Bayesian Neural Networks

James Allingham, Javier Antorán & Vincent Fortuin

MLG reading group – 22nd Feb 2023

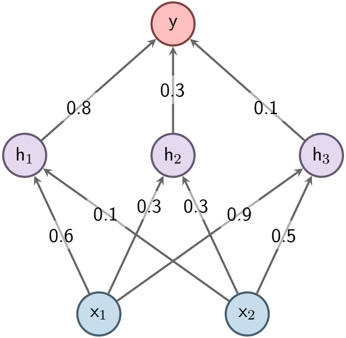
UNIVERSITY OF CAMBRIDGE

Outline

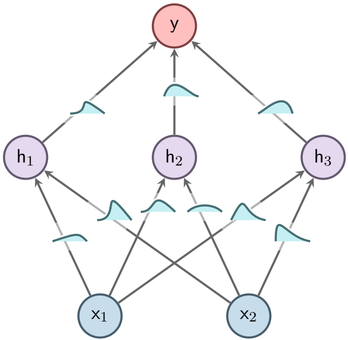
- Part 1 – James
  - BNN introduction
  - BNN challenges and solutions
  - BNN properties
- Part 2 – Javier
  - Laplace approximation
  - linear models
  - connections to infinite width limits
- Part 3 – Vincent
  - BNN priors
    - weight space
    - function space

Part 1

What are Bayesian Neural Networks?

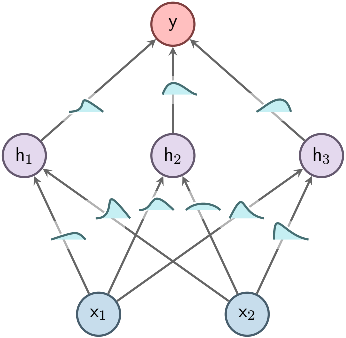


What are Bayesian Neural Networks?



$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(\mathcal{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \tag{1}$$

What are Bayesian Neural Networks?



$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(\mathcal{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \tag{1}$$

$$p(\mathbf{y}^* \mid \mathbf{x}^*) = \int_{\boldsymbol{\theta}} p(\mathbf{y}^* \mid \mathbf{x}^*,\, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{D}) d\boldsymbol{\theta} \tag{2}$$

Why BNNs?

1. NNs are poorly calibrated – they don't know when they don't know!
    - "Reject" uncertain predictions.
    - Exploration in RL / Bandits.
    - Active learning.
    - Combining different model's predictions.
    - Bet sizing.
    - etc.
2. Choosing hyper-parameters in NNs is hard (or expensive).
3. NNs can't naturally deal with missing data.

These are all problems that are solved by principled probabilistic models.

BNNs are hard

Some challenges:

- Integration!
- Choosing priors. (Part 3)
- High dimensionality.

BNNs are hard

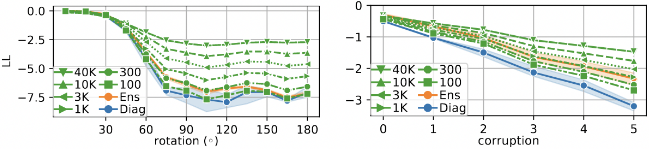
Some challenges:

- Integration!
- Choosing priors. (Part 3)
- High dimensionality.
  - Even *smaller* modern NNs have many parameters $> \mathcal{O}(10^6)$!
  - Storage of covariance matrices requires $\mathcal{O}(N^2)$ memory.
  - Makes approximation difficult.
  - Subspace [Izmailov et al., 2020] and subnetwork [Daxberger et al., 2021b] inference.

BNNs are hard

Some challenges:

- Integration!
- Choosing priors. (Part 3)
- High dimensionality.



(a) Rotated MNIST (b) Corrupted CIFAR10

Subnet Size	Memory
11.2M (100%)	500**TB**
40K (0.36%)	6.4**GB**
1K (0.01%)	4.0**MB**
100 (0.001%)	40**KB**

(c) Memory Footprints

Figure 1: Subnet inference with Laplace approx. on ResNet-18.

