DTU Compute
Department of Applied Mathematics and Computer Science

Approximate Inference for Bayesian Neural Networks

Silas Brack (s174433)

Kongens Lyngby 2022

Summary

The first move is optional and contains general background information about your key research variable or variables. The second move is the statement of the problem indicating your research hypothesis/ question. The third move is the methodology move representing your participants, if any ; your materials/instruments, procedures, and data analysis. The fourth move portrays your findings. Finally, in the last optional move, you talk about the likely implications of the study. (Maybe 2x this in terms of size.)
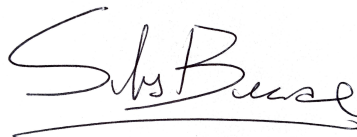
Preface

This Master's thesis was prepared at the department of Applied Mathematics and Computer Science at the Technical University of Denmark in fulfillment of the requirements for acquiring a Master's degree in Mathematical Modelling and Computation.

Kongens Lyngby, August 9, 2022

Silas Brack (s174433)

Acknowledgements

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Contents

Contents

List of Figures

List of Tables

Introduction

1.1 The problem

1.1.1 What is the problem?

1.1.2 Why is it interesting and important? (Motivation)

1.2 Current state of research in the area

1.2.1 Why is it hard? (E.g. why do naïve approaches fail?)

1.2.2 Why hasn't it been solved before? (Or, what's wrong with previous proposed solutions?)

1.3 My solution

1.3.1 How does my solution differ?

The ideas presented in this article essentially constitute a compilation of advice given by academics from multiple fields and institutions,

1.3.2 What are the key components of my approach and results? Are there any specific limitations?

1.4 Research objectives

1.4.1 What are the goals for this project?

This should be discussed in a bit more detail in a thesis, since there are certain objectives discussed in the project plan. In an article, a quick sentence which summarises the rest of the introduction and states succinctly exactly which problem will be tackled and which approach will be made should suffice.

CHAPTER 2

Literature Review

2.1 Baselines

2.1.1 Variational Adaptive-Newton

[Kha+17]

2.1.2 Noisy Natural Gradient

[Zha+18]

Theory

3.1 Bayes' Theorem

Bayes' theorem describes the probability of an event occurring with respect to prior knowledge of that event. Specifically for modelling, it can be used to relate latent variables with observed variables, allowing distributions on these latent variables to be inferred. We therefore want to find the posterior $p(\boldsymbol{\theta} \,|\, \boldsymbol{y})$ by According to Bayes' theorem, the posterior is given by

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{y}) = \frac{p(\boldsymbol{y} \,|\, \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})} \tag{3.1}$$?eq:bayes?

where $p(\boldsymbol{y} \,|\, \boldsymbol{\theta})$ is the likelihood, parameterised by the model parameters $\boldsymbol{\theta}$, $p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}; \boldsymbol{\alpha})$ is the prior, parameterised by its hyper-parameters $\boldsymbol{\alpha}$, and $p(\boldsymbol{y}) = p_{\boldsymbol{\theta}}(\boldsymbol{y})$ is the marginal likelihood (or evidence).

Furthermore, the posterior predictive probability, i.e., the probability of observing some new data, is given by

$$p(\boldsymbol{y}_{\text{new}} \,|\, \boldsymbol{y}) = \int p(\boldsymbol{y}_{\text{new}} \,|\, \boldsymbol{\theta})p(\boldsymbol{\theta} \,|\, \boldsymbol{y}) \, d\boldsymbol{\theta}. \tag{3.2}$$ {?}

3.2 Variational Inference

In variational inference (VI), we minimise the KL divergence between the posterior distribution $p(\boldsymbol{\theta} \,|\, \boldsymbol{y})$ and a variational distribution $q_{\boldsymbol{\phi}}(\boldsymbol{\theta})$ parameterised by its variational parameters $\boldsymbol{\phi}$ by optimising with respect to these parameters. The KL divergence is defined as

$$\begin{aligned} D_{\text{KL}}\left[q(\boldsymbol{\theta}) \,\|\, p(\boldsymbol{y} \,|\, \boldsymbol{\theta})\right] &\equiv \mathbb{E}_q\left[\log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{y} \,|\, \boldsymbol{\theta})}\right] \\ &= \mathbb{E}_q\left[\log q(\boldsymbol{\theta})\right] - \mathbb{E}_q\left[\log p(\boldsymbol{y} \,|\, \boldsymbol{\theta})\right] \end{aligned} \tag{3.3}$$?eq:kld?

