Generalised Variational Inference posteriors in Probabilistic Deep Learning

Giorgos Felekis
MSc Machine Learning

Supervisors: Theo Damoulas, Brooks Paige

Structure of the talk

1. Motivation
2. Uncertainty in Deep Learning
3. Generalised Variational Inference
4. Experiments and Analysis
5. Conclusion

1. Motivation

Why should we care about uncertainty?

Many applications of machine learning depend on good estimation of the uncertainty:

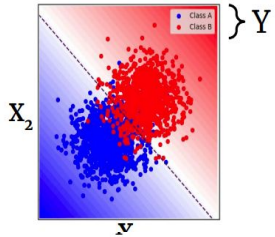
- Forecasting
- Decision making
- Learning from limited, noisy, and missing data
- Out-of-distribution detection
- Learning complex personalised models
- Automating scientific modelling, discovery, and experiment design
- Explore/Exploit dilemma in reinforcement learning

Nearly all applications!

What do we mean by Uncertainty?

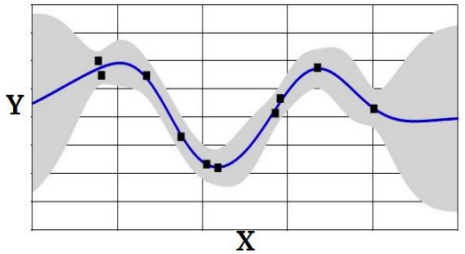
Return a distribution over predictions rather than a single prediction.

- Classification: Output label along with its confidence.



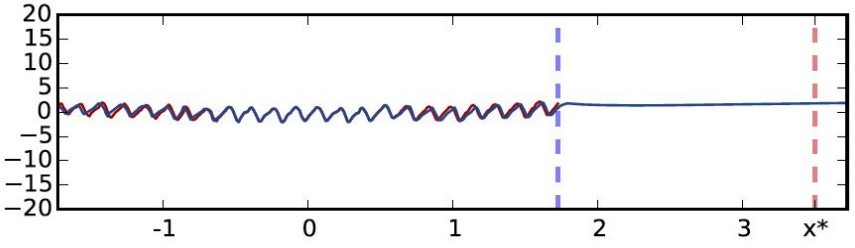
$$p(\mathbf{y}|\mathbf{x})$$

- Regression: Output mean along with its variance.

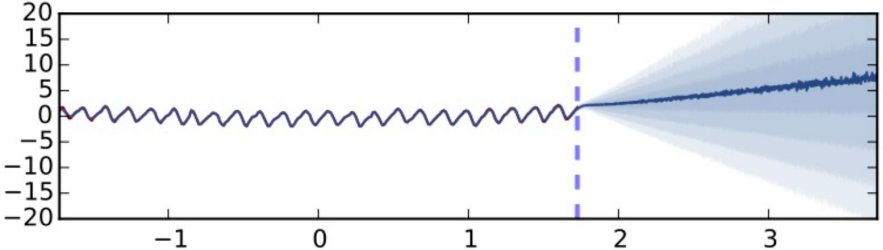


Good uncertainty estimates quantify when we can trust the model's predictions

What do we mean by Uncertainty?



(a) Standard deep learning model



(b) Probabilistic model

Y.Gal Uncertainty in Deep Learning, Thesis 2016

Types of Uncertainty

We have two types of uncertainty:

- Aleatoric uncertainty:

