Proceedings of Machine Learning Research 176:113–124, 2022 NeurIPS 2021 Competition and Demonstration Track

Evaluating Approximate Inference in Bayesian Deep Learning

Andrew Gordon Wilson[1] ANDREWGW@CIMS.NYU.EDU
Sanae Lotfi[1] SL8160@NYU.EDU
Sharad Vikram[2] SHARADMV@GOOGLE.COM
Matthew D. Hoffman[2] MHOFFMAN@GOOGLE.COM
Yarin Gal[3] YARIN@CS.OX.AC.UK
Yingzhen Li[4] LIYZHEN2@GMAIL.COM
Melanie F. Pradier[5] MELANIEF@MICROSOFT.COM
Andrew Foong[6] YKF21@CAM.AC.UK
Sebastian Farquhar[3] SEBASTIAN.FARQUHAR@CS.OX.AC.UK
Pavel Izmailov[1] PI390@NYU.EDU

[1] *New York University* [2] *Google Research* [3] *University of Oxford*
[4] *Imperial College London* [5] *Microsoft Research* [6] *University of Cambridge*

Editor: Douwe Kiela, Marco Ciccone, Barbara Caputo

Abstract

Uncertainty representation is crucial to the safe and reliable deployment of deep learning. Bayesian methods provide a natural mechanism to represent epistemic uncertainty, leading to improved generalization and calibrated predictive distributions. Understanding the fidelity of approximate inference has extraordinary value beyond the standard approach of measuring generalization on a particular task: if approximate inference is working correctly, then we can expect more reliable and accurate deployment across any number of real-world settings. In this competition, we evaluate the fidelity of approximate Bayesian inference procedures in deep learning, using as a reference Hamiltonian Monte Carlo (HMC) samples obtained by parallelizing computations over hundreds of tensor processing unit (TPU) devices. We consider a variety of tasks, including image recognition, regression, covariate shift, and medical applications. All data are publicly available, and we release several baselines, including stochastic MCMC, variational methods, and deep ensembles.

The competition resulted in hundreds of submissions across many teams. The winning entries all involved novel multi-modal posterior approximations, highlighting the relative importance of representing multiple modes, and suggesting that we should not consider deep ensembles a "non-Bayesian" alternative to standard unimodal approximations. In the future, the competition will provide a foundation for innovation and continued benchmarking of approximate Bayesian inference procedures in deep learning. The HMC samples will remain available through the competition website.

Keywords: Bayesian inference, Bayesian deep learning, approximate inference, Hamiltonian Monte Carlo

1. Competition Description

While deep learning has been revolutionary for machine learning, most modern deep learning models cannot represent their uncertainty nor take advantage of the well-studied tools of probability theory. The community has been immensely active in addressing this gap, with the introduction of new deep learning models that use Bayesian inference techniques, and Bayesian models that incorporate deep learning elements. The broad and consistent community interest in these topics is clearly evidenced by the NeurIPS Bayesian Deep Learning workshop being the second largest workshop at the conference every year since 2016. This broad interest is also clear from major tutorials on Bayesian deep learning and uncertainty representation in deep learning at NeurIPS 2019, ICML 2020, and NeurIPS 2021 (Khan, 2019; Wilson, 2020; Tran et al., 2020).

The use of Bayesian techniques in deep learning can be traced back to the 1990s, in seminal works by Radford Neal (Neal, 1996) and David MacKay (MacKay, 1995). These works gave rise to tools to reason about deep models' confidence, and achieved state-of-the-art performance on many tasks at the time. With a resurgence of deep learning, there has been extraordinary progress in the last five years for scaling approximate Bayesian inference procedures to modern architectures and datasets.

