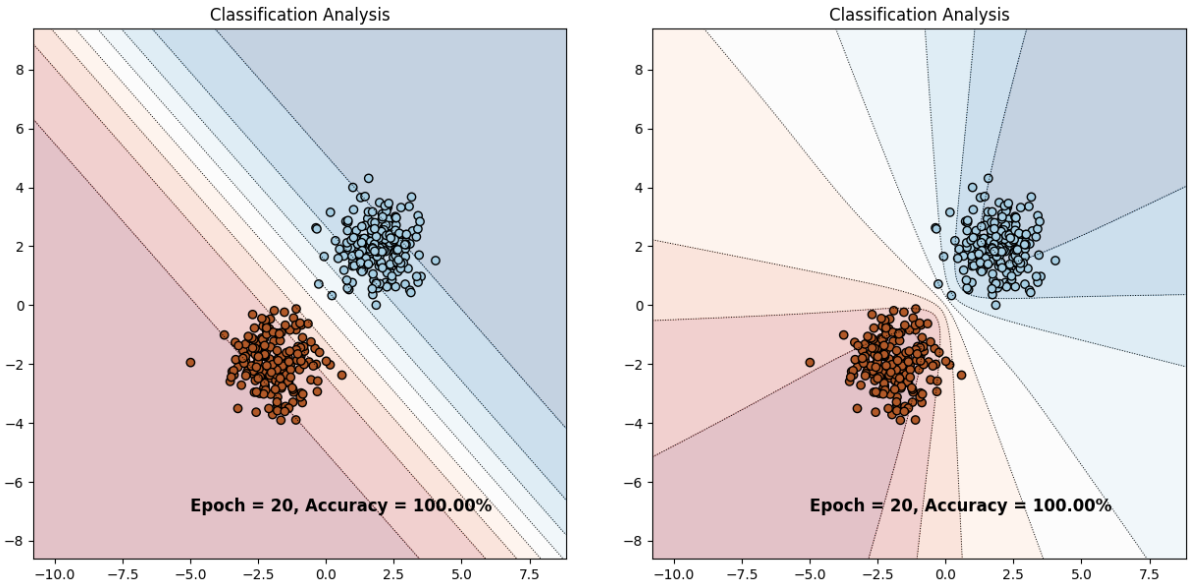
TP2: Approximate Inference in Classification

Fondements Théoriques de l'Apprentissage Profond (MVA)

Aziz Bacha

Laplace's approximation results



(a) MAP (b) Laplace

MAP estimate. For inputs far away from the training distribution (e.g. points such as $(8, 10)$), the model still outputs predictive probabilities very close to 1. This occurs because, under the MAP approximation, the posterior collapses to a single weight vector w_{MAP}, so the predictive probability is

$$p(y = 1 \mid x) = \sigma\left(w_{\text{MAP}}^{\top} x\right).$$

As $\|x\|$ increases, the inner product $w_{\text{MAP}}^{\top} x$ grows in magnitude, causing the sigmoid function to saturate. As a result, the model remains overconfident outside the data region, since no uncertainty is represented when using a point estimate for the weights.

Laplace approximation. In contrast, the Laplace approximation accounts for weight uncertainty by integrating over a Gaussian approximation of the posterior. This results in a smoother, curved decision boundary, rather than the single sharp linear boundary produced by MAP. Moreover, predictive uncertainty increases far from the training data, with predicted probabilities moving toward 0.5, whereas the MAP estimate remains overconfident.

Part I.3 « Variational inference »

Commentary on the `LinearVariational` and `VariationalLogisticRegression` classes.

1. `LinearVariational` layer. This layer implements a Bayesian linear map with a mean-field Gaussian posterior

$$q(w_i) = \mathcal{N}(\mu_i, \sigma_i^2),$$

parameterized by learnable (μ_i, ρ_i), where

$$\sigma_i = \log(1 + e^{\rho_i})$$

ensures $\sigma_i > 0$. During the forward pass, weights are sampled using the reparameterization trick

$$w = \mu + \sigma \odot \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I),$$

which allows gradients to backpropagate through the stochastic sampling. The layer also computes the KL divergence to a Gaussian prior $p(w) = \mathcal{N}(0, \sigma_p^2 I)$:

$$\mathrm{KL}(q(w) \,\|\, p(w)) = \sum_i \left(\log \frac{\sigma_p}{\sigma_i} + \frac{\sigma_i^2 + \mu_i^2}{2\sigma_p^2} - \frac{1}{2} \right).$$

Finally, the sampled weights are used in a standard linear mapping.