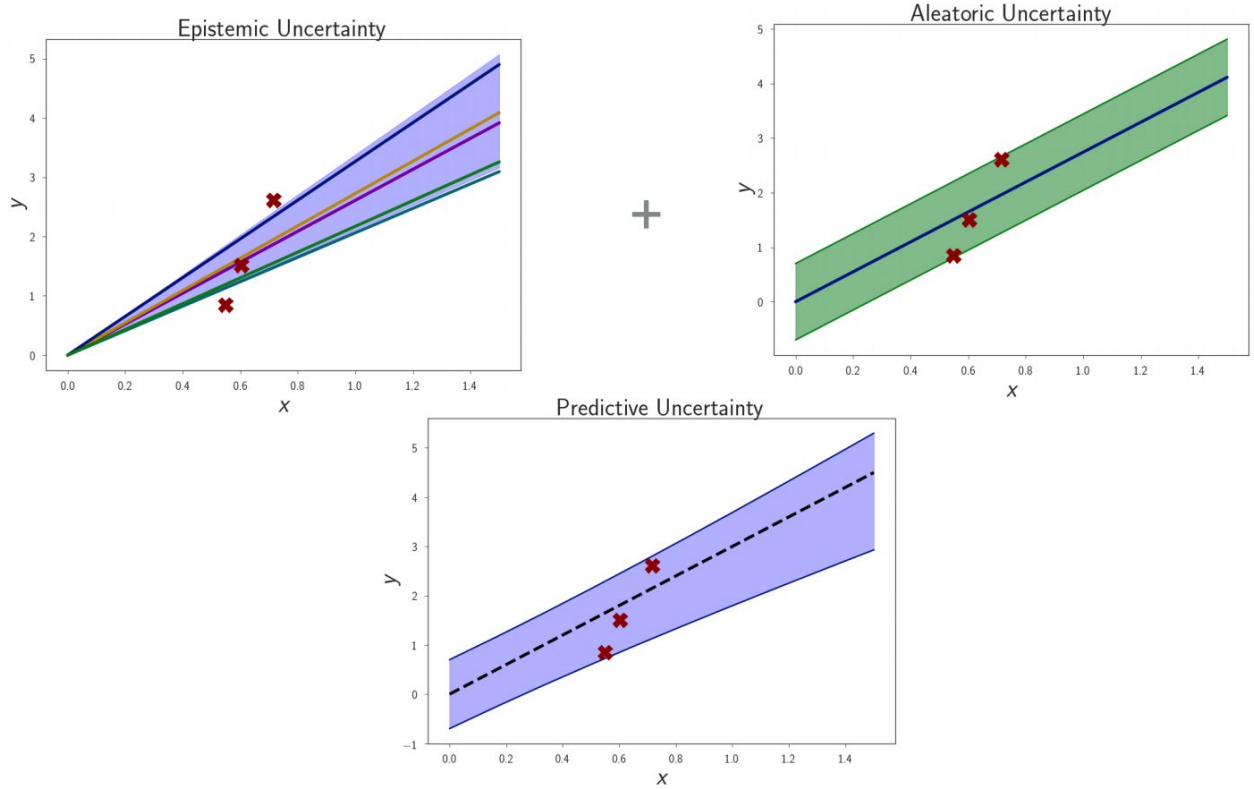
  Arises from the natural stochasticity of observations. In other words, it is the uncertainty that is increased when our observed labels are noisy.

- Epistemic uncertainty:

  Accounts for uncertainty in the model and is due to limited data and knowledge. A larger number of data points is able to decrease this kind of uncertainty.

Overall, aleatoric and epistemic uncertainty can then together express the overall predictive uncertainty, and hence the confidence we have in our predictions.

Types of Uncertainty



Overall,

Uncertainty information is really important for the practitioner!

You might use deep learning models on a daily basis to solve different tasks in vision or linguistics.

Understanding if your model is under-confident or falsely overconfident can help you get better performance out of it by answering the following questions:

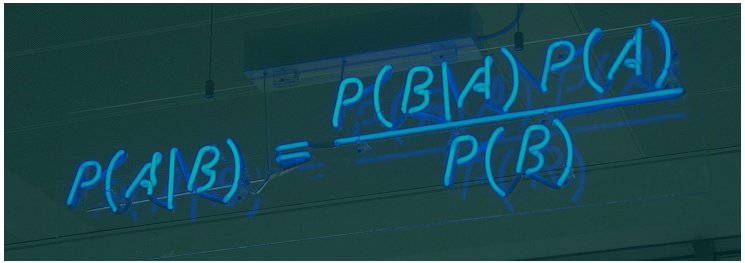
- How reliable are my observations?
- Should we use more diverse data?
- Should we change the model structure?

