

# Bayesian deep learning

## Bayesian inference for deep neural networks :)

*June 26, 2022*

*As I read papers in more details, I will add more details to the slides. Slides for today are created based on Part 1 and 3 of the 2020 ICML tutorial on bayesian deep learning.*

In many machine learning applications, we care a lot of about predicting a quantity. This includes both classification and regression problems.

This is most naturally\* captured by the following integral:

$$p(y|x, D) = \int p(y|x, \theta) p(\theta|D) d\theta$$

where  $p(\theta|D)$  is the posterior over the parameter vector given observed data  $D = \{(x^{(i)}, y^{(i)})\}$ .

**Interpretation.** Rather than finding one solution, average across several solutions

Other names: **marginalization** over  $\theta$

$$p(y|x, D) = \int p(y, \theta|x, D) d\theta$$

If we want to approximate the posterior by **a point estimate**, then one can argue that we should use the MAP. When the prior is uniform, then we have the MLE.

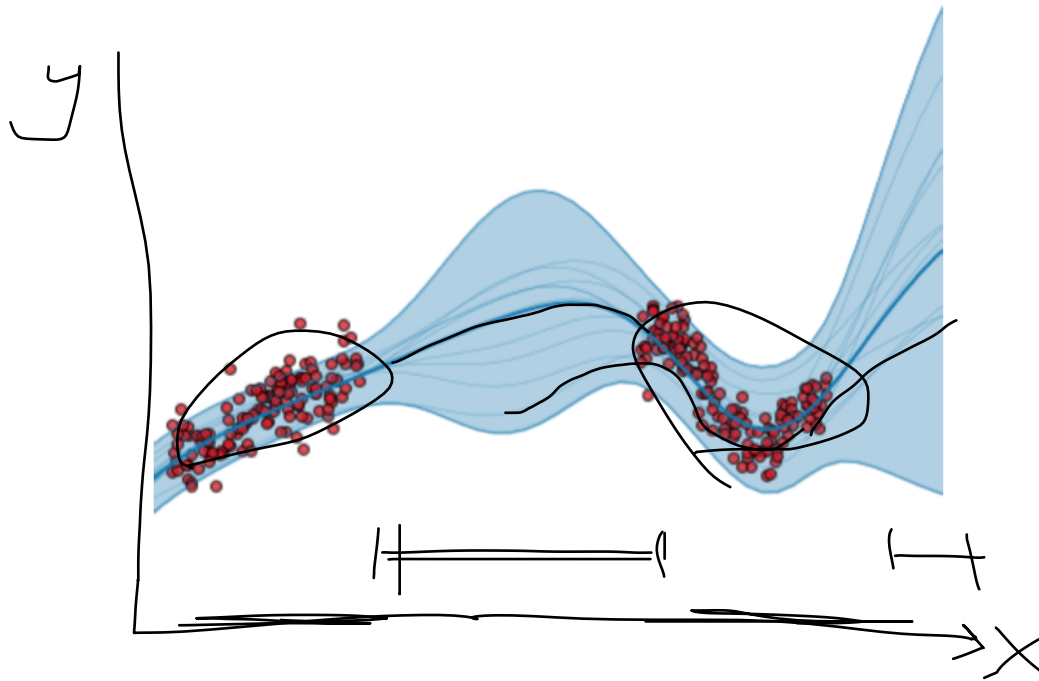
$$p(\theta | D) = \delta_{\hat{\theta}_{\text{MAP}}}(\theta) = \begin{cases} \infty & \text{if } \theta = \hat{\theta}_{\text{MAP}} \\ 0 & \text{otherwise} \end{cases}$$

If we do this approximation, then the posterior predictive becomes

$$\begin{aligned} p(y | x, D) &= \int p(y | x, \theta) p(\theta | D) d\theta \\ &= \int p(y | x, \theta) \delta(\hat{\theta}_{\text{MAP}}) d\theta \\ &= \underline{p(y | x, \hat{\theta}_{\text{MAP}})}. \end{aligned}$$

**Posterior is diffused**, so point estimation would give different result (in terms of the predictive distribution) compared to doing the integral (which is arguably the correct thing to do?)

- Otherwise, underestimation of epistemic uncertainty
- High variance problem (“variance” as in bias-variance tradeoff) → bad generalization



Also: Architecture agnostic! Empirically leads to *better* performance than MLE training!



Challenge:

- Lots of parameters (e.g., intractable for standard MCMC methods) (i.e.,  $|\theta|$  is large)

Rough idea:

- Learn approximate posterior
- Efficiently generate samples from the posterior

By year

- ICML 2015: Bayes by Backprop: factored Gaussian prior over parameters; finding approximate posterior by optimizing evidence lower bound using SGD

does dropout induce more bias

- ICML 2016: MC Dropout: similar to ensemble? average across dropout

- NeurIPS 2017: Deep Ensembles: retrain model multiple times to find different SGD solutions

- UAI 2018: Subspace inference: approximate posterior in lower dimensional space

- NeurIPS 2019: Stochastic Weighted Average (Gaussian): model SGD iterates

- 11, 14, 20: Stochastic MCMC

few-shot context