4th Symposium on Advances in Approximate Bayesian Inference, 2021 1–16

Probabilistic Deep Learning with Generalised Variational Inference

Giorgos Felekis GEORGIOS.FELEKIS.19@UCL.AC.UK
University College London

Theodoros Damoulas T.DAMOULAS@WARWICK.AC.UK
University of Warwick

Brooks Paige B.PAIGE@UCL.AC.UK
University College London

Abstract

We study probabilistic Deep Learning methods through the lens of Approximate Bayesian Inference. In particular, we examine Bayesian Neural Networks (BNNs), which usually suffer from multiple ill-posed assumptions such as prior and likelihood misspecification. In this direction, we investigate a recently proposed approximate inference framework called Generalised Variational Inference (GVI) in comparison to state-of-the-art methods including standard Variational Inference, Monte-Carlo Dropout, Stochastic gradient Langevin dynamics and Deep Ensembles. Also, we expand the original research around GVI by exploring a broader set of model architectures and mathematical settings on both real and synthetic data. Our experiments demonstrate that approximate posterior distributions derived from such a method offer attractive properties with respect to uncertainty quantification, prior specification robustness and predictive performance, especially in the case of BNNs. The code for all the experiments can be found in the following public Github repository: https://github.com/gfelekis/GVI-posteriors-in-Probabilistic-Deep-Learning

1. Introduction

Bayesian methods provide the gold standard method to capture uncertainty in deep learning models through their natural probabilistic representation. The success of these methods in practice, including the recent advances of Bayesian Deep Learning field, relies on Approximate Inference methods. These approximation schemes perform Bayesian reasoning efficiently, as they are approximating the posterior distribution either in a stochastic way (MCMC) or in a deterministic one (Variational Inference), thus allowing Bayesian modelling to be applied to many practical tasks. Recently, there's a lot of progress on these ideas mainly through variational techniques (3),(4),(14) and (17). In recent years, due to the availability of massive datasets, the main focus of Variational Inference is on scalable approaches (15), (5), black box algorithms (21) and Bayesian deep learning architectures such as the Variational Autoencoders (VAE) (18). In our work, we motivate the recently proposed framework of Generalised Variational Inference (19) which is a generalisation of standard Bayesian and Variational methods and seems to be able to overcome a lot of drawbacks and pathogenies of them regarding the prior, the likelihood and the computational needs. In this work we advance the research on GVI by exploring a broader set of divergences on more complex Bayesian Neural Network architectures and comparing these with state-of-the-art approximate inference methods.

© G. Felekis, T. Damoulas & B. Paige.

Contributions: We carry out an extensive comparative analysis of GVI among different discrepancy settings and model complexity. We also compare these against multiple Approximate Inference methods ranging from standard Variational Inference to Monte Carlo Dropout (13), Stochastic Gradient Langevin Dynamics (22) and Deep Ensembles. We found certain settings of different divergences, hyperparameters and neural network depths that markedly improve the performance of GVI and provide empirical evidence of the superiority of it over traditional approximate inference methods in the framework of Bayesian Neural Networks. Finally, we also conduct a detailed empirical analysis of uncertainty quantification in a controlled synthetic setting for GVI with a variety of different divergences. We finally evaluate their epistemic and aleatoric uncertainty and asses model calibration.

2. Theoretical Background - The many views of Variational Inference

2.1. ELBO view (Model selection)

It has been shown by Csiszar (6) and Donsker (9) that the Bayesian inference objective can be seen as the solution to an infinite dimensional optimisation problem and therefore every posterior distribution is the result of a well defined problem of this nature (Zellner in 1998 (23)). A few years later, Bissiri et. al (2) extensively discussed a generalised solution inspired by the Fenchel's conjugate of KL divergence and restated the optimisation problem over $P(\Theta)$, about a parameter θ, specified via an abstract loss function l (compared to the previous log-likelihood one) and regularised by the KL divergence, as follows:

$$q^*(\theta) = \underset{q \in P(\Theta)}{\operatorname{argmin}} \left\{ \mathbb{E}_{q(\theta)} \left[\sum_{i=1}^{n} l(\theta, x_i) \right] + KL(q||\pi) \right\}$$

where $q^*(\theta) = q(\theta|\kappa^*)$ for some optimal parameter $\kappa^* \in K$, the variational parameter space. The later is the standard way of deriving Variational Inference, by maximizing the evidence in the data and picking the element from the approximation family $\mathcal{Q}$ that maximizes the well-known Evidence Lower Bound (ELBO).

2.2. Discrepancy-minimisation view (DVI)

The Bayesian posterior is not a unique solution to the optimisation problem and hence, we could see its solution as the minimisation of a distance metric (usually in the form of divergences). Specifically, if we want to approximate the standard Bayesian posterior $q_B^*(\theta)$ with a variational distribution $q(\theta)$ we could write the solution of this problem as: $q^*(\theta) = \underset{q \in \mathcal{Q}}{\operatorname{argmin}} KL(q||q_B^*)$. The fact that the same objective function that maximizes the ELBO is the one that minimizes the distance of $\mathcal{Q}$ and $q_B^*(\theta)$ in the KL divergence sense by just rearranging the terms of the ELBO equation has motivated a large body of research (20), (1), (8) that tries to approximate the posterior by minimizing divergences, different from the KL, between the family $\mathcal{Q}$ and $q_B^*(\theta)$. In particular, for a divergence measure $D : P(\Theta) \times P(\Theta) \to \mathbb{R}_+$ we can define a new class of methods called *Discrepancy Variational Inference (DVI) methods*. DVI methods' objective is of the following form:

$$q^*_{\text{DVI}}(\theta) = \underset{q \in \mathcal{Q}}{\operatorname{argmin}} D(q||q_B^*), \quad \text{D} \neq \text{KL}$$

2.3. Constrained optimisation view (GVI)

Under the traditional Bayesian inference framework, modern machine learning models suffer from certain ill-posed assumptions mainly regarding prior misspecification (11). In order to tackle this Knoblauch et al. (19) proposed the *Rule of Three (RoT)* framework which split the inference problem into three elements: A loss, a divergence and a space of feasible solutions. The RoT foundation provides enough flexibility to tackle most of the inappropriate assumptions especially by its modularity nature. The authors define a generalised representation of Bayesian inference as follows:
Definition: For given observations $x_{1:n}$, a prior $\pi(\theta)$, a space $\Pi \subseteq \mathcal{P}(\Theta)$, a loss function $l : \Theta \times \mathcal{X} \to \mathbb{R}$ and a divergence $D(\cdot||\pi) : \Pi \to \mathbb{R}_+$ we say that posteriors have been constructed via the *Rule of Three* if they can be written as [1]:

$$q^*(\theta) = \underset{q \in \Pi}{\operatorname{argmin}} \left\{ \mathbb{E}_{q(\theta)} \left[\sum_{i=1}^{n} l(\theta, x_i) \right] + D(q||\pi) \right\} := P(l, D, \Pi)$$