Surprisingly, we can use Bayesian modelling to answer the questions above!

2. Uncertainty in Deep Learning

Bayes Rule

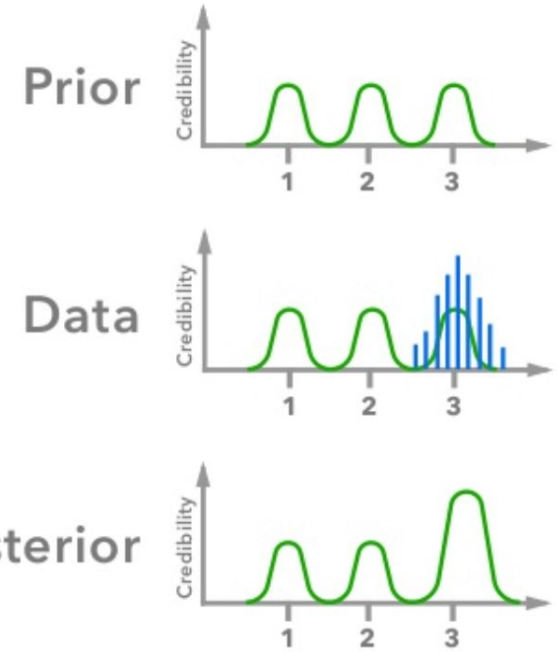
1. We have a prior about the world (a prior distribution over parameters).
2. We update our understanding of the world with the likelihood of events.
3. We obtain a new belief about the world (posterior) by computing it or approximating it (sampling).



Prior, Likelihood, Posterior

$$\overbrace{p(\theta|\mathbf{X})}^{\text{Posterior}} \propto \underbrace{p(\mathbf{X}|\theta)}_{\text{Likelihood}} \overbrace{\pi(\theta)}^{\text{Prior}}$$

posterior is proportional to likelihood times prior!



Bayesian Learning in one slide

- Observed inputs $\mathbf{X} = \{\mathbf{x}_i\}_{i=1}^N$ and outputs $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^N$
- Capture stochastic process believed to have generated outputs
- Def. ω model parameters as r.v.
- Prior dist. over ω: $p(\omega)$
- Likelihood: $p(\mathbf{Y}|\omega, \mathbf{X})$
- Posterior: $p(\omega|\mathbf{X}, \mathbf{Y}) = \frac{p(\mathbf{Y}|\omega,\mathbf{X})p(\omega)}{p(\mathbf{Y}|\mathbf{X})}$ (Bayes' theorem)
- Predictive distribution given new input $\mathbf{x}^*$

$$p(\mathbf{y}^*|\mathbf{x}^*, \mathbf{X}, \mathbf{Y}) = \int p(\mathbf{y}^*|\mathbf{x}^*, \omega) \underbrace{\boxed{p(\omega|\mathbf{X}, \mathbf{Y})}}_{\text{posterior}} \mathrm{d}\omega$$

- But... $p(\omega|\mathbf{X}, \mathbf{Y})$ is often intractable

Everything follows from the sum and product rules:

$$p(X) = \sum_Y p(X, Y)$$

$$p(X, Y) = p(Y|X)p(X)$$

Bayesian Deep Learning

Theorem: A neural network with one hidden layer, infinitely many hidden units and Gaussian priors on the weights is a Gaussian process (Neal, 1994).

Also, in the case that the parameter space is finite then we can still obtain model uncertainty in the form of a model called **Bayesian Neural Network (BNN).**

- In Bayesian deep learning we model posterior distribution over the weights of neural networks.
- In theory, leads to better predictions and well-calibrated uncertainty.

In other words in a Bayesian Neural Networks we introduce a prior distribution on the weights $\mathbf{p(W)}$ and obtain the posterior distribution $\mathbf{p(W|D)}$ through Bayesian learning.