However, the KL divergence contains the evidence term $\log p(\boldsymbol{y})$, which is intractable. Instead, we define a lower bound on this marginal likelihood term that is surrogate to the KL divergence (known as the ELBO). Since they are surrogate, ELBO fulfils

the property $\mathcal{L}[q_\phi] = \log p(\boldsymbol{y}) - D_{\mathrm{KL}}[q(\boldsymbol{\theta}) \parallel p(\boldsymbol{y} \mid \boldsymbol{\theta})]$, minimising the KL divergence is equivalent to maximising the ELBO, which is defined as

$$
\begin{aligned}
\mathcal{L}[q_\phi] &\equiv \mathbb{E}_q[\log p(\boldsymbol{y}, \boldsymbol{\theta})] - \mathbb{E}_q[\log q(\boldsymbol{\theta})] \\
&= \mathbb{E}_q[\log p(\boldsymbol{y} \mid \boldsymbol{\theta})] + \mathbb{E}_q[\log p(\boldsymbol{\theta})] - \mathbb{E}_q[\log q(\boldsymbol{\theta})] \\
&= \mathbb{E}_q[\log p(\boldsymbol{y} \mid \boldsymbol{\theta})] - \underbrace{-\mathbb{E}_q[\log p(\boldsymbol{\theta})]}_{\text{Cross-entropy}} + \underbrace{-\mathbb{E}_q[\log q(\boldsymbol{\theta})]}_{\text{Entropy}}
\end{aligned}
\qquad (3.4)\,\texttt{?eq:elbo?}
$$

where the first term $\mathbb{E}_q[\log p(\boldsymbol{y} \mid \boldsymbol{\theta})]$ is the data (or likelihood) term, $-\mathbb{E}_q[\log p(\boldsymbol{\theta})]$ is the cross-entropy of the prior with respect to the variational approximation, and $-\mathbb{E}_q[\log q(\boldsymbol{\theta})]$ is the entropy term of the variational approximation. The last two terms can be interpreted as the regularising KL divergence (or relative entropy) from the prior to the variational approximation, $D_{\mathrm{KL}}[q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta})] = \mathbb{E}_q[\log q(\boldsymbol{\theta})] - \mathbb{E}_q[\log p(\boldsymbol{\theta})]$. Furthermore, the likelihood term in the ELBO can be decomposed further by assuming independence in the observations, calculated as the term given by $\mathbb{E}_q[\log p(\boldsymbol{y} \mid \boldsymbol{\theta})] = \frac{1}{N}\sum_i^N \mathbb{E}_q[\log p(y_i \mid \boldsymbol{\theta})]$. Overall, BNNs have been found to be relatively ineffective unless the number of observations is greater than the number of model parameters.

3.3   Methods

Multiple approximate inference methods were implemented. Specifically, *maximum a posteriori* estimation, Laplace approximation, VI with mean-field, full-rank, low-rank and radial variational families, deep ensembles, and MultiSWAG. These methods will be described in the following sections.

3.3.1   Maximum a Posteriori Estimation

Maximum a posteriori estimation finds a point estimate of the posterior given by its maximum. It can therefore be interpreted as a Delta distribution estimate of the posterior $p(\boldsymbol{\theta} \mid \boldsymbol{y}) \approx \delta(\boldsymbol{\theta}_{\mathrm{MAP}})$, where $\boldsymbol{\theta}_{\mathrm{MAP}}$ is defined as

$$
\begin{aligned}
\boldsymbol{\theta}_{\mathrm{MAP}} &= \max_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid \boldsymbol{y}) \\
&= \max_{\boldsymbol{\theta}} \log p(\boldsymbol{\theta} \mid \boldsymbol{y}) \\
&= \max_{\boldsymbol{\theta}} [\log p(\boldsymbol{y} \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta})] \\
&= \max_{\boldsymbol{\theta}} \left[\sum_{i=1}^{N} \log p(y_i \mid \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \right].
\end{aligned}
\qquad (3.5)\,\texttt{\{?\}}
$$

As an optimisation problem, maximising the posterior corresponds to minimising the regularised loss,

$$\boldsymbol{\theta}_{\mathrm{MAP}} = \arg\min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{y}; \boldsymbol{\theta}) \tag{3.6}\{?\}$$

$$= \arg\min_{\boldsymbol{\theta}} \left[-\sum_{i=1}^{N} \log p(y_i \,|\, \boldsymbol{\theta}) - \log p(\boldsymbol{\theta}) \right]. \tag{3.7}\{?\}$$

In this formulation, the term $-\sum_{i=1}^{N} \log p(y_i \,|\, \boldsymbol{\theta})$ is known as the empirical loss or reconstruction loss and the term $\log p(\boldsymbol{\theta})$ is the regulariser.