The Rule of Three $P(l, D, \Pi)$ has a modular interpretation and is decomposed into three separate parts, each of them serving a specific and separate from each other purpose. Many of the existing approximation methods can be interpreted by RoT e.g. standard VI solves a problem specified by the Rule of Three $P(l, D, \Pi)$ where: $q^*_{\text{VI}}(\theta) := P(\log p(x_i|\theta), KL(q||\pi), \mathcal{Q})$. All the above lead us to the third view of VI, the constrained optimisation view (19) which treats the VI solution as the best $\mathcal{Q}$-constrained solution, where $\mathcal{Q}$ is a variational family.
Definition: Any Bayesian inference method solving a RoT form $P(l, D, \mathcal{Q})$ for $\mathcal{Q} = \{q(\theta|\kappa) : \kappa \in K\} \subseteq P(\Theta)$ is a procedure called *Generalazided Variational Inference (GVI)*. Hence, GVI *like VI* has the form $P(l, D, \mathcal{Q})$ which satisfies the RoT modularity property and *like DVI* targets non-standard posteriors without conflating l and D. Changing the divergence in the GVI-sense affects only the uncertainty quantification and cannot interfere with the way that the best parameter is found (e.g. interfere with the loss function).

3. Experiments and Results

The black box nature of Bayesian Neural Networks (BNN) lurks a high risk of having a misspecified priors. Even in small BNN architectures we define priors over weights and biases without having proper intuition about what they actually express. As a result, such inappropriate assumptions could lead many times to strange posterior distributions. In that sense approximate inference methods which can capture robustness to prior misspecification might actually be "better", at least for the BNN case compared to asymptotically exact inference algorithms. For all these reasons, we motivate the use of GVI to achieve such robustness, which we then test experimentally across a range of different discrepancies and real/synthetic datasets. Our experiments evaluate to what extent switching to a Generalised framework through GVI can actually improve prior misspecification, uncertainty quantification (UQ) and predictive performance on BNNs.

1. We use the symbol P to describe the optimization Problem defined by these three elements.

3.1. Regression on UCI data sets

Initially, we conduct multiple regression experiments on four different data sets from the UCI Machine Learning repository (10) in order to compare different divergence measures and different neural network depths on the GVI setting. Thanks to the modularity of RoT, and as a result of GVI, in order to tackle the issue of prior misspecification in Bayesian Neural Networks we just need to focus on trying different discrepancy measures while keeping the loss function fixed. Indeed, in the experiments that we have conducted we expand the research of (19) to almost all the members of the f-divergence family and also Fisher distance, while we kept the loss function fixed to the usual log-likelihood one. Consequently, we made a straight comparison between the best performing divergences that we found on the previous step, the Standard VI (KL Divergence) and three approximate inference methods: Monte-Carlo Dropout, Deep Ensemble and Stochastic Gradient Langevin Dynamics. In Figure 1, we can see this comparison of the methods across different number of hidden layers for a fixed dataset.

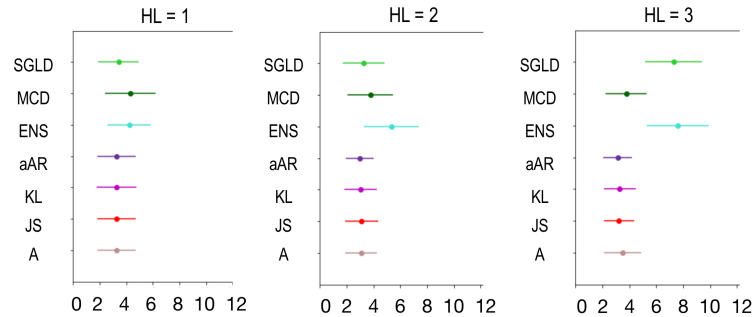

Figure 1: *RMSE values (x-axis) of different methods/divergences across a different number of hidden layers (HL) for the Boston Housing dataset.* We have three approximate inference methods: Stochastic gradient Langevin Dynamics, SGLD (light green), Monte-Carlo Dropout, MCD (green), and Deep Ensembles, ENS (light blue), three GVI based divergences: Parametrized α-Rényi Divergence, aAR (purple) for $\alpha = 2.50$, Jensen-Shannon Divergence, JS (red), α-Divergence, A (light brown) for $\alpha = 2.75$ and the standard VI approach via the KL Divergence, KL (pink). The complete results can be found in Appendix C.

Overall, we empirically demonstrate the superiority of GVI compared to the the other methods in most of the scenarios that we created. From the complete results that can be found in the Appendix C we can observe that GVI outperforms in every divergence the other methods in 9 out of 12 experiments that we conduct on different datasets and model architectures. The model settings can be found in Appendix A.

3.2. Regression on Gaussian Process ground truth

We evaluate aleatoric (noise) and epistemic (model) uncertainty on a GVI setting across different discrepancies, network depths and compare this with the prior art. From (16) we regress on a heteroscedastic toy data set generated from a GP as the ground truth. Based on (7) we decompose uncertainty into aleatoric ($\mathcal{U}_A$) and epistemic ($\mathcal{U}_E$) as follows:

$$\mathcal{U}_A = \mathbb{E}[\sigma^2_{\text{pred}}] = \mathbb{E}_{q(w)}[\mathcal{U}(y'|x', w)], \quad \mathcal{U}_E = Var_{q(w)}(\mu_{\text{pred}}) = \mathcal{U}(y'|x') - \mathcal{U}_A$$

In Figure 2 we focus on the GVI case of Fisher distance and we can observe that in the

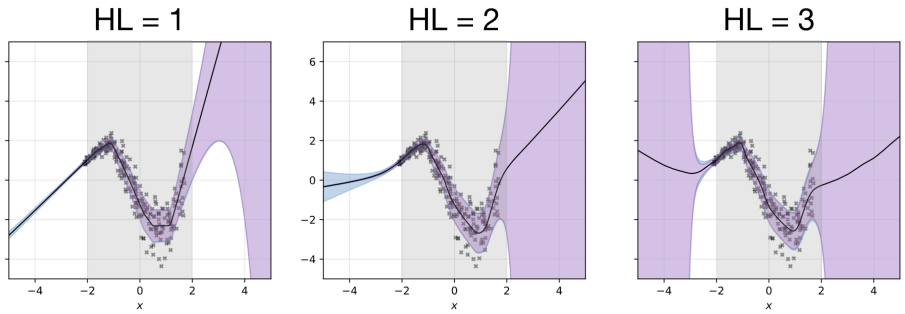

Figure 2: *Aleatoric (purple) and epistemic (blue) uncertainty decomposition across different number of hidden layers (HL) for Fisher distance.* The black line represents the mean of prediction with respect to the variational posterior. The complete results can be found in Appendix C.

interpolation region all the models seem to fit the data really well and performance improves when the number of hidden layers increases. In the *extrapolation regions* we observe two different trends. First, for $\underline{x < -2}$ the models seem to be heavily overconfident in the one hidden layer case, because the heteroscedacity of the data lowers noise variance, and UQ improves as the number of hidden layers increases. For $\underline{x > 2}$ we can observe underconfidence especially when the number of hidden layers is increasing. In this region aleatoric uncertainty dominates the epistemic one.

