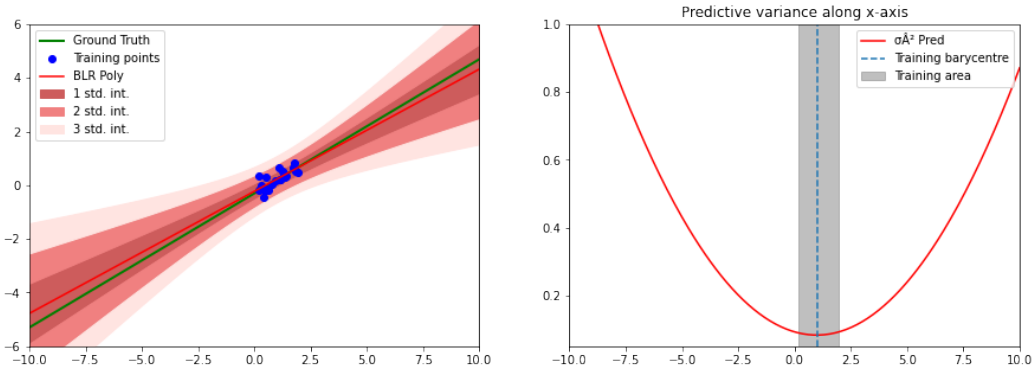


Figure 1: Predictive variance tends toward infinity far from training samples.

1.2 Why predictive variance increases far from training distribution?

The linear model holds high uncertainty for out of distribution samples, as seen in Figure 1. Indeed we have

$$\sigma^2_{\text{pred}}(\mathrm{x}) = \frac{1}{\beta} + \Phi(\mathrm{x})^T \Sigma \Phi(\mathrm{x})$$

Thus, for linear Φ and if Σ is not degenerated, $\sigma^2_{\text{pred}} \xrightarrow[\|x\| \to +\infty]{} +\infty$

In the particular case where $\alpha = 0$ and $\beta = 1$, it reduces to $\sigma^2_{\text{pred}}(\mathrm{x}) = 1 + \Sigma x^2 = O(x^2)$

1.3 Comment results on sinusoidal dataset using Gaussian basis functions.

We see in Figure 2 when using Gaussian basis functions that predictive variance converges far from training samples.

Recall we have $\sigma^2_{\text{pred}}(\mathrm{x}) = \frac{1}{\beta} + \Phi(\mathrm{x})^T \Sigma \Phi(\mathrm{x})$

But in the localized basis functions case, we have $\Phi(x) \xrightarrow[\|x\| \to +\infty]{} 0$, thus $\sigma^2_{\text{pred}} \xrightarrow[\|x\| \to +\infty]{} \frac{1}{\beta}$

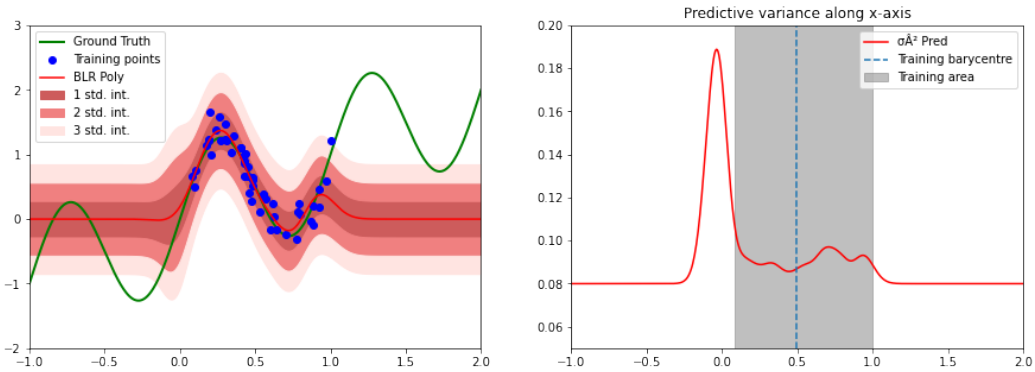
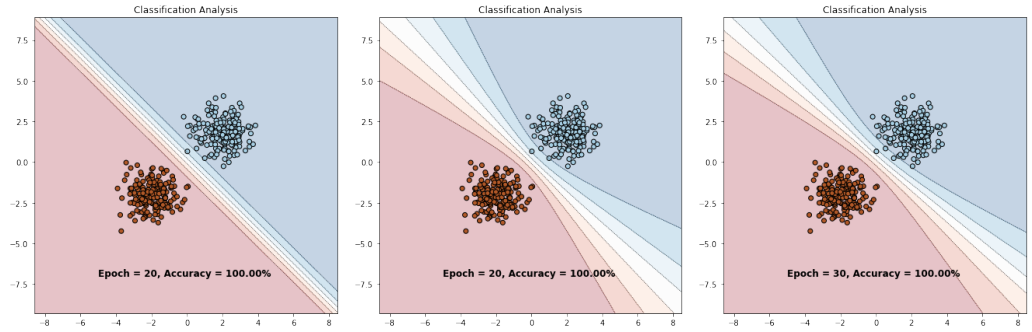