2. `VariationalLogisticRegression` model. This model wraps a single `LinearVariational` layer to form Bayesian logistic regression. Each forward pass samples weights and outputs

$$p(y = 1 \mid x, w) = \sigma(w^\top x).$$

The model KL term is the KL divergence returned by the variational layer.

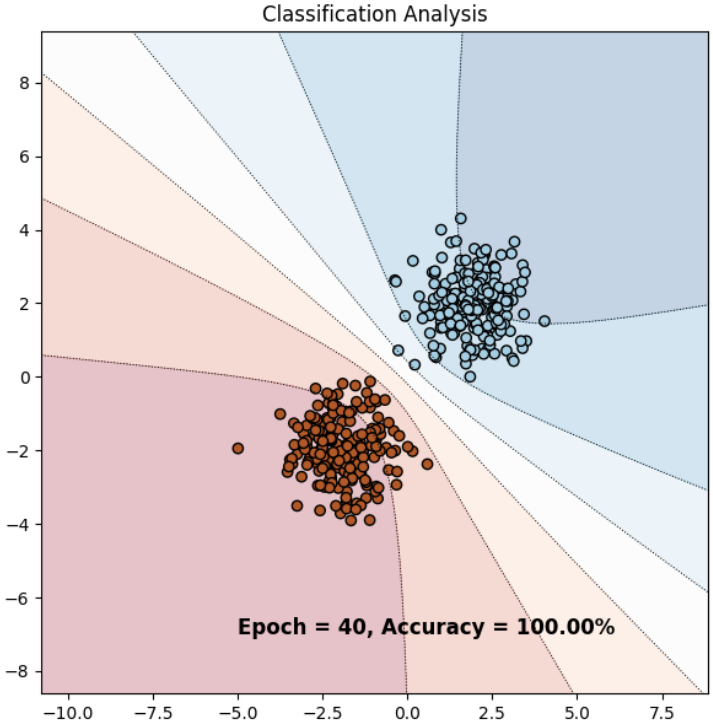


Figure 2: Variational Inference

Loss / ELBO. Since the forward pass samples $w_s \sim q_\theta(w)$, the binary cross-entropy corresponds to a Monte-Carlo estimate of the negative log-likelihood:

$$\mathrm{NLL}(\theta; \mathcal{D}) \approx -\sum_n \log p(y_n \mid x_n, w_s)\,.$$

The regularization term is

$$\mathrm{KL}(q_\theta(w) \,\|\, p(w)),$$

and the optimized objective is the (negative) ELBO:

$$\mathcal{L}_{\mathrm{ELBO}} = \mathrm{NLL} + \mathrm{KL}.$$

Predictive distribution. At test time, predictions are obtained by averaging over multiple weight samples $w_s \sim q_\theta(w)$:

$$\hat{p}(y = 1 \mid x) \approx \frac{1}{S} \sum_{s=1}^{S} p(y = 1 \mid x, w_s).$$

Behaviour vs. MAP estimate. Compared to the deterministic MAP model (single w_{MAP}), the variational model averages over $w \sim q_\theta(w)$, yielding a smoother decision boundary and reduced overconfidence. Far from the data, epistemic uncertainty is higher (probabilities closer to 0.5 and a fuzzier boundary) than for MAP.