Many of these procedures have provided promising performance on tasks of public interest, such as medical diagnoses and more reliable autonomous driving (Leibig et al., 2017; Filos et al., 2019). For example, in medical diagnosis, it is not sufficient to simply label an image as pathological or healthy. Instead, we need to make a decision about treatment based on *probabilities* of class labels. For this purpose, Bayesian methods represent *epistemic uncertainty* over different hypotheses for the data, in order to provide a full predictive distribution. This predictive distribution is then crucial in decision making, as it can be combined with a loss function that recognizes asymmetry in outcomes, and that rare mistakes can be extraordinarily costly. A false negative, for instance, is often much more costly than a false positive.

However, there has been no mechanism for understanding whether approximate inference procedures in deep learning are working as intended, and providing a faithful approximation of a Bayesian predictive distribution. Indeed, standard metrics, such as generalization accuracy, or negative log likelihood, provide no way of separating the effects of model specification and inference procedure (Yao et al., 2019).

It is helpful to consider an analogy with optimization. Optimization procedures, in principle, are intended to minimize our training objective, not to provide good test set generalization on a particular benchmark task. A method that is good at optimization will presumably be useful on a wide variety of problems. While it is easy to query our training loss in optimization, we do not have direct access to the Bayesian posterior predictive distribution.

In this competition we provide a unique opportunity to measure the *fidelity* of approximate inference procedures in deep learning through comparison to Hamiltonian Monte Carlo (HMC) (Neal et al., 2011). HMC is a highly efficient and well-studied Markov Chain Monte Carlo (MCMC) method that is guaranteed to asymptotically produce samples from the true posterior, but is prohibitively expensive in modern deep learning: HMC can take tens of thousands of training epochs to produce a single sample from the posterior. To

address this computational challenge, we have parallelized the computation over hundreds of tensor processing unit (TPU) devices. We provide extensive details for our procedure, as well as comparisons to several popular baselines, in Izmailov et al. (2021b).

This competition provides a standardized mechanism for evaluating the fidelity of a wide variety of approaches to approximate inference in deep learning. Each participant is given access to our HMC samples for a variety of architectures across several reference datasets. Participants only need to provide access to the predictive distribution from their procedure, and our evaluation framework then creates a leaderboard of all methods. In the evaluation phase of the competition, participants submit code for their inference procedures on evaluation datasets, and we locally evaluate the fidelity of inference. We consider problems for image recognition, regression, covariate shift, and healthcare. Total runtime of the approximate inference procedures is constrained to no more than ten times standard SGD training, which covers essentially all modern approximate inference procedures, but is several orders of magnitude less expensive than full HMC.

The competition will provide an enormous resource for understanding the efficacy of many approximate inference procedures, separating model specification and inference in evaluation, and the design of new inference algorithms which can provide reliable inference at a much lower cost than HMC, which is otherwise inaccessible to machine learning practitioners. More broadly, the development of high fidelity Bayesian inference procedures in deep learning is a crucial component of building safe and robust systems for automatic decision making — which requires faithful representations of uncertainty, and reliable predictive distributions.

In this section we outline the details of the competition setup, and in Sections 2 and 3 we discuss the results and conclusions. The competition had hundreds of entries, inspiring new approximate inference procedures and conceptual insights. One notable theme in the leading entries, discussed further in the next section, is the use of multimodal posterior approximations. The competition will form a lasting benchmark to evaluate approximate inference procedure in deep learning, with the latest samples and data available through the competition website: https://izmailovpavel.github.io/neurips_bdl_competition/.

1.1. Tasks

Bayesian deep learning methods have countless potential applications precisely because inferring the Bayesian posterior distribution is such a powerful principled way to incorporate the information contained in a training dataset. These scenarios include:

- Safe medical diagnostics: automatically handling clear-cut diagnoses while elevating difficult decisions to medical professionals who can request further scans.
- Rare or under-represented inputs: recognizing the uncertainty present when individuals come from groups that are under-represented in datasets and seeking guidance from experts.
- Covariate shift: identifying situations where covariate shift makes model predictions unreliable, for example in autonomous driving.