In Figure 3 we provide an extended comparison among the different methods for the

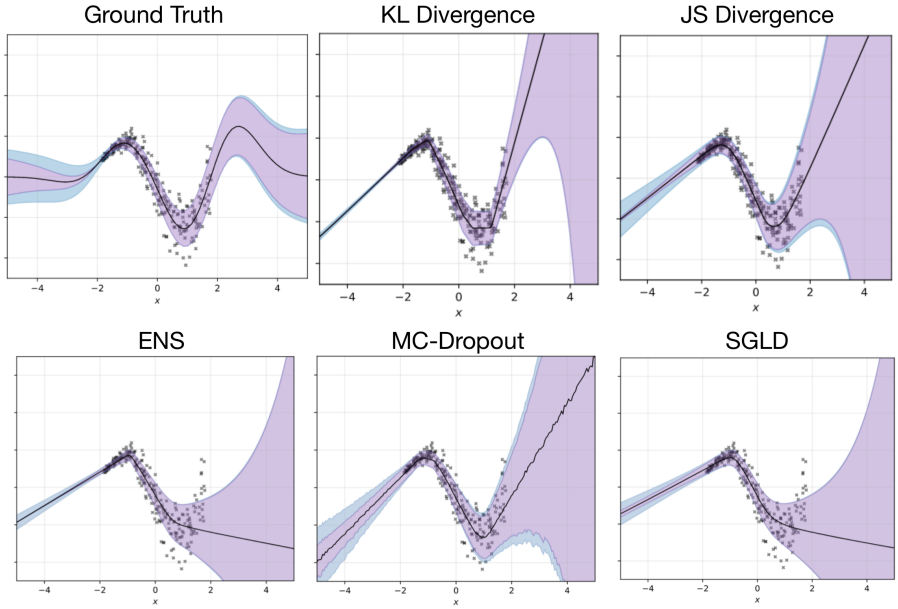

Figure 3: *Aleatoric (purple) and epistemic (blue) uncertainty decomposition across methods for the one hidden layer case.* The black line represents the mean of prediction with respect to the variational posterior. **Top:** From left to right we have, the GP ground truth, Standard VI with KL divergence (KL), GVI with Jensen-Shannon divergence (JS). **Bottom:** From left to right we have, Deep Ensemble (ENS), Monte Carlo Dropout (MC-Dropout) and Stochastic Gradient Langevin Dynamics (SGLD). The complete results can be found in Appendix C.

one hidden layer case and see how GVI behaves compared to other approximate inference methods regarding the UQ. As mentioned before UQ is performed poorly for one hidden layer case in general due to the heteroscedacity of the data which makes all the models overconfident in the left extrapolation area. As expected epistemic uncertainty, which expresses model uncertainty, dominates in this region and thus a more complex model will be more suitable for this evaluation (see Fig. 2). We observe that compared to KL, JS divergence seems to offer both better UQ and balance of aleatoric-epistemic uncertainty in all the regions. It is worth mentioning that KL and JS divergences both belong to the f-family and the later is a normalized and symmetrical version of the first which makes it smoother. For the three approximate inference methods, both Deep Ensembles and SGLD are badly fit the data. They offer a poor UQ and predictive performance (black line) even in the interpolation area, contrary to the MC Dropout model which performs significantly better. MC Dropout efficiently captures aleatoric uncertainty also in the left extrapolation region and does not collapse it compared to other methods apart from JS (GVI). This is probably due to the fact that is know that it's effectively introduces noise (12).
In Table 1 we report the predicted probability difference to ground truth and the effective coverage at the 1σ interval for GVI and MC Dropout methods. Except KL divergence JS, Fisher and MC Dropout provide nice calibration measures. Overall, GVI with JS divergence and MC Dropout seem to perform better compared to all the other methods in UQ, predictive performance and calibration.

Divergence	**No. of hidden layers**	**Predicted probability**	**Difference to $\mathcal{N}(\mu, \sigma^2)$**
F	1	**0.56**	**0.12**
	2	**0.56**	**0.12**
	3	0.83	0.15
KL	1	**0.55**	**0.13**
	2	0.81	0.13
	3	0.53	0.15
JS	1	**0.60**	**0.08**
	2	0.53	0.15
	3	0.52	0.16
MC Dropout	1	**0.60**	**0.08**

Table 1: *Table of differences between the predicted probability and the true probability for each of the divergences for F, KL and JS divergences and MC Dropout.* Note here that under the Gaussian assumption the predicted probability should be 0.68 for 1σ interval. The complete table alongside the calibration curves can be found in Appendix B.2

4. Conclusions

This work offers a comparative analysis of a generalised Bayesian inference framework among different discrepancy settings, datasets, network architectures and other approximate inference methods. We have empirically shown the usefulness of GVI on BNNs in certain settings regarding the model complexity and provide extensive insights of the performance of different divergence measures and its benefits over traditional approximate inference methods. Overall, GVI offers an alternative approach to dealing with the challenging model and prior specification tasks in BNNs. It handles model misspecification via robust scoring functions and prior misspecification through alternative divergences and we argue that these might be preferred in Bayesian Deep Learning.

References

[1] Luca Ambrogioni, Umut Güçlü, Yağmur Güçlütürk, Max Hinne, Eric Maris, and Marcel AJ van Gerven. Wasserstein variational inference. *arXiv preprint arXiv:1805.11284*, 2018.

[2] Pier Giovanni Bissiri, Chris C Holmes, and Stephen G Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 78(5):1103, 2016.

[3] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[4] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622. PMLR, 2015.

[5] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[6] Imre Csiszár. I-divergence geometry of probability distributions and minimization problems. *The annals of probability*, pages 146–158, 1975.

[7] Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning. In *International Conference on Machine Learning*, pages 1184–1193. PMLR, 2018.

[8] Adji B Dieng, Dustin Tran, Rajesh Ranganath, John Paisley, and David M Blei. Variational inference via *chi*-upper bound minimization. *arXiv preprint arXiv:1611.00328*, 2016.