BNNs are hard – integration

- The integrals in eqs. (1) and (2) are intractable!
    - BNNs in practice require approximations.

BNNs are hard – integration

- The integrals in eqs. (1) and (2) are intractable!
  - BNNs in practice require approximations.
- The integral in the predictive distribution

$$p(\mathbf{y}^* \mid \mathbf{x}^*) = \int_{\boldsymbol{\theta}} p(\mathbf{y}^* \mid \mathbf{x}^*, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{D}) d\boldsymbol{\theta} \tag{2}$$

is easily approximated using Monte Carlo:

$$p(\mathbf{y}^* \mid \mathbf{x}^*) \approx \frac{1}{N} \sum_{n=1}^{N} p\left(\mathbf{y}^* \,\middle|\, \mathbf{x}^*, \boldsymbol{\theta}^{(n)}\right), \quad \boldsymbol{\theta}^{(n)} \sim p(\boldsymbol{\theta} \mid \mathcal{D}). \tag{3}$$

BNNs are hard – integration

- The integrals in eqs. (1) and (2) are intractable!
  - BNNs in practice require approximations.
- The integral in the predictive distribution

$$p(\mathbf{y}^* \mid \mathbf{x}^*) = \int_{\boldsymbol{\theta}} p(\mathbf{y}^* \mid \mathbf{x}^*, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathcal{D}) d\boldsymbol{\theta} \tag{2}$$

  is easily approximated using Monte Carlo:

$$p(\mathbf{y}^* \mid \mathbf{x}^*) \approx \frac{1}{N} \sum_{n=1}^{N} p\left(\mathbf{y}^* \,\middle|\, \mathbf{x}^*, \boldsymbol{\theta}^{(n)}\right), \quad \boldsymbol{\theta}^{(n)} \sim p(\boldsymbol{\theta} \mid \mathcal{D}). \tag{3}$$

- However, approximating the posterior distribution

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(\mathcal{D} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \tag{1}$$

  is slightly trickier...

Approximating the posterior

Two main approaches:

1. Assuming a simplified form for the posterior $p(\boldsymbol{\theta} \mid \mathcal{D})$, allowing us to avoid (or simplify) calculating the evidence $p(\mathcal{D})$.

Approach 1 – Simplified Posteriors

- Laplace approx. [MacKay, 1992, Daxberger et al., 2021a].
- VI [Hinton and Van Camp, 1993, Graves, 2011, Blundell et al., 2015, Osawa et al., 2019].
- EP [Hernández-Lobato and Adams, 2015].
- MC Dropout [Gal and Ghahramani, 2016].
  - Some issues [Osband, 2016].

Approximating the posterior

Two main approaches:

1. Assuming a simplified form for the posterior $p(\boldsymbol{\theta} \mid \mathcal{D})$, allowing us to avoid (or simplify) calculating the evidence $p(\mathcal{D})$.
2. Using MCMC methods to sample directly from $p(\boldsymbol{\theta} \mid \mathcal{D})$ without ever calculating $p(\mathcal{D})$.

Approach 2 – Sampling

- Pionered by Neal [1995], who used HMC [Duane et al., 1987, Neal, 2012]. "Gold standard".
- SGLD [Welling and Teh, 2011] & SGHMC [Chen et al., 2014].
  - Biased [Betancourt, 2015].
  - No rejection sampling [Garriga-Alonso and Fortuin, 2021].

A simple baseline – MFVI

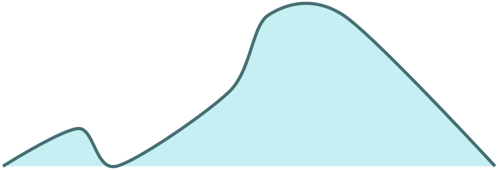
High level idea: approximate $p(\boldsymbol{\theta} \mid \mathcal{D})$ with $q_{\boldsymbol{\phi}}(\boldsymbol{\theta})$.