In recent work, researchers have made great progress on specific metrics associated with these sorts of applications of Bayesian deep learning (e.g., Lakshminarayanan et al., 2017;

Table 1: **Required Tasks.** Submissions to the competition are requested to demonstrate their effectiveness at approximating the predictive posterior across a range of easily-accessible applications and architectures. For CIFAR datasets, both the standard test set and corrupted versions (Hendrycks and Dietterich, 2019) are used. For UCI, we use the regression datasets chosen in Hernández-Lobato and Adams (2015).

Prediction Type	Dataset	Architecture	Metrics Assessed
	Development Datasets		
Classification	CIFAR-10-(C)	ResNet-20	Top-1 Agreement & Total Variation
	CIFAR-100-(C)	ResNet-20	Top-1 Agreement & Total Variation
	IMDB	CNN-LSTM	Top-1 Agreement & Total Variation
	Evaluation Datasets		
Classification	CIFAR-10-(C)	AlexNet	Top-1 Agreement & Total Variation
	MedMNIST: DermaMNIST (Yang et al., 2020)	LeNet-5	Top-1 Agreement & Total Variation
Regression	UCI-Gap: Energy (Foong et al., 2019)	3x200 FCNN	Wasserstein Distance

Leibig et al., 2017; Maddox et al., 2019; Filos et al., 2019; Ovadia et al., 2019; Izmailov et al., 2021a).

However, prior work has mostly focused on simple metrics like accuracy, log-likelihood, and rejection accuracy. It is possible to score highly on these in individual cases even if a method generalizes poorly to other tasks due to issues with approximate inference. Rather than focus on generalization error for a set of tasks, we instead wish to understand which approximate inference methods are performing as intended to produce *high quality posterior approximations*, leading to more calibrated expectations of their applicability across a broad range of settings.

To this end, we ask the participants to implement approximate inference for the application scenarios listed in Table 1. For the datasets listed as *development datasets*, we provide the HMC checkpoints and the corresponding predictive distributions to the participants. These datasets were used to develop and calibrate the solutions in the first stage of the competition (development phase). The datasets listed under *evaluation datasets* were used to evaluate the submissions. The participants submited the training scripts to produce the predictive distributions on these datasets, which we then compare to the predictive distributions of the private HMC checkpoints.

These datasets cover different types of supervised learning tasks (regression and classification), and the evaluation is conducted for both in-distribution and out-of-distribution scenarios. Also the datasets and reference architectures in Table 1 were selected to enable the collection of reliable HMC simulation results. We believe these are comprehensive benchmarks that allow to draw conclusive results from the competition.

1.2. Metrics

The submissions were evaluated based on the similarity of their predictive distribution to the predictive distribution approximated by a long run of multiple Hamiltonian Monte Carlo chains. Let us denote the target predictive distribution approximated by HMC for an input x by $\hat{p}(y|x)$, and let $p(y|x)$ be the predictive distribution from a submission to the competition.

For classification tasks, we consider two primary metrics: *agreement* and *total variation*. Let $D_{test} = \{x_i\}_{i=1}^n$ be the test dataset. Then we define the agreement between $\hat{p}$ and p as the fraction of the test data points for which the top-1 predictions of $\hat{p}$ and p agree:

$$\text{agreement}(\hat{p}, p) = \frac{1}{n} \sum_{i=1}^{n} I[\arg\max_j \hat{p}(y = j|x_i) = \arg\max_j p(y = j|x_i)], \tag{1}$$

where $I[\cdot]$ is the indicator function. The agreement metric measures how well the submission is able to capture the top-1 predictions of the Bayesian model average. Higher is better.

We define the total variation metric between $\hat{p}$ and p as the total variation distance between the predictive distributions averaged over the test data points:

$$\text{TV}(\hat{p}, p) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \sum_j \left| \hat{p}(y = j|x_i) - p(y = j|x_i) \right|. \tag{2}$$

The total variation metric captures how well the full predictive distributions of $\hat{p}$ and p agree. In order to achieve a low total variation score, the submission has to capture not only the top-1 prediction of HMC, but also all of the class-probabilities. Lower is better.