[9] Monroe D Donsker and SR Srinivasa Varadhan. Asymptotics for the wiener sausage. *Communications on Pure and Applied Mathematics*, 28(4):525–565, 1975.

[10] D Dua and C Graff. Uci machine learning repository [http://archive. ics. uci. edu/ml]. irvine, ca: University of california, school of information and computer science. failure to detect type 2 diabetes early costing $700 million per year, diabetes australia, 8 july 2018. *Google Scholar*, 2019.

[11] Vincent Fortuin. Priors in bayesian deep learning: A review. *arXiv preprint arXiv:2105.06868*, 2021.

[12] Y. Gal. What my deep model doesn't know. *http://mlg.eng.cam.ac.uk/yarin/blog*$_3$*d*801*aa*532*c*1*ce.html*, 2015.

[13] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, page 1050–1059, 2016.

[14] Alex Graves. Practical variational inference for neural networks. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper/2011/file/7eb3c8be3d411e8ebfab08eba5f49632-Paper.pdf.

[15] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(5), 2013.

[16] E. Markou X. Zheng J. Antorán, X. Liu. Uncertainty in bayesian neural networks. *https://github.com/JavierAntoran/Bayesian-Neural-Networks*, 2019.

[17] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[18] Diederik P Kingma and Max Welling. An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*, 2019.

[19] Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. Generalized variational inference: Three arguments for deriving new posteriors. *arXiv preprint arXiv:1904.02063*, 2019.

[20] Yingzhen Li and Richard E Turner. Renyi divergence variational inference. *arXiv preprint arXiv:1602.02311*, 2016.

[21] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial intelligence and statistics*, pages 814–822. PMLR, 2014.

[22] Max Welling and Yee Whye Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 681–688, 2011. ISBN 9781450306195.

[23] Arnold Zellner. Optimal information processing and bayes's theorem. *The American Statistician*, 42(4):278–280, 1988.

Appendix A. Regression on UCI data sets

A.1. Model settings & data

Across all the experiments we kept the structure of the neural network fixed in order to make the comparisons as fair as possible. In particular, we use a Multi-Layer Perceptron with [100] hidden units for the one hidden layer case, [100, 100] hidden units for the two hidden layer case and [100, 100, 100] hidden units for the three hidden layer case. The activation function was a ReLU function and inference was performed via Bayes by backprop and the Adam optimiser. For the training of each model we run 100 epochs and perform 30 random splits of each data set with a split of 90%-10% (train-test). All the models were evaluated on the test sets using the average negative log likelihood (NLL) as well as the average root mean square error (RMSE). Also, for each of the 30 splits, the predictions are

computed based on 100 samples from the variational posterior. Note that the priors and variational posteriors are both fully factorised normal distributions and thus our model was predicting the regression mean $\mu(x)$ and the log-standard deviation $\log \sigma(x)$. Also, that helped us of having all of our divergences in a closed Gaussian form. Below we present the closed form of the divergences that were used:

For a prior $\pi \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and approximate posterior $q \sim \mathcal{N}(\mu_2, \sigma_2^2)$ distributions we have the following closed forms:

- Kullback-Leibler Divergence:

$$KL(\pi||q) = \frac{1}{2\sigma_2^2}\big((\mu_1 - \mu_2)^2 + \sigma_1^2 - \sigma_2^2\big) + \ln \frac{\sigma_2}{\sigma_1}$$

- Reverse Kullback-Leibler Divergence:

$$RKL(\pi||q) = KL(q||\pi)$$

- α-Divergence:

$$D_A^{(\alpha)}(\pi||q) = \frac{1}{\alpha(1-\alpha)}\left(1 - \frac{\sigma_2^{\alpha}\sigma_1^{1-\alpha}}{\sqrt{\alpha\sigma_2^2 + (1-\alpha)\sigma_1^2}} e^{-\frac{\alpha(1-\alpha)}{\alpha\sigma_2^2+(1-\alpha)\sigma_1^2}\frac{(\mu_1-\mu_2)^2}{2}}\right)$$

- Jensen-Shannon Divergence:

$$JS(\pi, q) = KL\Big(\pi||\frac{\pi+q}{2}\Big) + KL\Big(q||\frac{\pi+q}{2}\Big)$$

- Total Variation Distance:
  For the TV distance there is no closed form and hence we approximate it by its bounds. In particular, the following inequality holds:

$$\frac{1}{200}\min\left\{1, \max\left\{\frac{|\sigma_1^2 - \sigma_2^2|}{\sigma_1^2}, \frac{40|\mu_1 - \mu_2|}{\sigma_1}\right\}\right\} \leq TV(\pi, q) \leq \frac{3|\sigma_1^2 - \sigma_2^2|}{2\sigma_1^2} + \frac{|\mu_1 - \mu_2|}{2\sigma_1}$$

  Hence we defined an approximation for each bound (upper U, lower L):

$$TVU(\pi, q) = \frac{1}{200}\min\left\{1, \max\left\{\frac{|\sigma_1^2 - \sigma_2^2|}{\sigma_1^2}, \frac{40|\mu_1 - \mu_2|}{\sigma_1}\right\}\right\}$$

$$TVL(\pi, q) = \frac{3|\sigma_1^2 - \sigma_2^2|}{2\sigma_1^2} + \frac{|\mu_1 - \mu_2|}{2\sigma_1}$$

- α-Rényi Divergence:

$$D_{AR}^{(\alpha)}(p||q) = \ln \frac{\sigma_2}{\sigma_1} + \frac{1}{2(\alpha-1)}\ln\big(\frac{\sigma_2^2}{(\sigma^2)_\alpha^*}\big) + \frac{1}{2}\frac{\alpha(\mu_1 - \mu_2)^2}{(\sigma^2)_\alpha^*}$$

  where:

$$(\sigma^2)_\alpha^* = \alpha\sigma_2^2 + (1-\alpha)\sigma_1^2 > 0$$

- Fisher Distance:

$$F(\pi, q) = \sqrt{2} ln\big(\frac{\mathcal{F}((\mu_1, \sigma_1^2), (\mu_2, \sigma_2^2)) + (\mu_1 - \mu_2)^2 + 2(\sigma_1^2 + \sigma_2^2)}{4\sigma_1\sigma_2}\big)$$

where:

$$\mathcal{F}((\mu_1, \sigma_1^2), (\mu_2, \sigma_2^2)) = \sqrt{((\mu_1 - \mu_2)^2 + 2(\sigma_1 - \sigma_2)^2)((\mu_1 - \mu_2)^2 + 2(\sigma_1 + \sigma_2^2))}$$

Appendix B. Regression on Gaussian Process ground truth

B.1. Model settings & data

Here, the heteroscedastic data was generated by a Gaussian Process with an RBF kernel ($l = 1, \sigma_n = 0.3|x + 2|$). Also a Multi-Layer Perceptron was used as the regressor with [100] ReLU hidden units for the one hidden layer case, [100, 200] ReLU hidden units for the two hidden layer case and [100, 200, 100] ReLU hidden units for the three hidden layer case and in all cases it was trained for 500 epochs. We compute epistemic and aleatoric uncertainty and we also investigate the behaviour of all the models across different divergences and different depths by not only visual inspection but also some model selection and information criteria that we are going to see below.

