Approximate Inference Turns Deep Networks into Gaussian Processes, Khan et al. (2019)

P. Clavier, E. Cohen, J. Linhart

Bayesian Machine Learning, MVA 2020/2021
March 24, 2021

école normale supérieure paris–saclay
université PARIS-SACLAY

Motivation: DNNs vs. GPs

Two powerful and complementary ML-models:

Scalability (SGD) vs. Interpretability (closed form)

Relation: GPs seen as *infinitly wide* DNNs, Neal (1997)
→ How to make DNNs behave like GPs in practice?

Bayesian DNNs: Learn a distribution (not a point estimate):

$$p(y_*|x_*, \mathcal{D}) = \int \underbrace{p(y_*|\mathbf{w}, x_*)}_{\text{likelihood}} \underbrace{p(\mathbf{w}|\mathcal{D})}_{\text{posterior}} d\mathbf{w} \tag{1}$$

Posterior approximation:

- MC-sampling
- Gaussian Approximation (Laplace or VI)

Contribution: Relating Bayesian DNN posteriors to GPs!

DNN2GP Framework, Khan et al. (2019)

Gaussian approximation → Find a linear model → Find a GP

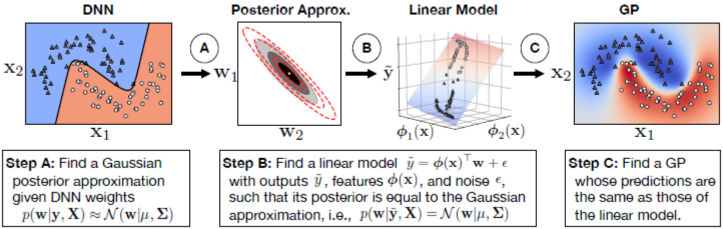
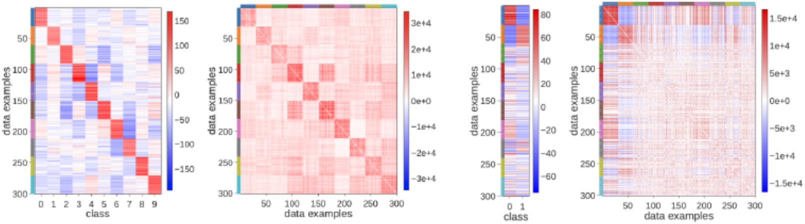


Figure 1: Turning a DNN into a GP, Khan et al. (2019)

- **Laplace Approximation**: behavior *after* training (DNN2GP)
- **VI Approximation**: behavior *during* training (VOGGN)
- Use Generalised Gauss-Newton (GGN) Approximation
- Covariance Matrix of the GP $\simeq$ **NTK** (Jacot et al. (2020))

Experimental Results on MINST



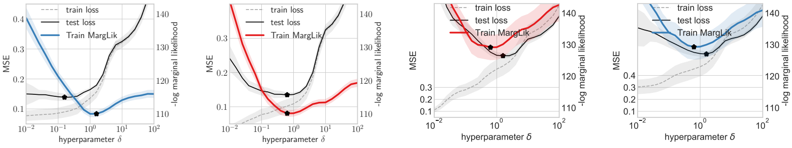
(a) MNIST GP (left: Posterior mean; right: Kernel matrix)

(b) Binary-MNIST (left: Posterior mean; right: Kernel matrix)

- **DNN2GP**-GP reflects the behavior of trained DNN, Fig (a)
- **VOGN** trains Bayesian DNN capable of estimating its uncertainty, Fig (b)

Experimental Results - Ours

- The GP marginal likelihood is very similar for training data obtained with both, Laplace and VI approximations
- Its evolution is close to the test loss behavior
- The chosen hyperparameters are not exactly equal to the ones according to the test-loss



(a) Laplace Paper (b) VI Paper (c) Laplace Ours (d) VI Ours

Figure 2: Model fit on higher variance simulated data, and hyperparameter optimization δ

Conclusion and Perspectives

Take away: Bayesian NNs can be obtained by VOGN and the GP constructed via DNN2GP correctly reflects its behavior!

Limitations:

- Better result analysis with more difficult data
- GGN-approximation (expensive computation)
- VI for deep architectures → SWAG, Maddox et al. (2019)

GP relation: Interpretation and vizualisation of DNN behavior

- GP kernel defines a known function sapce
- Analyse **NTK-RKHS** for smoothness and stability properties, Fort et al. (2020)
- Regularize DNN with RKHS-norm, Bietti et al. (2019)

References I

Alberto Bietti, Grégoire Mialon, Dexiong Chen, and Julien Mairal. A kernel perspective for regularizing deep neural networks. pages 664–674, 2019.

Stanislav Fort, Gintare Karolina Dziugaite, Mansheej Paul, Sepideh Kharaghani, Daniel M Roy, and Surya Ganguli. Deep learning versus kernel learning: an empirical study of loss landscape geometry and the time evolution of the neural tangent kernel. *arXiv preprint arXiv:2010.15110*, 2020.

Arthur Jacot, Franck Gabriel, and Hongler Clementj. Neural tangent kernel: Convergence and generalization in neural networks. *arXiv:1806.07572v4*, 2020.

References II

Mohammad Emtiyaz Khan, Alexander Immer, Ehsan Abedi, and Maciej Korzepa. Approximate inference turns deep networks into gaussian processes. *arXiv preprint arXiv:1906.01930*, 2019.

Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning. pages 13153–13164, 2019.

Radford M. Neal. *Bayesian Learning for Neural Networks*. 1997.