For regression tasks, we consider the Wasserstein-2 distance between $\hat{p}$ and p. Since p is provided as a set of sampled predictions for each example, the metric is calculated as a point-wise W_2 distance:

$$W_2(\hat{p}, p) = \inf_I \sqrt{\sum_{i \in I, j} |p_i - \hat{p}_j|^2}, \tag{3}$$

where I are possible orderings of points. Lower is better.

In addition to these scores, the submissions were constrained in training time (s): the time taken for the training script to perform approximate inference. We require the training time to be no more than the equivalent of 1000 SGD training epochs for each of the tasks. We will request the participants to provide the scripts for the several winning submissions.

Submissions received a performance score which is a weighted average of (1−agreement), total variation, and Wasserstein distance averaged over the test problems. For each of the metrics we rank the submissions and compute the average ranking accross the problem. For the *light track* we only consider the CIFAR-10 dataset, and for the *extended track* we average the rankings across all the datasets.

1.3. Baselines, code, and material provided

As baselines, we provide Deep Ensembles, Mean-Field Variational Inference, Monte-Carlo Dropout and several variations of Stochastic Gradient MCMC (SG-MCMC) methods. For these methods we provide the results on the development phase datasets (see Table 2)

Table 2: **Baselines and example results.** The agreement and total variation metrics for the deep ensembles and SG-MCMC variations. The methods were trained on the CIFAR-10 train set, and we report the results on the original CIFAR-10 test set and the corrupted test sets from CIFAR-10-C. For CIFAR-10-C we report the mean and standard deviation of the metrics over the different corruptions and corruption intensities.

			SG-MCMC			
Dataset	Metric	Deep Ensembles	SGLD	SGHMC	SGHMC-CLR	SGHMC-CLR-Prec
CIFAR-10 Test	Agreement	91.7	91.8	92.2	**92.8**	**92.8**
	Total Variation	0.104	0.106	0.105	0.095	**0.092**
CIFAR-10-C	Agreement	80.1 ± 9.5	78.9 ± 10.9	79.9 ± 10.3	81.77 ± 8.8	**82.5 ± 8.4**
	Total Variation	0.204 ± 0.073	0.205 ± 0.076	0.193 ± 0.069	0.183 ± 0.064	**0.172 ± 0.06**

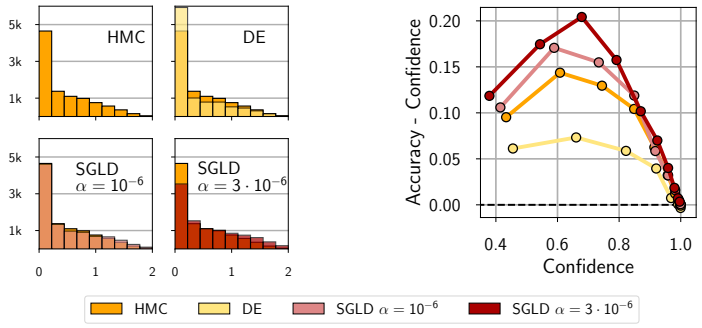

Figure 1: Distribution of predictive entropies (**left**) and calibration curve (**right**) of posterior predictive distributions for HMC, SGLD (with learning rates $\alpha = 10^{-6}$ and $3 \cdot 10^{-6}$), and deep ensemble on ResNet20-FRN on CIFAR-10. On the left, for all methods, except HMC we plot a pair of histograms: for HMC and for the corresdponding method. Deep ensembles provide more confident predictions than HMC, SGLD with high learning rate is undercondident, while SGLD with $\alpha = 10^{-6}$ matches HMC well.

and an implementation (starting kit) in the JAX framework. In Figure 1 we visualize the predictive entropy distribution and the calibration curve for HMC, deep ensembles and SGLD at different learning rates. For more detail, see Izmailov et al. (2021b).

1.4. Tutorial and documentation

We describe our HMC implementation in detail, as well as baseline comparisons, in Izmailov et al. (2021b). We additionally provide a detailed documentation of the API for submissions to the competition, and tutorial resources on Bayesian deep learning at the competition website (https://izmailovpavel.github.io/neurips_bdl_competition/). We provided a tutorial on Bayesian deep learning at ICML 2020 (Wilson, 2020).