B.2. Model Calibration

Here is the complete table of the differences between the predicted probability and the true probability for each of the divergences.

Divergence	**No. of hidden layers**	**Predicted probability**	**Difference**
KL	1	0.55	0.13
KL	2	0.81	0.13
KL	3	0.53	0.15
RKL	1	0.56	0.12
RKL	2	0.54	0.14
RKL	3	0.52	0.16
A	1	**0.58**	0.10
A	2	**0.66**	0.02
A	3	**0.75**	0.07
AR	1	0.55	0.13
AR	2	0.57	0.11
AR	3	0.84	0.16
αAR	1	0.56	0.12
αAR	2	**0.60**	0.10
αAR	3	**0.68**	0.00

Divergence	No. of hidden layers	Predicted probability	Difference
JS	1	**0.58**	0.10
JS	2	0.53	0.15
JS	3	0.52	0.16
TVL	1	**0.58**	0.10
TVL	2	0.89	0.21
TVL	3	**0.65**	0.03
TVU	1	0.57	0.11
TVU	2	0.56	0.12
TVU	3	0.82	0.14
F	1	0.55	0.13
F	2	0.56	0.12
F	3	0.83	0.15

We highlight the divergence measures which correspond to the two best calibrated models for each neural network depth. Note here that we define a well-calibrated model to be one whose predicted probability differs no more than 0.1 from the true probability:

- For the **one hidden layer** case: TVL and αAR
- For the **two hidden layers** case: KL, AR and A
- For the **three hidden layers** case: TVL, A and JS.

In Figure 4 we can see the calibration curves for each divergence for different number of hidden layers. Each curve represents how confidence varies across different sigma intervals, indicating either underconfidence if the curve is above the dashed line or overconfidence if it falls below it. Two key observations here are that the confidence increases together with the number of layers in almost all the models and that there is a general overconfidence trend on lower sigma intervals i.e. $(0, 0.7)$.

B.3. Model Selection

Finally, one standard way to perform model selection in Probabilistic Deep Learning is to compute the values of certain criteria from the Information Theory field. Specifically here we tested three of them for each model:

1. **Bayesian Information Criterion - BIC**

$$BIC(\mathrm{M}) = p \ln(n) - 2 \ln(\hat{L})$$

2. **Akaike Information Criterion - AIC**

$$AIC(\mathrm{M}) = 2p - 2 \ln(\hat{L})$$

3. **Hannan–Quinn Information Criterion - HQC**

$$HQC(\mathrm{M}) = 2p \ln(\ln(n)) - 2\hat{L}$$

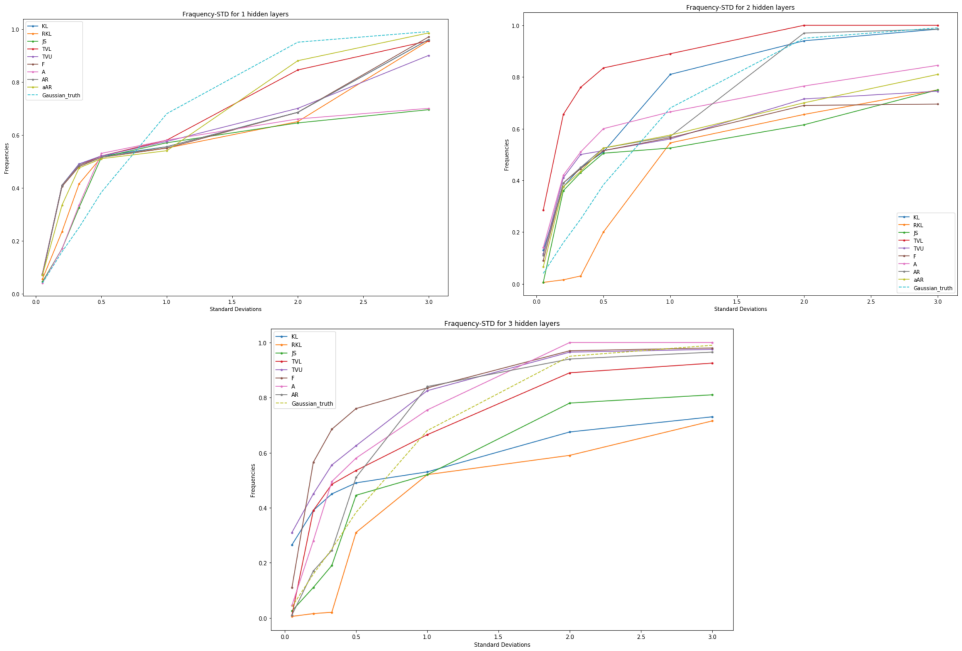

Figure 4: *Calibration curves of different divergences.* The x-axis represents the σ-interval and the y-axis represents the probability inside a σ-interval. The dashed line represents the expected values under a Gaussian distribution ($1\sigma \approx 0.68$, $2\sigma \approx 0.955$ and $3\sigma \approx 0.997$).

where $\hat{L}$ is the maximized value of the likelihood function of the model M, p is the number of parameters estimated by the model and n the sample size. Information Criteria (ICs) are really useful as they can measure the efficiency of a model in terms of predicting the data and also the complexity of the model. We can observe that all three of them are constituted of two terms: a complexity term, which is the one dependent to the number of model's parameters, and a data fit term. Generally, the computation of ICs is a trade-off between these two terms as they try to answer the question: Do I need more data or a more complex model? Therefore, when as a next step we computed all the values of the ICs for all the models we ended up observing that the complexity penalty dominated the IC value and hence the one hidden layer case was always the preferable choice (see table in the Appendix). Here it is important to mention a few things about the information criteria. It would be naive to conclude that the information criteria, which were used here, are useless. Instead, they serve their purpose as they penalise the model for trying to fit more parameters than the needed ones for this specific problem/data set. In the example here, where the data set was generated from an RBF kernel of a Gaussian process which is a stationary kernel, the penalty term dominates. It is not necessarily true that the IC value has to increase as long as the number of parameters grows. As we said, all the ICs perform a trade-off between model complexity and data fit. Here, the data fit component is roughly equivalent for all the different neural network depths, thus the only thing that changes is the number of parameters. Consequently, it makes perfect sense to see this monotonic deterioration. For neural networks, an advantage of having more than one hidden layers is that they can capture non stationary kernels (Neural Network kernel, Polynomial kernel, etc). So we

encourage the reader to examine the same results for a non-stationary case where the ICs would be much more informative than here. Overall, it is crucial to understand that the results have to do with the specific problem, which is simple, and the fact that number of parameters is not a really good complexity value for an MLP. Having said that, it would probably be more fruitful to compare just the differences between the log-likelihoods of the models. In the table below we present the log-likelihood value for each model and we highlight in bold the best for each divergence.