Bayesian Deep Learning

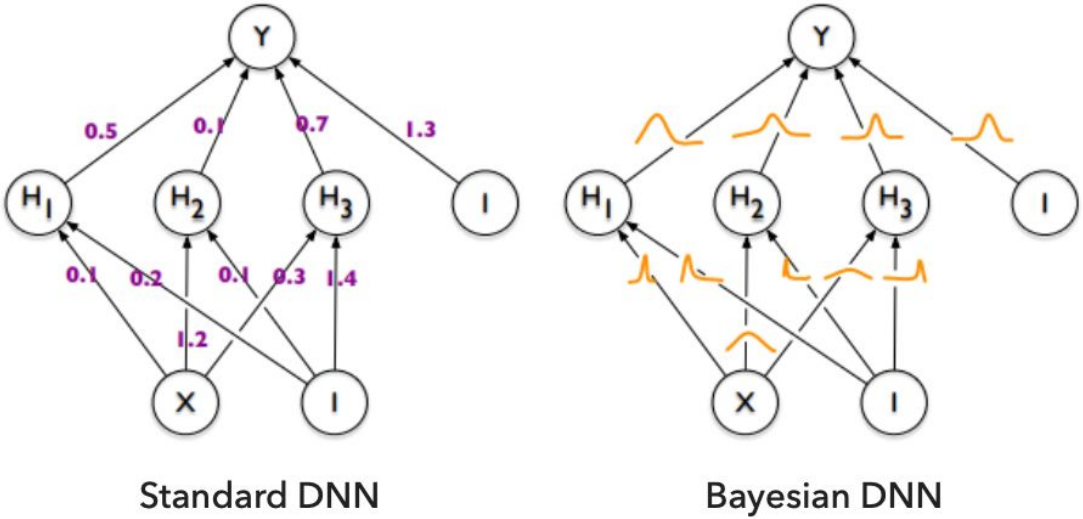
So, instead of having the parameters and the predictions to be represented by single, fixed values (point estimates), they are represented by distributions.

BNN == Marginalisation

Finally, in Bayesian Neural Networks regularisation arises naturally through the prior $\mathbf{p(W)}$. Specifically, it has been proved that:

- A **Uniform** prior choice leads to a Maximum Likelihood training.
- A **Laplace** prior choice leads to a training with L1-regularisation.
- A **Gaussian** prior choice leads to a training with L2-regularisation.

Bayesian Deep Learning



Standard DNN

Bayesian DNN

Variational Inference

As we said the denominator (posterior) of

$$p(\omega|\mathbf{X}, \mathbf{Y}) = \frac{p(\mathbf{Y}|\omega,\mathbf{X})p(\omega)}{p(\mathbf{Y}|\mathbf{X})}$$

is not tractable.

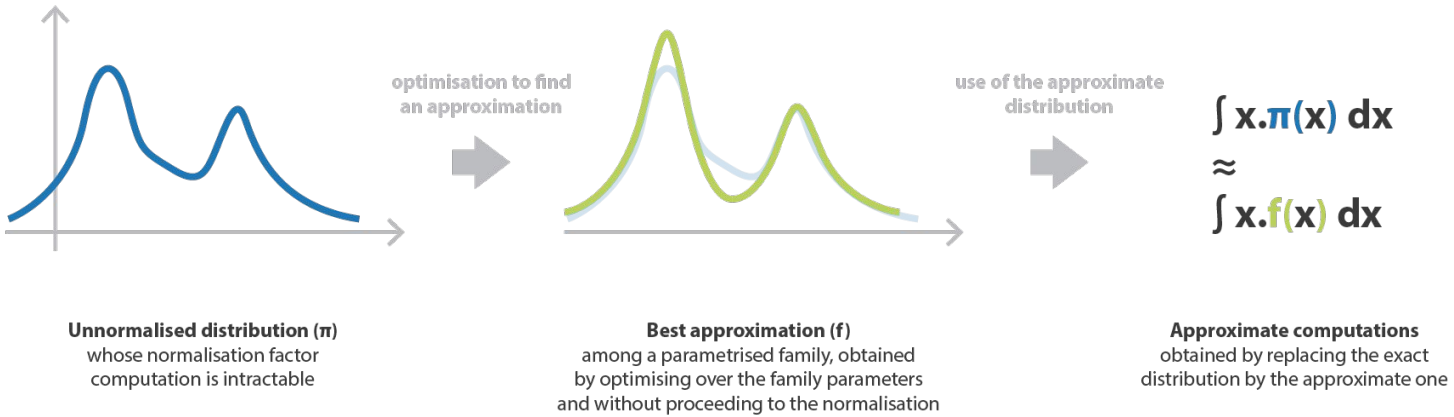
Thus, it is important to find other ways to approximate the true posterior distribution.