- q is a NN parameterised by $\boldsymbol{\phi}$.
- Mean-field assumption: dimensions of $\boldsymbol{\theta}$ are independent.

A simple baseline – MFVI

High level idea: approximate $p(\boldsymbol{\theta} \mid \mathcal{D})$ with $q_{\phi}(\boldsymbol{\theta})$.

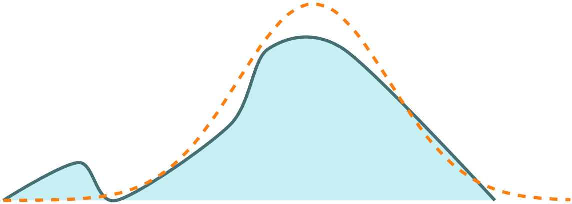
- q is a NN parameterised by ϕ.
- Mean-field assumption: dimensions of $\boldsymbol{\theta}$ are independent.



A simple baseline – MFVI

High level idea: approximate $p(\boldsymbol{\theta} \mid \mathcal{D})$ with $q_{\phi}(\boldsymbol{\theta})$.

- q is a NN parameterised by ϕ.
- Mean-field assumption: dimensions of $\boldsymbol{\theta}$ are independent.



A simple baseline – MFVI

High level idea: approximate $p(\boldsymbol{\theta} \mid \mathcal{D})$ with $q_{\boldsymbol{\phi}}(\boldsymbol{\theta})$.

- q is a NN parameterised by $\boldsymbol{\phi}$.
- Mean-field assumption: dimensions of $\boldsymbol{\theta}$ are independent.

$$\boldsymbol{\phi} = \arg\min_{\boldsymbol{\phi}} D_{\mathrm{KL}}\left[q_{\boldsymbol{\phi}}(\boldsymbol{\theta}) \,||\, p(\boldsymbol{\theta} \mid \mathcal{D})\right] \tag{4}$$

$$= \arg\max_{\boldsymbol{\phi}} \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{\theta})}\left[\log p(\mathcal{D} \mid \boldsymbol{\theta})\right] - D_{\mathrm{KL}}\left[q_{\boldsymbol{\phi}}(\boldsymbol{\theta}) \,||\, p(\boldsymbol{\theta})\right] \tag{5}$$

$$= \arg\max_{\boldsymbol{\phi}} \mathcal{L}(\boldsymbol{\phi}) \tag{6}$$

A simple baseline – MFVI

High level idea: approximate $p(\boldsymbol{\theta} \mid \mathcal{D})$ with $q_{\boldsymbol{\phi}}(\boldsymbol{\theta})$.

- q is a NN parameterised by $\boldsymbol{\phi}$.
- Mean-field assumption: dimensions of $\boldsymbol{\theta}$ are independent.

$$\boldsymbol{\phi} = \arg\min_{\boldsymbol{\phi}} D_{\mathrm{KL}}\left[q_{\boldsymbol{\phi}}(\boldsymbol{\theta}) \,||\, p(\boldsymbol{\theta} \mid \mathcal{D})\right] \tag{4}$$

$$= \arg\max_{\boldsymbol{\phi}} \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{\theta})}\left[\log p(\mathcal{D} \mid \boldsymbol{\theta})\right] - D_{\mathrm{KL}}\left[q_{\boldsymbol{\phi}}(\boldsymbol{\theta}) \,||\, p(\boldsymbol{\theta})\right] \tag{5}$$

$$= \arg\max_{\boldsymbol{\phi}} \mathcal{L}(\boldsymbol{\phi}) \tag{6}$$

1. $\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{\theta})}\left[\log p(\mathcal{D} \mid \boldsymbol{\theta})\right]$ – data fit term
2. $D_{\mathrm{KL}}\left[q_{\boldsymbol{\phi}}(\boldsymbol{\theta}) \,||\, p(\boldsymbol{\theta})\right]$ – complexity term

A simple baseline – MFVI

High level idea: approximate $p(\boldsymbol{\theta} \mid \mathcal{D})$ with $q_{\phi}(\boldsymbol{\theta})$.