3.3.2 Laplace Approximation

In the Laplace approximation (LA) [Dax+21], the posterior is approximated by a Gaussian, similarly to mean-field VI. However, instead of finding the optimal Gaussian distribution locations and scales by maximising the ELBO, the LA finds the location by computing the MAP solution $\boldsymbol{\theta}_{\mathrm{MAP}}$ and the scale by approximating the log posterior with a second degree Taylor expansion around this solution ($\boldsymbol{\theta}_0 = \boldsymbol{\theta}_{\mathrm{MAP}}$) and determining the curvature via its Hessian matrix $\boldsymbol{H} = \nabla_{\boldsymbol{\theta}}^2 \log p(\boldsymbol{\theta} \,|\, \boldsymbol{y}) \big|_{\boldsymbol{\theta}_{\mathrm{MAP}}}$. Since the Taylor expansion is performed around the MAP solution, the first order derivative is zero, and the expansion is simply given by

$$\ln p(\boldsymbol{\theta} \,|\, \boldsymbol{y}) \approx \ln p(\boldsymbol{\theta}_0 \,|\, \boldsymbol{y}) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\mathsf{T}} \boldsymbol{H}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \tag{3.8}?\underline{\texttt{eq:laplace-t}}$$

$$\tilde{p}(\boldsymbol{\theta} \,|\, \boldsymbol{y}) \approx p(\boldsymbol{\theta}_0 \,|\, \boldsymbol{y}) \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\mathsf{T}} \boldsymbol{H}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \right).$$

Normalising this unnormalised posterior yields the Laplace approximation of the posterior

$$p(\boldsymbol{\theta} \,|\, \boldsymbol{y}) \approx \sqrt{\frac{\det \boldsymbol{H}}{(2\pi)^D}} \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\mathsf{T}} \boldsymbol{H}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \right)$$

$$\approx \mathcal{N}(\boldsymbol{\theta}_0, \boldsymbol{H}^{-1}) = \mathcal{N}(\boldsymbol{\theta}_{\mathrm{MAP}}, \boldsymbol{H}^{-1}). \tag{3.9}?\underline{\texttt{eq:laplace?}}$$

3.3.3 Mean-Field Variational Approximation

?⟨ssec:mfvi⟩? One common variational family used to approximate the real posterior is a product of independent Gaussian distributions (the mean-field approximation) such that each model parameter is sampled from a normal distribution and is independent of all other model parameters, yielding a Gaussian with a diagonal covariance matrix. In mean-field VI, model weights are sampled from an approximate posterior $q(\boldsymbol{\theta}) \approx p(\boldsymbol{\theta} \,|\, \boldsymbol{y})$ as

$$q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}) \Leftrightarrow$$
$$\boldsymbol{\theta} = \boldsymbol{\mu} + \boldsymbol{\sigma} \circ \boldsymbol{\epsilon} \tag{3.10}?\underline{\texttt{eq:meanfield}}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. For this type of variational approximation, the entropy term is given by $\mathbb{H}\left[q(\theta)\right] = -\sum_i \log \sigma_i$ and the cross-entropy of the prior relative to the variational approximation is calculated via Monte Carlo simulation by taking the prior log probability with respect to mean-field posterior samples as

$$
\begin{aligned}
\mathbb{H}\left[q(\boldsymbol{\theta}), p(\boldsymbol{\theta})\right] &= -\int q(\boldsymbol{\theta}) \log(p(\boldsymbol{\theta}))\, d\boldsymbol{\theta} \\
&\approx \frac{1}{S} \sum_{s=1}^{S} \log p(\boldsymbol{\theta}^{(s)})
\end{aligned}
\tag{3.11}\{?\}
$$

where $\boldsymbol{\theta}^{(s)} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$.