Table 3: **Results of the competition.** The results for each of the metrics on each of the datasets for the top 6 participating teams.

	CIFAR-10		MedMNIST		UCI
Team	Agreement ↑	TVD ↓	Agreement ↑	TVD ↓	W_2 ↓
moellenh	**0.787**	**0.198**	0.884	0.099	**0.094**
adelaunoy	0.773	0.21	0.875	0.107	0.116
nkotelevskii & achille.thin	0.778	0.219	0.89	0.098	0.166
Bodacious	0.722	0.267	0.146	0.699	0.968
pierre-simon	0.719	0.276	0.819	0.151	0.486
Santosh	0.721	0.288	**0.891**	**0.094**	-

2. Competition Results

Here, we discuss the winning solutions and the results of the competition. In the evaluation phase we received a total of **337 submissions** from 12 different teams. We report the results of the top 6 teams in Table 3. Full results are available on the competition website (https://izmailovpavel.github.io/neurips_bdl_competition/).

2.1. Team `moellenh`

Team members: Thomas Möllenhoff, Yuesong Shen, Gian Maria Marconi, Peter Nickl, Mohammad Emtiyaz Khan.
Light track: first place.
Extended track: first place.

Team `moellenh` designed an innovative method based on the Bayesian learning rule (BLR) introduced by Khan and Rue (2021). The idea of their approach is to construct a global approximation to the posterior over the parameters with a structured approximation around multiple independent modes. They construct a Gaussian mixture posterior approximation, where each Gaussian has a diagonal covariance, and is estimated using an Adam-like optimizer, called the improved Variational Online-Newton (iVON) method, originally proposed by Lin et al. (2020). Each of the Gaussians is constructed from an independent run of the method, and the Gaussians are weighted uniformly in the mixture.

Team `moellenh` reported that on the development phase CIFAR-10 dataset, the proposed iVON method with a single mode approximation achieved an agreement of 93.2%, compared to the baselines: SGD with 91% agreement and VoGN (Osawa et al., 2019) with 91.9% agreement. By considering a mixture of Gaussian approximations to 8 modes they managed to further improve the agreement to 95%.

Team `moellenh` achieved the best results among all teams on the CIFAR and UCI datasets at the evaluation phase. For each dataset, they used a mixture of 6 independently

trained Gaussians to approximate the posterior. They also reported that they could outperform the best solution on MedMNIST data by using a more expensive approximation with 16 Gaussians, which exceeded the training-time constraints.

2.2. Team `nkotelevskii & achille.thin`

Team members: Nikita Kotelevskii and Achille Thin.
Light track: second place (shared).
Extended track: second place.

The approach of team `nkotelevskii & achille.thin` is based on the MutiSWAG method (Wilson and Izmailov, 2020). The idea of MultiSWAG is to construct a Gaussian mixture posterior approximation, where each Gaussian is a local approximation to the posterior around a different mode. In standard MultiSWAG, the Gaussian mean and covariance are approximated from the optimization trajectory: the mean is the mean of the weights of the optimization iterates, and the covariance matrix is a low-rank plus diagonal approximation to their empirical covariance.

Team `nkotelevskii & achille.thin` modified MultiSWAG to use SGLD (Welling and Teh, 2011), a stochastic Gradient MCMC sampler, instead of a standard optimizer. Furthermore, they explored several novel ways of representing the covariance matrix for each of the Gaussians: a layer-wise factorization and an even more fine-grained factorization with multiple components per layer. For their final solution, they used an ensemble of multiple Gaussians with different types of covariance matrix representation.

Team `nkotelevskii & achille.thin` reported that their results are sensitive to the choice of hyper-parameters. They leveraged the publicly available development data to tune these hyper-parameters to achieve optimal results.