- q is a NN parameterised by ϕ.
- Mean-field assumption: dimensions of $\boldsymbol{\theta}$ are independent.

$$\begin{aligned} \boldsymbol{\phi} &= \arg\min_{\boldsymbol{\phi}} D_{\mathrm{KL}}\left[q_{\phi}(\boldsymbol{\theta}) \,||\, p(\boldsymbol{\theta} \mid \mathcal{D})\right] && (4) \\ &= \arg\max_{\boldsymbol{\phi}} \mathbb{E}_{q_{\phi}(\boldsymbol{\theta})}\left[\log p(\mathcal{D} \mid \boldsymbol{\theta})\right] - D_{\mathrm{KL}}\left[q_{\phi}(\boldsymbol{\theta}) \,||\, p(\boldsymbol{\theta})\right] && (5) \\ &= \arg\max_{\boldsymbol{\phi}} \mathcal{L}(\boldsymbol{\phi}) && (6) \end{aligned}$$

1. $\mathbb{E}_{q_{\phi}(\boldsymbol{\theta})}\left[\log p(\mathcal{D} \mid \boldsymbol{\theta})\right]$ – data fit term
2. $D_{\mathrm{KL}}\left[q_{\phi}(\boldsymbol{\theta}) \,||\, p(\boldsymbol{\theta})\right]$ – complexity term

$$\mathcal{L}(\boldsymbol{\phi}) \approx \frac{1}{N} \sum_{n=1}^{N} \left[\log p\left(\mathcal{D} \,\middle|\, \boldsymbol{\theta}^{(n)}\right) - \log q_{\phi}\left(\boldsymbol{\theta}^{(n)}\right) + \log p\left(\boldsymbol{\theta}^{(n)}\right)\right], \quad \boldsymbol{\theta}^{(n)} \sim q_{\phi}(\boldsymbol{\theta}). \qquad (7)$$

Problems with MFVI for BNNs

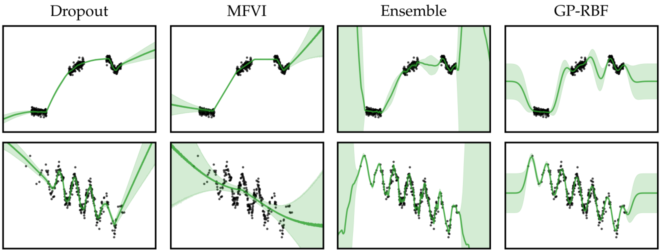


Figure 2: MFVI doesn't provide "in-between" uncertainty, and underfits!

Problems with MFVI for BNNs

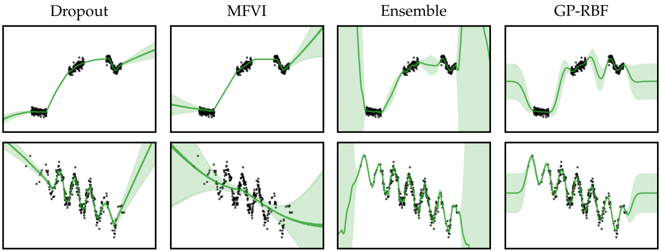


Figure 2: MFVI doesn't provide "in-between" uncertainty, and underfits!

- Foong et al. [2020] prove that MFVI (and MC Dropout) cannot capture "in-between" uncertainty for single hidden layer BNNs.
- They demonstrate this is a problem of approximate inference.
- They show empirically that this also occurs for deeper BNNs (despite proving that they are universal approximators for μ and σ).
- Farquhar et al. [2020] argue that MFVI is less restrictive with depth.