The use of the mean-field approximation for BNNs has been found to be unreliable [Wu+18] and, for increasingly wide neural networks, converges towards the prior [CPD21]. However, for *deep* BNNs, the mean-field assumption may be reasonable [FSG20].

3.3.4   Full-Rank Variational Approximation

As an alternative to the mean-field approximation, a full-rank approximation doesn't assume independence of the model parameters, instead being sampled as

$$
q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\mathrm{FR}}).
\tag{3.12}\text{?eq:fullrank?}
$$

For D model parameters, while a mean-field approximation has scale parameters $\boldsymbol{\sigma}^D$, the full-rank approximation has scale parameters $\boldsymbol{\Sigma}_{\mathrm{FR}}^{D \times D}$. The variational parameters scale quadratically with the number of model parameters, which constitutes a significant limitation of this type of approximation. Specifically, since NNs are typically overparameterised, possessing from thousands to millions to billions of parameters, the quadratic number of variational parameters makes this type of VI infeasible for BNNs. As such, full-rank VI is not applied in any of the experiments in this paper.

3.3.5   Low-Rank Variational Approximation

As a compromise between the last two methods, a low-rank variational approximation uses a low-rank approximation of the covariance matrix $\boldsymbol{\Sigma}_{\mathrm{LR}}$.

$$
q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}_{\mathrm{LR}}).
\tag{3.13}\text{?eq:lowrank?}
$$

This approximation is obtained by reconstructing a covariance matrix paremeterised by the covariance factor $\tilde{\boldsymbol{\Sigma}}$ and diagonal element vector $\boldsymbol{\sigma}_{\mathrm{diag}}$ as

$$
\begin{aligned}
\hat{\boldsymbol{\Sigma}}_{\mathrm{LR}} &= \tilde{\boldsymbol{\Sigma}} \tilde{\boldsymbol{\Sigma}}^{\mathsf{T}} + \boldsymbol{\sigma}_{\mathrm{diag}} \\
\boldsymbol{\Sigma}_{\mathrm{LR}} &= \min_{\hat{\boldsymbol{\Sigma}}_{\mathrm{LR}}} \left\| \boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}_{\mathrm{LR}} \right\|_F
\end{aligned}
\tag{3.14}\text{?eq:lowrank-covar}
$$

3.3.6 Radial Variational Approximation

Farquhar, Osborne, and Gal [FOG20] introduces the Radial approximate posterior, For each NN layer, this posterior is sampled similarly to the mean-field ((3.10)), but by normalising the $\boldsymbol{\epsilon}$ term (projecting it onto a hypersphere) yielding a direction term $\boldsymbol{\epsilon}/\|\boldsymbol{\epsilon}\|$, and scaling it by the distance r. Therefore, for each layer, the radial posterior is sampled as

$$q(\boldsymbol{\theta}) = \text{Radial}(\boldsymbol{\mu}, \boldsymbol{\sigma})$$
$$\boldsymbol{\theta} = \boldsymbol{\mu} + \boldsymbol{\sigma} \circ \frac{\boldsymbol{\epsilon}}{\|\boldsymbol{\epsilon}\|} r \qquad (3.15)\,\text{?eq:radial?}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, $r \sim \mathcal{N}(0,1)$. The entropy term of the Radial approximate posterior is given $\mathbb{H}\left[q(\theta)\right] = -\sum_i \log \sigma_i + \text{const}$ (and is therefore approximately equal to the entropy of the mean-field approximation up to a constant) and the cross-entropy term is calculated via Monte Carlo simulation as for mean-field VI (Section 3.3.3), but by sampling from a radial posterior $\boldsymbol{\theta}^{(s)} \sim \text{Radial}(\boldsymbol{\mu}, \boldsymbol{\sigma})$ as in (3.15).

3.3.7 Deep Ensembles

In deep ensembles [LPB17], M neural networks are trained with different initialisations. In this way, multiple local MAP solutions $\boldsymbol{\theta}^{(m)}$ are obtained and equally considered, such that $p(\boldsymbol{\theta} = \boldsymbol{\theta}^{(m)} \mid \boldsymbol{y}) = 1/M$ for every $\boldsymbol{\theta}^{(m)} \in \{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(M)}\}$ (and 0 elsewhere). To make predictions with a deep ensemble, we use the posterior predictive given by

$$p(\boldsymbol{y}_{\text{new}} \mid \boldsymbol{y}) = \int p(\boldsymbol{y}_{\text{new}} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{y}) \, d\boldsymbol{\theta}$$
$$= \frac{1}{M} \sum_{m=1}^{M} p(\boldsymbol{y}_{\text{new}} \mid \boldsymbol{\theta}^{(m)}) \qquad (3.16)\,\{?\}$$

where $p(\boldsymbol{y}_{\text{new}} \mid \boldsymbol{\theta}^{(m)})$ are the normalised logits, or predicted probabilities, from model m in the ensemble.