2.3. Team `adelaunoy`

Team members: Arnaud Delaunoy. **Advisor**: Gilles Louppe.
Light track: second place (shared).
Extended track: third place.

Team `adelaunoy` developed an extended version of the *Anchored Ensembles* initially proposed by Pearce et al. (2020). The idea of anchored ensembles is to inject noise in the training procedure for an ensemble of neural networks in such a way that the ensemble components converge to samples from the posterior. To do so, the solutions are regularized to be close to points sampled from the prior, known as anchors.

Team `adelaunoy` modified the anchored ensembles to improve their computational efficiency: instead of training each ensemble component from scratch, they initialized new ensemble components from a previously obtained solution. In this case, anchors are drawn with a Guided Walk Metropolis-Hastings MCMC procedure which produces correlated samples from the prior, to ensure that the anchors are close to each other, but at the same time cover the prior distribution. The resulting procedure can construct an ensemble of 21 members in the time needed for standard anchored ensembles to train just 2 members.

The final algorithm used by team `adelaunoy` used a combination of independently trained and sequential anchored ensembles to achieve an optimal trade-off of computational efficiency and ensemble diversity.

3. Observations and Takeaways

In this section we discuss the important takeaways and observations we made from the results of the competition.

Most notably, all the teams that finished in the top-5 used some form of ensembling to construct a multimodal approximation to the posterior. Despite the fact that in this competition we measure the fidelity of approximate inference and not the generalization performance, multiple participants found ensembling to lead to significant improvements. Furthermore, team `moellenh` reported that standard deep ensembles provide a very strong baseline for matching the HMC predictive distributions. The same observation has been made in Wilson and Izmailov (2020) and Izmailov et al. (2021b) and discussed in detail in the blogpost *Deep Ensembles as Approximate Bayesian Inference* (`https://cims.nyu.edu/~andrewgw/deepensembles/`). These results highlight that we should stop treating standard deep ensembles as a "non-Bayesian" alternative to procedures such as unimodal variational inference. Indeed, multi-modal approximations to the posterior should become a new standard in Bayesian deep learning, and the multi-modality may even be more important than the quality of approximation within each of the modes.

Another observation shared by multiple participants is that the quality of the posterior approximation achieved by their methods is highly dependent on the hyper-parameters. We believe that high-quality HMC references that we developed for this competition can be used to tune Bayesian deep learning methods to achieve high quality approximate inference in practice.

Infinite width limits of Bayesian neural networks are sometimes also used as a proxy for "ground truth" inference, because under certain conditions in regression these limits converge to a closed-form Gaussian process predictive distribution (e.g., Foong et al., 2020; He et al., 2020). However, these limits give rise to a different model than the parametric Bayesian neural network analogues, thus making it unclear how closely we would expect high quality inference within the parametric Bayesian neural network to match the exact inference in the infinite limit network. On the UCI benchmark (the only benchmark here that permits closed form inference for the infinite limit) we found that the infinite limit ranked last in its similarity to HMC. Comparing these metrics further is an interesting direction for future work.

Finally, we found that all the submitted methods struggled to match the predictions of HMC on corrupted data: on the corrupted CIFAR-10-C data the best agreement achieved by any team was 78.7% compared to 91.6% on the clean CIFAR-10 test set. We believe that there is still a large room for improvement for approximate inference methods and our competition provides a unique benchmark to measure the progress in this area.

In conclusion, we believe the competition will provide a foundation for innovation and continued benchmarking of approximate Bayesian inference procedures in deep learning. We received over 300 submissions from 12 active teams, each developing a unique approach to the competition. The top 3 submissions all developed novel ideas providing state-of-the-art

methods in Bayesian deep learning. The data and HMC samples used in the competition will be hosted on the competition website, establishing a permanent benchmark for approximate inference methods in Bayesian deep learning.

If you find this benchmark or the HMC samples useful in your research, please cite Izmailov et al. (2021b) and this competition summary.

Acknowledgements

We would like to thank all participants, and the NeurIPS competition chairs and organizing committee, for a wonderful event.