- Markov Chain Monte Carlo (MCMC) sampling methods

They are computationally expensive!

Variational Inference

- VI is often used as an alternative to MCMC;
- Can be used to approximate the posterior of Bayesian models with a simpler distribution;
- Faster than MCMC for complex models and larger datasets;
- Instead of sampling, VI is based on optimisation;



Variational Inference - Main Idea

For observations $D = \{x_1, x_2, ..., x_n\}$ and latent variables $Z = \{z_1, z_2, ..., z_m\}$:

1. We pick a restricted family of distributions $\mathbf{Q}$ (approximate densities) over the latent variables with each own parameters,

$$\mathbf{q(Z|k)}$$

2. Then, we seek the member of the family that makes $\mathbf{q}$ close enough to the true posterior.
3. Finally, we approximate the posterior with the optimal member of the family $\mathbf{q^*(\ )}$.

Variational Inference - KL Divergence

In Variational Inference we measure the closeness of two distributions with a statistical measure of discrepancy called Kullback-Leibler Divergence (KL).

The KL divergence for Variational inference is given by the following equation:

$$KL(q||p) = \mathbb{E}_q \left[log \frac{q(Z)}{P(Z|D)} \right] = \int_Z q(z) \log \frac{q(z)}{p(z|x)} dz$$

We can easily see that: $KL(q||p) = 0 \iff q \equiv p$

Hence, we want to find the member of $\mathbf{Q}$ that minimizes the KL divergence to the exact posterior:

$$q^*(z) = \underset{q \in \mathcal{Q}}{\operatorname{argmin}} KL(q(Z)||p(Z|D))$$

Variational Inference - ELBO

The truth is that we actually cannot minimize the KL divergence direct, and thus we minimize a function that is equal to it up to a constant. This is the **evidence lower bound (ELBO).**