Divergence	**No. of hidden layers**	**Log-likelihood**
KL	1	352
KL	2	347
KL	3	**343**
RKL	1	**380**
RKL	2	390
RKL	3	727
A	1	**398**
A	2	413
A	3	636
AR	1	352
AR	2	347
AR	3	**343**
αAR	1	**348**
αAR	2	361
αAR	3	360
JS	1	**357**
JS	2	368
JS	3	400
TVL	1	**344**
TVL	2	345
TVL	3	347
TVU	1	**343**
TVU	2	**343**
TVU	3	344
F	1	351
F	2	349
F	3	**343**

From this table we can notice that, although some divergences achieve the best log-likelihood value at the maximum number of hidden layers, it is not always the case and it does not justify the increase of the number of model parameters. It is interesting to see that the best model based on the log-likelihood value agrees in most of the cases with our visual inspection conclusion about each layer.

Appendix C. Analytical results

C.1. Regression on UCI data sets

In Figure 4 we can see the comparative analysis of the different GVI settings for the multiple divergences and in Figure 5 the comparison of GVI's top performers against common approximate inference methods:

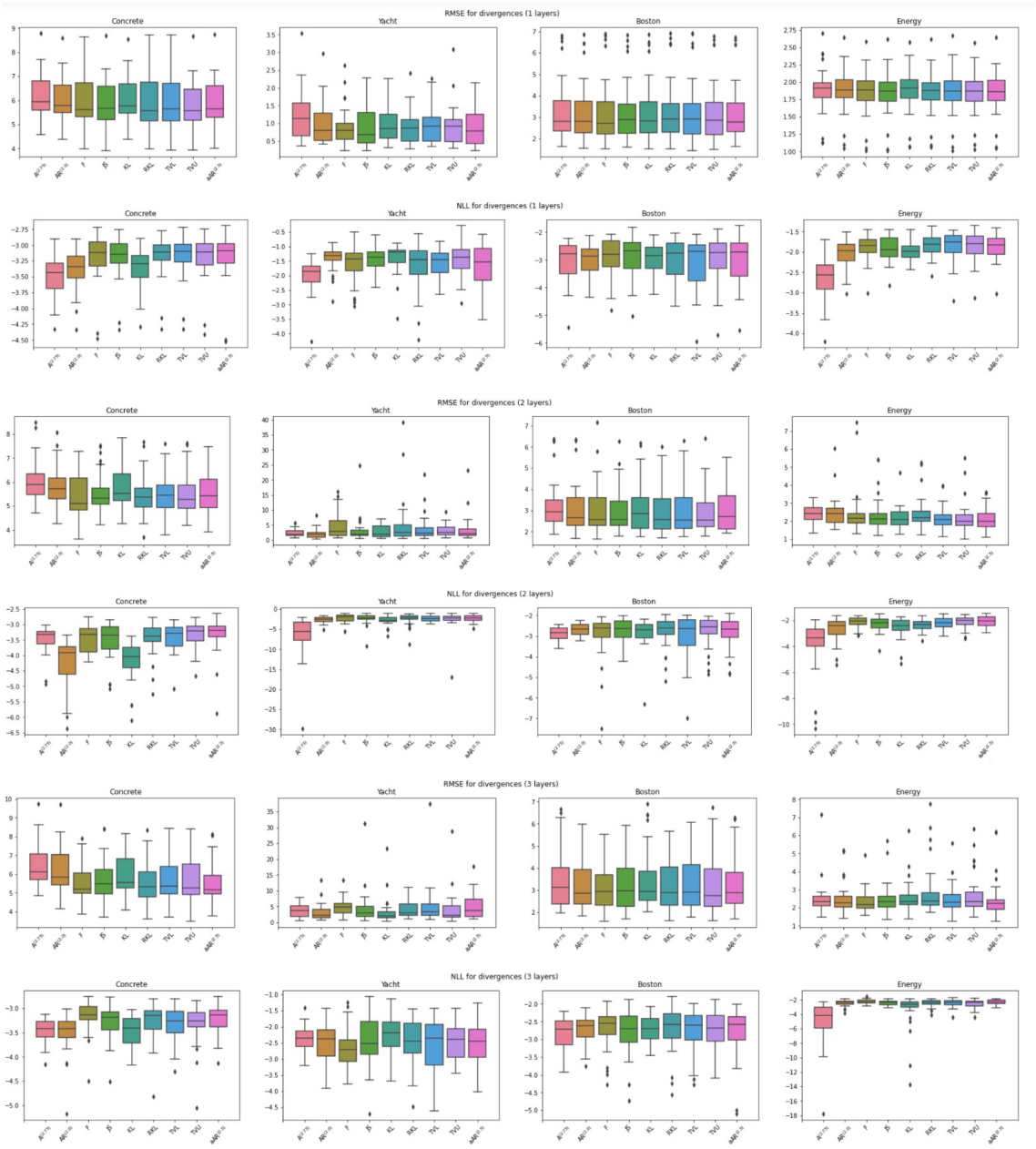

Figure 5

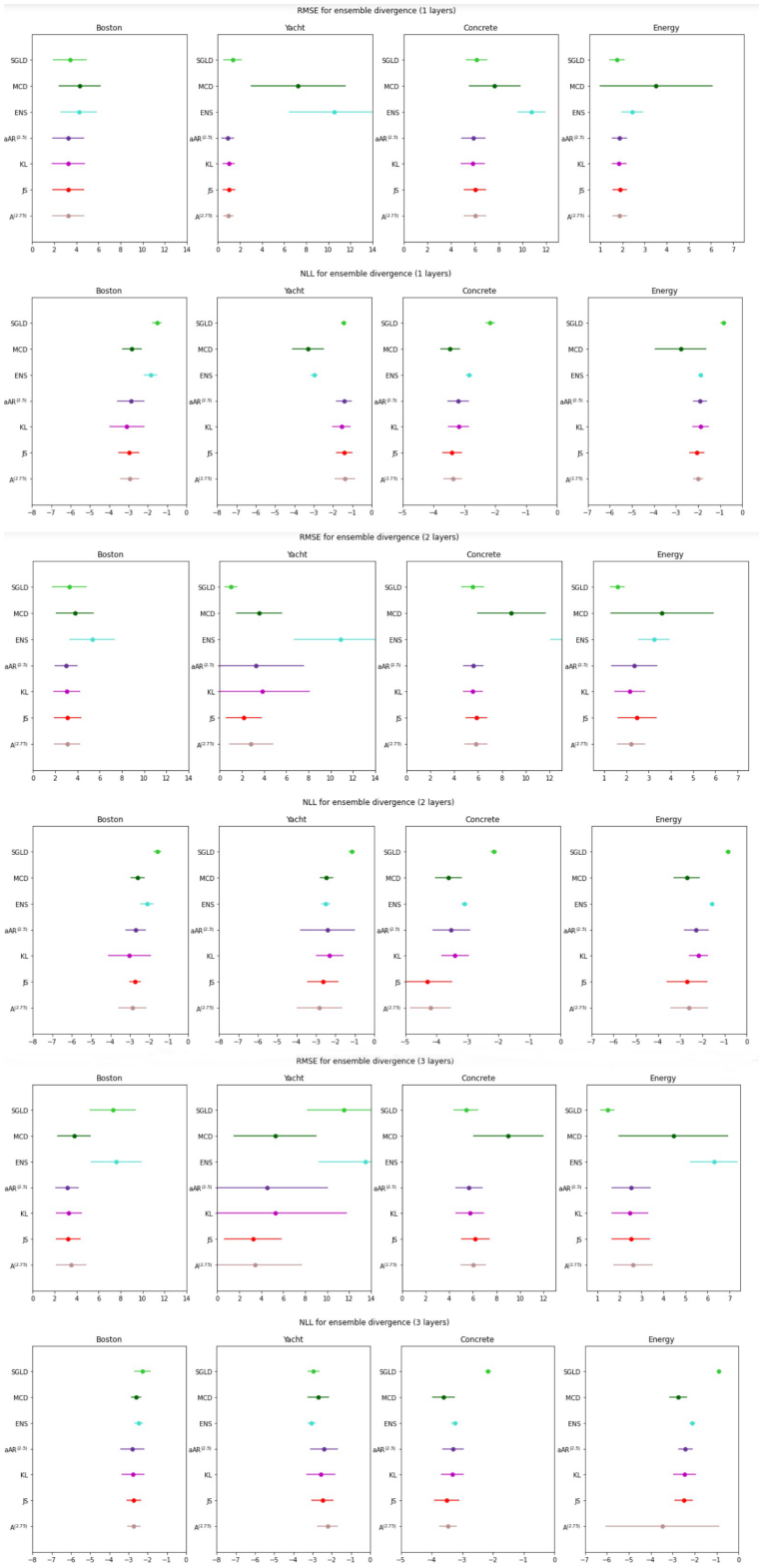
RMSE for ensemble divergence (1 layers)
NLL for ensemble divergence (1 layers)
RMSE for ensemble divergence (2 layers)
NLL for ensemble divergence (2 layers)
RMSE for ensemble divergence (3 layers)
NLL for ensemble divergence (3 layers)
Boston
Yacht
Concrete
Energy
SGLD
MCD
ENS
KL
JS

Figure 6

We can make the following remarks from Figure 6:

- For the **one hidden layer** case: The GVI methods seem to outperform MC-Dropout and Deep Ensembles and be equally good with the SGLD method especially for the RMSE values. MC-Dropout seems to be the worst performer here.
- For the **two hidden layers** case: Here, all the approximate inference methods seem to improve and come closer to the GVI RMSE values in most of the cases. However, we see that Ensemble and SGLD outperform all the other methods when it comes to NLL values.
- For the **three hidden layers** case: We can notice that the uncertainty value (variance) increases in all the methods. GVI remains the best performer and in some cases alongside SGLD and also improves its performance in the NLL scores.

C.2. Regression on Gaussian Process ground truth

In Figures 6 and 7 are the complete results from the regression workaround on the Gaussian Process ground truth data set. Also, here we provide the best models based on visual inspection:

- For the **one hidden layer** case: JS, KL and A and αAR
- For the **two hidden layers** case: KL, F and TVU
- For the **three hidden layers** case: TVU, A and JS.