References

Angelos Filos, Sebastian Farquhar, Aidan N Gomez, Tim G J Rudner, Zachary Kenton, Lewis Smith, Milad Alizadeh, Arnoud de Kroon, and Yarin Gal. Benchmarking bayesian deep learning with diabetic retinopathy diagnosis. 2019.

Andrew Foong, David Burt, Yingzhen Li, and Richard Turner. On the expressiveness of approximate inference in bayesian neural networks. *Advances in Neural Information Processing Systems*, 33:15897–15908, 2020.

Andrew YK Foong, Yingzhen Li, José Miguel Hernández-Lobato, and Richard E Turner. 'in-between'uncertainty in bayesian neural networks. *arXiv preprint arXiv:1906.11537*, 2019.

Bobby He, Balaji Lakshminarayanan, and Yee Whye Teh. Bayesian deep ensembles via the neural tangent kernel. *Advances in Neural Information Processing Systems*, 33:1010–1022, 2020.

Dan Hendrycks and Thomas Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *ICLR*, March 2019.

José Miguel Hernández-Lobato and Ryan P. Adams. Probabilistic Backpropagation for Scalable Learning of Bayesian Neural Networks. *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

Pavel Izmailov, Patrick Nicholson, Sanae Lotfi, and Andrew G Wilson. Dangers of bayesian model averaging under covariate shift. *Advances in Neural Information Processing Systems*, 34, 2021a.

Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Gordon Wilson. What are bayesian neural network posteriors really like? In *International Conference on Machine Learning*, pages 4629–4640, 2021b.

Emtiyaz Khan. Deep learning with Bayesian principles, 2019. URL https://www.youtube.com/watch?v=2wFb46Q8kmA.

Mohammad Emtiyaz Khan and Håvard Rue. The bayesian learning rule. *arXiv preprint arXiv:2107.04562*, 2021.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. 2017.

Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. 7 (1), 2017.

Wu Lin, Mark Schmidt, and Mohammad Emtiyaz Khan. Handling the positive-definite constraint in the bayesian learning rule. In *International Conference on Machine Learning*, pages 6116–6126. PMLR, 2020.

David JC MacKay. Probable networks and plausible predictions?a review of practical Bayesian methods for supervised neural networks. *Network: computation in neural systems*, 6(3):469–505, 1995.

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

Radford M Neal et al. Mcmc using hamiltonian dynamics. *Handbook of markov chain monte carlo*, 2(11):2, 2011.

R.M. Neal. *Bayesian Learning for Neural Networks*. Springer Verlag, 1996. ISBN 0387947248.

Kazuki Osawa, Siddharth Swaroop, Mohammad Emtiyaz E Khan, Anirudh Jain, Runa Eschenhagen, Richard E Turner, and Rio Yokota. Practical deep learning with bayesian principles. *Advances in neural information processing systems*, 32, 2019.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. 2019.

Tim Pearce, Felix Leibfried, and Alexandra Brintrup. Uncertainty in neural networks: Approximately bayesian ensembling. In *International conference on artificial intelligence and statistics*, pages 234–244. PMLR, 2020.

Dustin Tran, Jasper Snoek, and Balaji Lakshminarayanan. Practical uncertainty estimation and out-of-distribution robustness in deep learning, 2020. URL https://slideslive.com/38935801/practical-uncertainty-estimation-outofdistribution-robustness-in-deep-learning.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688. Citeseer, 2011.

Andrew Gordon Wilson. Bayesian deep learning and a probabilistic perpsective of model construction, 2020. URL https://www.youtube.com/watch?v=E1qhGw8QxqY.

Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. In *Advances in Neural Information Processing Systems*, 2020.

Jiancheng Yang, Rui Shi, and Bingbing Ni. Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. *arXiv preprint arXiv:2010.14925*, 2020.

Jiayu Yao, Weiwei Pan, Soumya Ghosh, and Finale Doshi-Velez. Quality of uncertainty quantification for bayesian neural network inference. *arXiv preprint arXiv:1906.09686*, 2019.