Variational vs. Laplace approximation. The Laplace approximation uses a local Gaussian around the MAP solution,

$$q_{\text{Lap}}(w) = \mathcal{N}\left(w \mid \mu_{\text{MAP}}, H^{-1}\right),$$

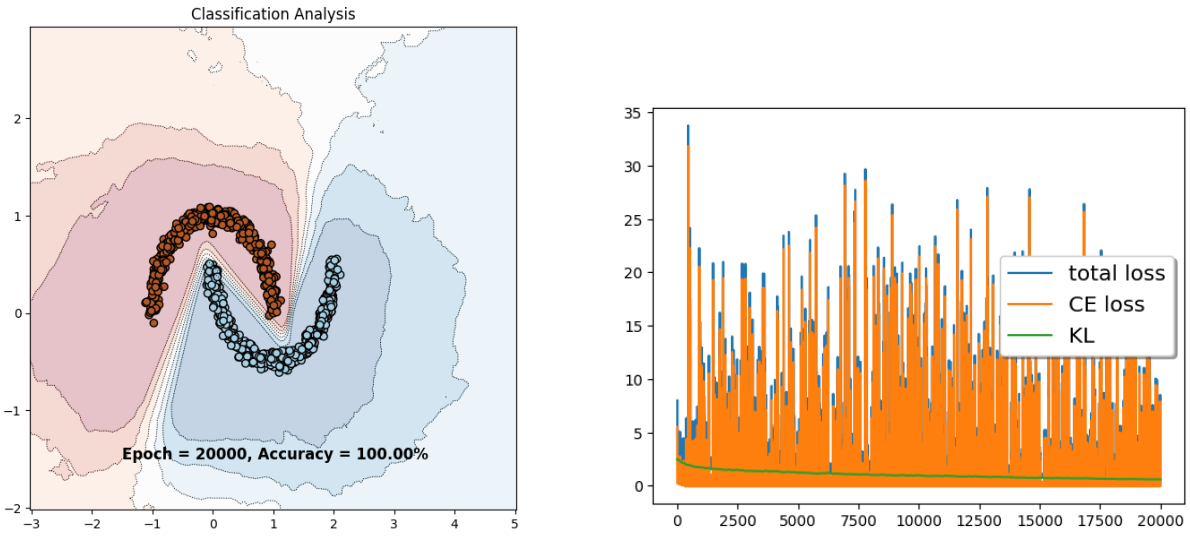
where H is the Hessian of the negative log-posterior at w_{MAP}. In contrast, the mean-field variational approximation optimizes a factorized Gaussian

$$q_\theta(w) = \prod_i \mathcal{N}(w_i \mid \mu_i, \sigma_i^2)$$

by maximizing the ELBO with SGD and the reparameterization trick, making it a global optimization within the mean-field family (not tied to w_{MAP}).

Bayesian Neural Networks

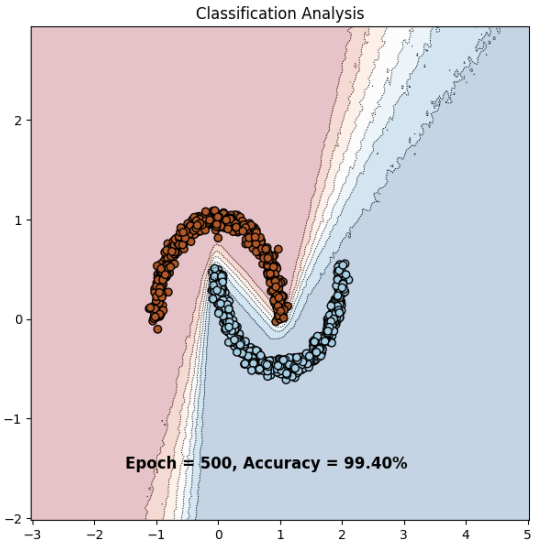
Variational Inference with Bayesian Neural Networks



(a) (b)

- The predictive distribution is uncertain far from the data: near the data, the classifier is confident and accurate, but away from the training region, predictions do not collapse to 0 or 1, showing proper epistemic uncertainty.
- The KL term decreases slowly, meaning the posterior distribution becomes more concentrated as training progresses. The CE term fluctuates strongly because each forward pass samples new weights, making the loss noisy but unbiased. Despite this noise, the total loss trends downward and the model converges.

0.1 Monte Carlo Dropout



Bayesian Logistic Regression can only learn a linear decision boundary and therefore cannot model complex datasets. MC Dropout produces richer non-linear boundaries, better uncertainty estimates, and scales easily to high-dimensional models without the heavy computations that BLR requires.