3.3.8 MultiSWAG

Stochastic weight averaging estimates the final point estimate of the weights of a neural network $\boldsymbol{\theta}_{\text{SWA}}$ as the average of the model parameters $\boldsymbol{\theta}$ at the end of each epoch obtained via gradient descent after convergence has been reached.

For SWAG [Mad+19], the standard deviation of the model parameters $\boldsymbol{\sigma}_{\text{SWA}}$ is also calculated. From the SWA estimates of the model parameter mean and standard deviations, the posterior is then estimated as

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) \approx \mathcal{N}(\boldsymbol{\theta}_{\text{SWA}}, \boldsymbol{\sigma}_{\text{SWA}}). \qquad (3.17)\,\text{?eq:swag?}$$

For MultiSWAG [WI20], M SWAG estimates are obtained from different initialisations, as in deep ensembles. Then, a Gaussian mixture model (GMM) is formed as a combination of each Gaussian SWAG posterior estimate $\mathcal{N}(\boldsymbol{\theta}_{\text{SWA}}^{(m)}, \boldsymbol{\sigma}_{\text{SWA}}^{(m)})$ where each component is equally weighed, as

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) \approx \frac{1}{M} \sum_{m=1}^{M} \mathcal{N}(\boldsymbol{\theta}_{\text{SWA}}^{(m)}, \boldsymbol{\sigma}_{\text{SWA}}^{(m)}). \qquad (3.18)\, \texttt{?eq:multiswag?}$$

Algorithm 1 Modified mini-batch K-means

`?edminibatch⟩?`

1: Given: K, mini-batch size B, iterations T, dataset X, correlation matrix P.
2: Initialize $C = \{\mathbf{c}^{(1)}, \mathbf{c}^{(2)}, \ldots, \mathbf{c}^{(K)}\}$ with random $\mathbf{x}$'es picked from X.
3: $A \leftarrow B \cdot T$ sorted random indexes to X, denoted $a_1, a_2, \ldots, a_{B \cdot T}$.
4: $X' \leftarrow \{\mathbf{x}^{(a_1)}, \mathbf{x}^{(a_2)}, \ldots, \mathbf{x}^{(a_{B \cdot T})}\}$ ▷ Cache all points
5: $\mathbf{size} \leftarrow 0$
6: **for** $i = 1$ to T **do**
7:       $M \leftarrow B$ examples picked randomly from X'
8:       **for** $\mathbf{x} \in M$ **do**   ▷ *Assignment step*
9:           $\mathbf{d}[\mathbf{x}] \leftarrow f(C, \mathbf{x}, \text{P})$ ▷ Cache closest center
10:      **end for**
11:      **for** $\mathbf{x} \in M$ **do**   ▷ *Update step*
12:          $\mathbf{c} \leftarrow \mathbf{d}[\mathbf{x}]$ ▷ Get cached center for current $\mathbf{x}$
13:          $\mathbf{size}[\mathbf{c}] \leftarrow \mathbf{size}[\mathbf{c}] + 1$ ▷ Update cluster size
14:          $\eta \leftarrow \frac{1}{\mathbf{size}[\mathbf{c}]}$ ▷ Get learning rate
15:          $\mathbf{c} \leftarrow (1 - \eta)\mathbf{c} + \eta \mathbf{x}$ ▷ Take gradient step
16:      **end for**
17: **end for**
18: **return** C, $\mathbf{size}$

3.4   Bayesian ADAM: Bayesian Adaptive Moment Estimation

[Kha+17; Zha+18; Mia+22]