Figure 2: Predictive variance converges far from training samples.

2 Week 2

2.1 Compared to MAP, how does predictive distribution for Laplace approximation behave?

MAP estimate is equivalent to approximating the posterior with $p(\mathrm{w} \mid \mathrm{X}, \mathrm{Y}) \approx \delta(\mathrm{w} - \mathrm{w}_{\mathrm{MAP}})$ which is very coarse. Moreover, uncertainty does not increase far from training data as posterior mean probability only depend on the scalar product with the vector line between the 2 clusters.



(a) MAP estimate (b) Laplace approximation (c) Variational inference

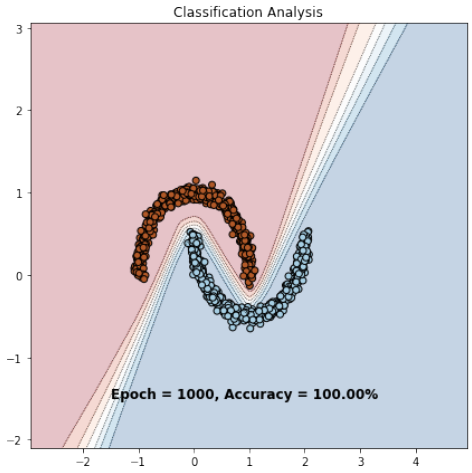
2.2 Comment the class `LinearVariational`. What is the main difference between Laplace's and VI's approximations?

`LinearVariational` class implements mean field approximation of nn.Linear which can similarly be stacked to build Bayesian version of neural networks.

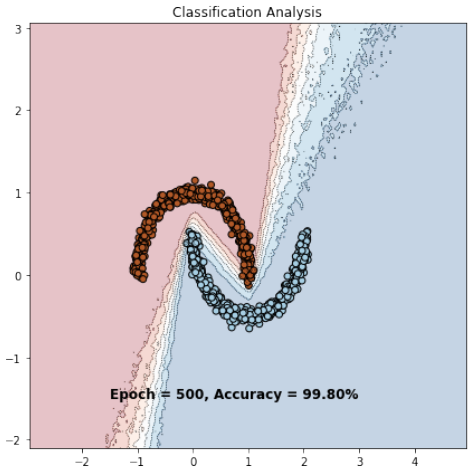
Laplace approximation and Variational Inference mainly differ in the nature of their approximation assumption. Laplace approximation assumes posterior is Gaussian centered on MAP, whereas Variational Inference admits mean field approximation which assumes weights being independently Gaussian, minimizing KL divergence in this family of possible distributions. Thus, there is no notion of best estimate in Laplace approximation contrary to VI where mean can be finely tuned.

2.3 What is the benefit of MC Dropout variational inference over Bayesian Logistic Regression with variational inference?

MC Dropout as the advantage of being extremely simple to implement as no difference with training classical dropout neural network is required. MC dropout admits also less parameters (just weights instead of weight means and weight variances). Finally, the structural hypothesis is less coarse than Variational Inference. It only assumes independence between layers and not independence between all neurons.