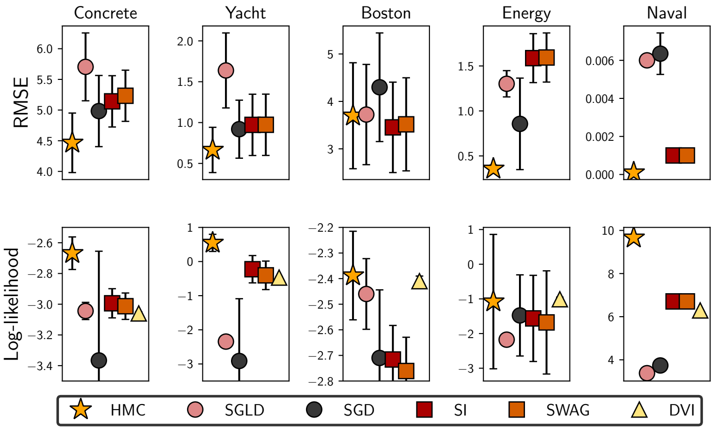
What are BNN posteriors really like?

Izmailov et al. [2021b] perform *full batch* 🔥 HMC for modern NNs to explore this question.

Note: this is not practical at all! But we can learn a lot.

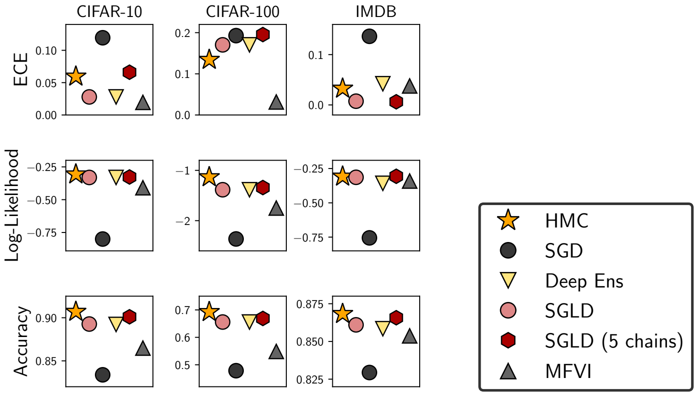
What are BNN posteriors really like?

Finding 1 – BNNs can achieve significant performance gains



What are BNN posteriors really like?

Finding 1 – BNNs can achieve significant performance gains

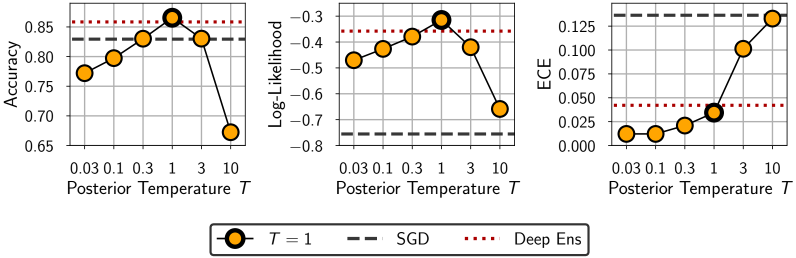


What are BNN posteriors really like?

Finding 2 – Posterior tempering is not needed

There is little evidence for a "cold posterior" effect [Wenzel et al., 2020], which seems to be largely caused by data augmentation.