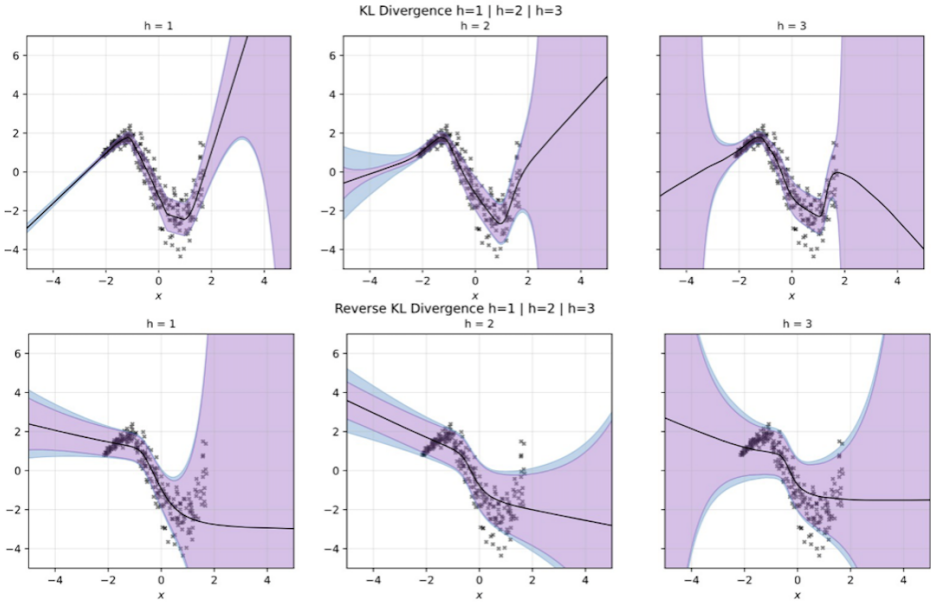

Figure 7

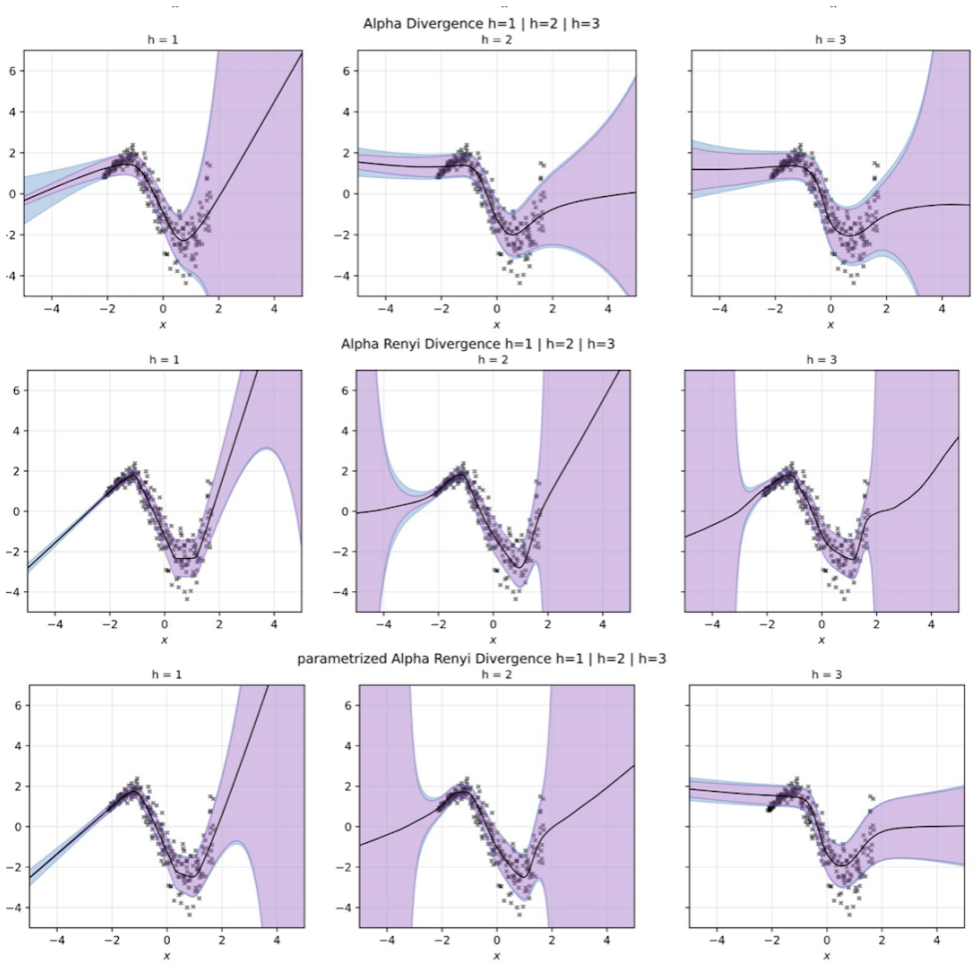
Alpha Divergence h=1 | h=2 | h=3
h = 1
h = 2
h = 3
Alpha Renyi Divergence h=1 | h=2 | h=3
h = 1
h = 2
h = 3
parametrized Alpha Renyi Divergence h=1 | h=2 | h=3
h = 1
h = 2
h = 3
x

Figure 8

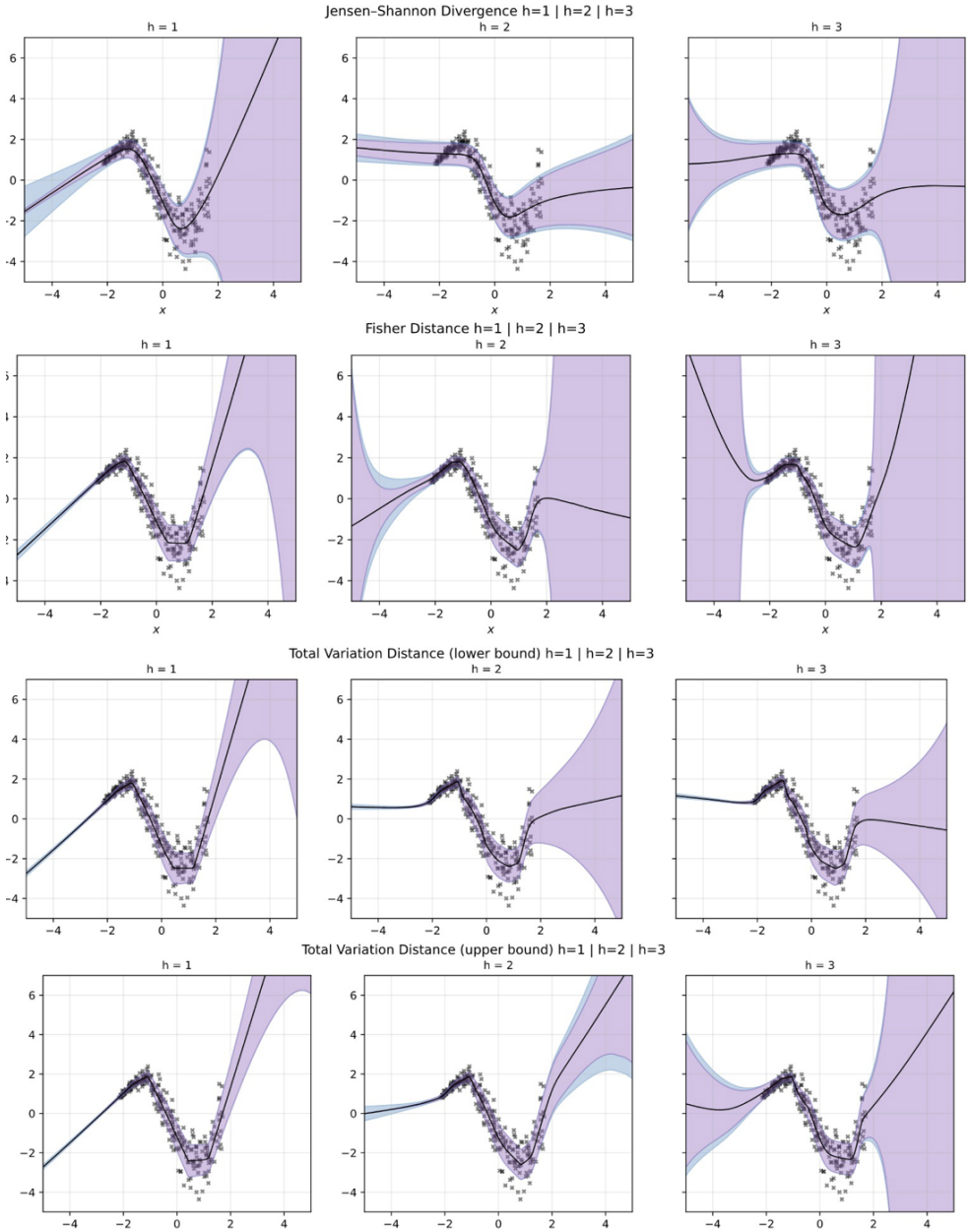
Jensen–Shannon Divergence h=1 | h=2 | h=3
h = 1
h = 2
h = 3
Fisher Distance h=1 | h=2 | h=3
Total Variation Distance (lower bound) h=1 | h=2 | h=3
Total Variation Distance (upper bound) h=1 | h=2 | h=3
x

Figure 9