(a) Variational Inference

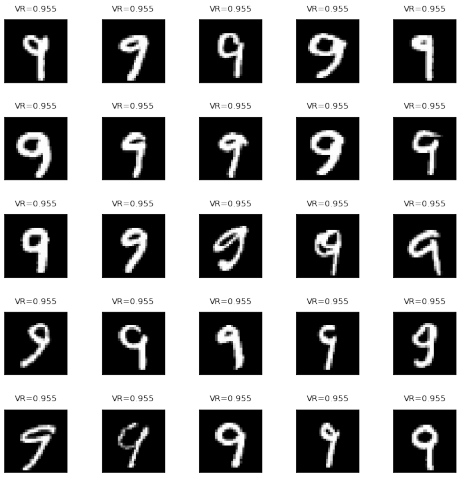


(b) MC Dropout

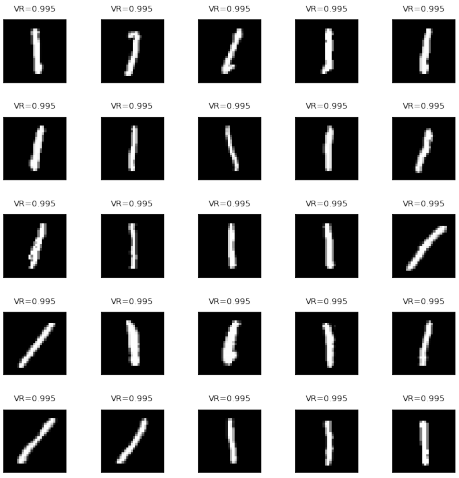
3 Week 3

3.1 Comment results for investigating most uncertain vs. confident samples

We clearly see in Figure 5a that our trainend model is highly confident about predicting 9 MNIST numbers, and in Figure 5b that it is highly uncertain about predicting 1 MNIST numbers.



(a) Confident samples



(b) Uncertain samples

3.2 Failure precition

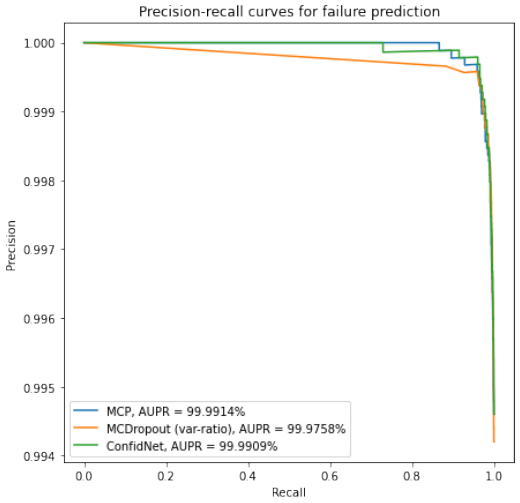
- **Explain the goal of failure prediction**

  Failure prediction aims at detecting if we can rely on a model for a given prediction. The goal is thus to have a calibrated measure of confidence that we can threshold to accept or reject the prediction.

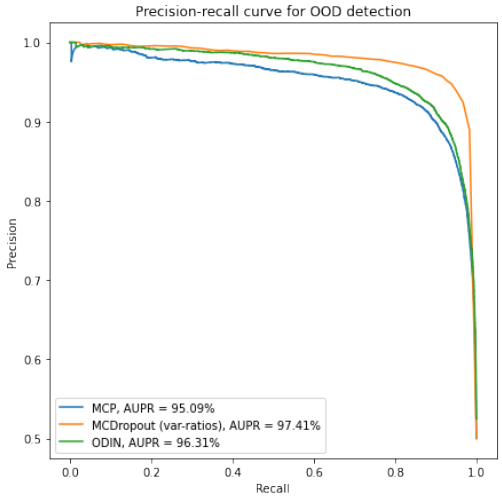
- **Comment the code of the `LeNetConfidNet` class**

  `LeNetConfidNet` consists in a LeNet network to which was added fully connected layers to predict its uncertainty. These are the layers that are trained accordingly to TCP criterion during confidence training.

- **Analyze results between MCP, MCDropout and ConfidNet**
  One can see on Figure 6a that in our tests, MCP performs a little bit better than Confidnet on AUPR metric. They both perform better than MC Dropout.



(a) Failure prediction



(b) Out of Distribution Detection

3.3 OOD detection: analyse results and explain the difference between the 3 methods

One can see on Figure 6b that in our tests, MC Dropout performs a little bit better than ODIN on AUPR metric. They both perform better than MCP.

Maximum Class Probability is the simplest confidence indicator where we fully rely on the network to predict the right class with the right probability.

MC Dropout consists a posteriori sample from our Bayesian dropout network and compute metrics on samples such as variation-ratio.

Finally, ODIN introduce temperature scaling or inverse adversarial perturbation in order to perturb prediction differently on in and out of distribution data. Thus it is a dataset-specific OOD methods.