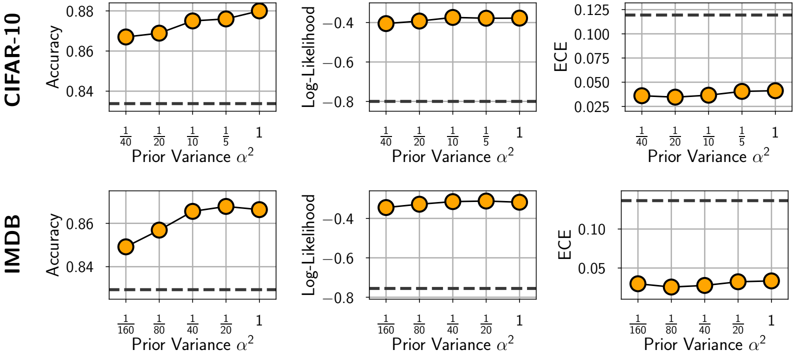
$$p_T(w|\mathcal{D}) \propto \left(p(\mathcal{D}|w) \cdot p(w)\right)^{1/T} \tag{8}$$



What are BNN posteriors really like?

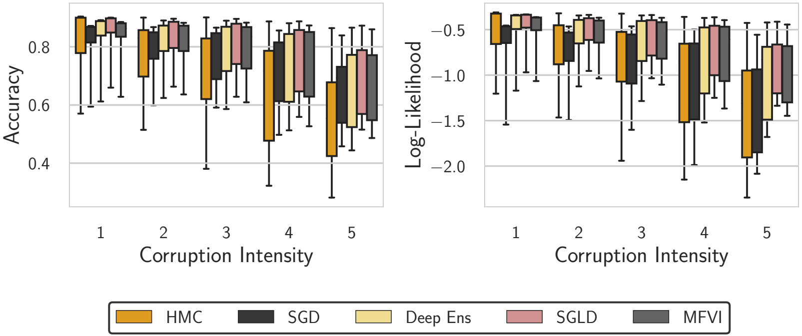
Finding 3 – Performance is robust to the choice of prior scale

... and pretty similar for diag. Gaussian, MoG, and logistic priors.



What are BNN posteriors really like?

Finding 4 – BNNs are surprisingly bad under dist. shift



Addressed in [Izmailov et al., 2021a] with prior choice.

What are BNN posteriors really like?

Finding 5 – Deep ens. and SGMCMC provide distinct predictive dists. from HMC

Deep ensembles and SGMCMC can provide good generalization. Deep ensemble predictive distributions are as to HMC as SGLD, and closer than VI.

Metric	HMC (reference)	SGD	Deep Ens	MFVI	SGLD	SGHMC	SGHMC CLR	SGHMC CLR-Prec
				CIFAR-10				
Accuracy	89.64 ±0.25	83.44 ±1.14	88.49 ±0.10	86.45 ±0.27	89.32 ±0.23	89.38 ±0.32	**89.63** **±0.37**	87.46 ±0.21
Agreement	94.01 ±0.25	85.48 ±1.00	91.52 ±0.06	88.75 ±0.24	91.54 ±0.15	91.98 ±0.35	**92.67** **±0.52**	90.96 ±0.24
Total Var	0.074 ±0.003	0.190 ±0.005	0.115 ±0.000	0.136 ±0.000	0.110 ±0.001	0.109 ±0.001	**0.099** **±0.006**	0.111 ±0.002

What are BNN posteriors really like?

(In an idealised setting...)

1. BNNs can achieve significant performance gains over standard training and deep ensembles.
2. Posterior tempering is not needed for near-optimal performance, with little evidence for a "cold posterior" effect (largely caused by data augmentation).
3. Performance is robust to the choice of prior scale, and relatively similar for diagonal Gaussian, MoG, and logistic priors.
4. BNNs show surprisingly poor generalization under distribution shift. Addressed in [Izmailov et al., 2021a] with prior choice.
5. Deep ensembles and SGMCMC can provide good generalization, but different predictive distributions from HMC. Notably, deep ensemble predictive distributions are as to HMC as SGLD, and closer than VI.

References I

M. Betancourt. The fundamental incompatibility of scalable hamiltonian monte carlo and naive data subsampling. In F. R. Bach and D. M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 533–540. JMLR.org, 2015. URL `http://proceedings.mlr.press/v37/betancourt15.html`.