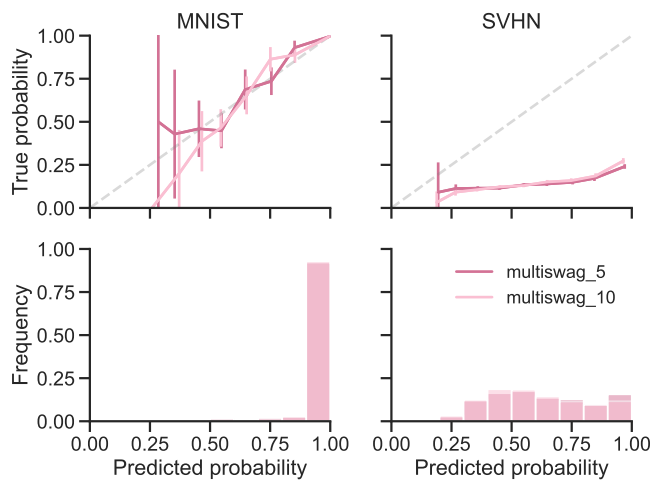## 3.5   Kronecker-Factored Approximate Curvature

[MG15]

Method

Results

Figure 5.1: In-distribution calibration curves for various variational families with variational inference and Laplace approximation trained on MURA..

⟨fig:mura-cc-vi⟩?

Table 5.1: ELBO and $\hat{k}$-statistic for variational inference using different variational families trained on the eight schools test problem, with three replicates for normalizing flows..

?⟨tab:poi⟩?

Type	Flows	$\mathcal{L}\left[q_\phi\right]$		$\hat{k}$		Params
		μ	σ	μ	σ	
Mean-field	—	-33.41	—	0.87	—	20
Full-rank	—	-32.60	—	0.82	—	110
Planar	4	-33.31	0.02	0.84	0.01	85
	8	-32.93	0.15	0.80	1e-3	169
	16	-32.25	0.11	0.73	0.03	337
	32	-31.84	0.06	0.65	0.05	673
Radial	4	-37.73	0.01	0.90	0.01	58
	8	-37.66	0.02	0.89	0.01	106
	16	-34.46	1.72	0.84	0.07	202
	32	-32.36	0.23	0.74	0.07	394
IAF	4	-31.53	0.09	0.57	0.03	63114
	8	-31.53	0.01	0.59	0.05	125918
	16	-31.61	0.06	0.54	0.01	251526
	32	-31.67	0.03	0.50	0.02	502742

CHAPTER 6

Experiments

CHAPTER 7

Conclusion

7.1   Conclusion and Outlook

7.1.1   Summary and key results

Remember to actually include some results and a quantitative look at the results that were obtained.

7.1.2   Outlook and future developments

Which avenues are most promising for future research? Either in order to solve the problem this paper tables or to solve the next problem.

APPENDIX A

An Appendix

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Bibliography

[Bin+18]   Eli Bingham et al. "Pyro: Deep Universal Probabilistic Programming."
            In: *Journal of Machine Learning Research* (2018).

[BN06]   Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and
            machine learning*. Volume 4. 4. Springer, 2006, pages 165–172.

[BS09]   José M Bernardo and Adrian FM Smith. *Bayesian theory*. Volume 405.
            John Wiley & Sons, 2009.

[BZG18]   Juhan Bae, Guodong Zhang, and Roger Grosse. "Eigenvalue corrected
            noisy natural gradient." In: *arXiv preprint arXiv:1811.12565* (2018).

[CGJ96]   David A Cohn, Zoubin Ghahramani, and Michael I Jordan. "Active learn-
            ing with statistical models." In: *Journal of artificial intelligence research*
            4 (1996), pages 129–145.

[Cha21]   Daniel T Chang. "Bayesian Neural Networks: Essentials." In: *arXiv preprint
            arXiv:2106.13594* (2021).

[CPD21]   Beau Coker, Weiwei Pan, and Finale Doshi-Velez. "Wide Mean-Field Vari-
            ational Bayesian Neural Networks Ignore the Data." In: *International
            Conference on Machine Learning* (2021).

[Dax+21]   Erik Daxberger et al. "Laplace Redux-Effortless Bayesian Deep Learn-
            ing." In: *Advances in Neural Information Processing Systems* (2021).

[FOG20]   Sebastian Farquhar, Michael A Osborne, and Yarin Gal. "Radial bayesian
            neural networks: Beyond discrete support in large-scale bayesian deep
            learning." In: *International Conference on Artificial Intelligence and Statis-
            tics*. PMLR. 2020.