C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR, 2015.

References II

T. Chen, E. B. Fox, and C. Guestrin. Stochastic gradient hamiltonian monte carlo. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1683–1691. JMLR.org, 2014. URL `http://proceedings.mlr.press/v32/cheni14.html`.

E. Daxberger, A. Kristiadi, A. Immer, R. Eschenhagen, M. Bauer, and P. Hennig. Laplace redux-effortless bayesian deep learning. *Advances in Neural Information Processing Systems*, 34:20089–20103, 2021a.

E. Daxberger, E. Nalisnick, J. U. Allingham, J. Antorán, and J. M. Hernández-Lobato. Bayesian deep learning via subnetwork inference. In *International Conference on Machine Learning*, pages 2510–2521. PMLR, 2021b.

References III

S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics letters B*, 195(2):216–222, 1987. tex.ids: duane1987hybrida publisher: Elsevier.

S. Farquhar, L. Smith, and Y. Gal. Liberty or depth: Deep bayesian neural nets do not need complex weight posterior approximations. *Advances in Neural Information Processing Systems*, 33:4346–4357, 2020.

A. Foong, D. Burt, Y. Li, and R. Turner. On the expressiveness of approximate inference in bayesian neural networks. *Advances in Neural Information Processing Systems*, 33:15897–15908, 2020.

Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

A. Garriga-Alonso and V. Fortuin. Exact langevin dynamics with stochastic gradients. *arXiv preprint arXiv:2102.01691*, 2021.

References IV

A. Graves. Practical variational inference for neural networks. *Advances in neural information processing systems*, 24, 2011.

J. M. Hernández-Lobato and R. Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International conference on machine learning*, pages 1861–1869. PMLR, 2015.

G. E. Hinton and D. Van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 5–13, 1993.

P. Izmailov, W. J. Maddox, P. Kirichenko, T. Garipov, D. Vetrov, and A. G. Wilson. Subspace inference for bayesian deep learning. In *Uncertainty in Artificial Intelligence*, pages 1169–1179. PMLR, 2020.

P. Izmailov, P. Nicholson, S. Lotfi, and A. G. Wilson. Dangers of bayesian model averaging under covariate shift. *Advances in Neural Information Processing Systems*, 34:3309–3322, 2021a.

References V

P. Izmailov, S. Vikram, M. D. Hoffman, and A. G. G. Wilson. What are bayesian neural network posteriors really like? In *International conference on machine learning*, pages 4629–4640. PMLR, 2021b.

D. J. MacKay. A practical Bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992. tex.ids: mackay1992practicala publisher: MIT Press.

R. M. Neal. *Bayesian learning for neural networks*. PhD thesis, University of Toronto, Canada, 1995. URL `https://librarysearch.library.utoronto.ca/permalink/01UTORONTO_INST/14bjeso/alma991106438365706196`.

R. M. Neal. MCMC using Hamiltonian dynamics. *arXiv:1206.1901 [physics, stat]*, June 2012. URL `http://arxiv.org/abs/1206.1901`. tex.ids: neal2011mcmc arXiv: 1206.1901.

References VI

K. Osawa, S. Swaroop, M. E. E. Khan, A. Jain, R. Eschenhagen, R. E. Turner, and R. Yokota. Practical deep learning with bayesian principles. *Advances in neural information processing systems*, 32, 2019.

I. Osband. Risk versus uncertainty in deep learning: Bayes, bootstrap and the dangers of dropout. In *NIPS workshop on bayesian deep learning*, volume 192, 2016.

M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient langevin dynamics. *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, 2011.

F. Wenzel, K. Roth, B. S. Veeling, J. Świątkowski, L. Tran, S. Mandt, J. Snoek, T. Salimans, R. Jenatton, and S. Nowozin. How good is the bayes posterior in deep neural networks really? *arXiv preprint arXiv:2002.02405*, 2020.