[Foo+20]   Andrew Foong et al. "On the expressiveness of approximate inference in
            bayesian neural networks." In: *Advances in Neural Information Processing
            Systems* 33 (2020), pages 15897–15908.

[For21]   Vincent Fortuin. "Priors in bayesian deep learning: A review." In: *arXiv
            preprint arXiv:2105.06868* (2021).

[FSG20]   Sebastian Farquhar, Lewis Smith, and Yarin Gal. "Liberty or depth: Deep
            bayesian neural nets do not need complex weight posterior approxima-
            tions." In: *Advances in Neural Information Processing Systems* (2020).

[GG16]   Yarin Gal and Zoubin Ghahramani. "Dropout as a bayesian approxima-
           tion: Representing model uncertainty in deep learning." In: *International
           Conference on Machine Learning*. PMLR. 2016.

[GIG17]   Yarin Gal, Riashat Islam, and Zoubin Ghahramani. "Deep bayesian ac-
           tive learning with image data." In: *International Conference on Machine
           Learning*. PMLR. 2017.

[Hua+17]   Gao Huang et al. "Densely connected convolutional networks." In: *Pro-
           ceedings of the IEEE conference on computer vision and pattern recogni-
           tion*. 2017.

[Izm+21]   Pavel Izmailov et al. "What are Bayesian neural network posteriors really
           like?" In: *International Conference on Machine Learning*. PMLR. 2021.

[KH+09]   Alex Krizhevsky, Geoffrey Hinton, et al. "Learning multiple layers of
           features from tiny images." In: (2009).

[Kha+17]   Mohammad Emtiyaz Khan et al. "Variational adaptive-Newton method
           for explorative learning." In: *arXiv preprint arXiv:1711.05560* (2017).

[KSW15]   Durk P Kingma, Tim Salimans, and Max Welling. "Variational dropout
           and the local reparameterization trick." In: *Advances in Neural Informa-
           tion Processing Systems* (2015).

[LeC+98]   Yann LeCun et al. "Gradient-based learning applied to document recog-
           nition." In: *Proceedings of the IEEE* (1998).

[LPB17]   Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. "Sim-
           ple and scalable predictive uncertainty estimation using deep ensembles."
           In: *Advances in neural information processing systems* (2017).

[Mac92]   David JC MacKay. "The evidence framework applied to classification
           networks." In: *Neural computation* (1992).

[Mad+19]   Wesley J Maddox et al. "A simple baseline for bayesian uncertainty in
           deep learning." In: *Advances in Neural Information Processing Systems*
           32 (2019).

[MG15]   James Martens and Roger Grosse. "Optimizing neural networks with
           kronecker-factored approximate curvature." In: *International conference
           on machine learning*. PMLR. 2015, pages 2408–2417.

[Mia+22]   Marco Miani et al. "Laplacian Autoencoders for Learning Stochastic Rep-
           resentations." In: *arXiv preprint arXiv:2206.15078* (2022).

[Net+11]   Yuval Netzer et al. "Reading digits in natural images with unsupervised
           feature learning." In: (2011).

[Ngu+17]   Cuong V Nguyen et al. "Variational continual learning." In: *arXiv preprint
           arXiv:1710.10628* (2017).

[Pas+]   Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance
           Deep Learning Library." In: *Advances in Neural Information Processing
           Systems*.

[Raj+17]   Pranav Rajpurkar et al. "Mura: Large dataset for abnormality detection in musculoskeletal radiographs." In: *arXiv preprint arXiv:1712.06957* (2017).

[RK21]   Hippolyt Ritter and Theofanis Karaletsos. "TyXe: Pyro-based Bayesian neural nets for Pytorch." In: *arXiv preprint arXiv:2110.00276* (2021).

[Sha48]   Claude Elwood Shannon. "A mathematical theory of communication." In: *The Bell system technical journal* 27.3 (1948).

[WI20]   Andrew G Wilson and Pavel Izmailov. "Bayesian deep learning and a probabilistic perspective of generalization." In: *Advances in neural information processing systems* (2020).

[Wu+18]   Anqi Wu et al. "Deterministic variational inference for robust bayesian neural networks." In: *arXiv preprint arXiv:1810.03958* (2018).

[Zha+18]   Guodong Zhang et al. "Noisy natural gradient as variational inference." In: *International Conference on Machine Learning*. PMLR. 2018, pages 5852–5861.