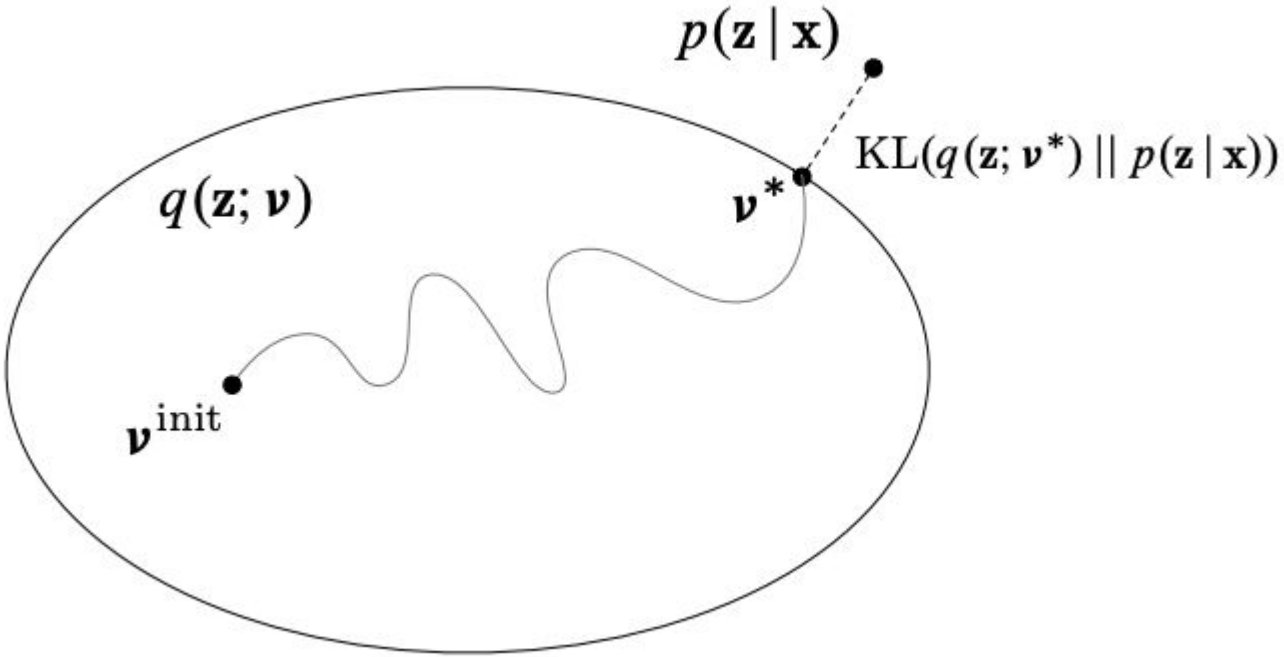
$$KL(q||p) = \mathbb{E}_q\left[\log \frac{q(Z)}{P(Z|D)}\right] = \mathbb{E}_q[\log q(Z)] - \mathbb{E}_q[\log p(Z|D)]$$

$$= \mathbb{E}_q[\log q(Z)] - \mathbb{E}_q[\log p(Z, D)] + \log p(D)$$

$$ELBO(q) = \mathbb{E}_q[\log p(Z, D)] - \mathbb{E}_q[\log q(Z)]$$

Minimizing KL == Maximizing ELBO

Variational Inference - Summary



3. Generalised Variational Inference

Generalized Variational Inference: Three arguments for deriving new Posteriors

Jeremias Knoblauch J.KNOBLAUCH@WARWICK.AC.UK
The Alan Turing Institute
Dept. of Statistics
University of Warwick
Coventry, CV4 7AL, UK

Jack Jewson J.E.JEWSON@WARWICK.AC.UK
The Alan Turing Institute
Dept. of Statistics
University of Warwick
Coventry, CV4 7AL, UK

Theodoros Damoulas T.DAMOULAS@WARWICK.AC.UK
The Alan Turing Institute
Depts. of Computer Science & Statistics
University of Warwick
Coventry, CV4 7AL, UK

Editor: Leslie Pack Kaelbling

Abstract

In this paper we advocate an optimization-centric view on Bayesian statistics and introduce a novel generalization of Bayesian inference. On both counts, our inspiration is the representation of Bayes' rule as an infinite-dimensional optimization problem as shown independently by Csiszár (1975); Donsker and Varadhan (1975); Zellner (1988). First, we use this representation to prove a surprising optimality result of standard Variational Inference (VI) methods: Under the proposed view, the standard Evidence Lower Bound (ELBO) maximizing VI posterior is always preferable to alternative approximations of the Bayesian posterior. Next, we argue for an optimization-centric generalization of standard Bayesian inference. The need for this generalization arises in situations of severe misalignment between reality and three assumptions underlying the standard Bayesian posterior: (1) Well-specified priors, (2) well-specified likelihood models and (3) the availability of infinite computing power. In response to this observation, our generalization is defined by three arguments and named the Rule of Three (RoT). Each of its three arguments relaxes one of the assumptions underlying standard Bayesian inference. We axiomatically derive the RoT and recover existing methods as special cases, including the Bayesian posterior and its approximation by standard Variational Inference (VI). In contrast, alternative approximations to the Bayesian posterior maximizing other ELBO-like objectives violate these axioms. Finally, we introduce a special case of the RoT that we call Generalized Variational Inference (GVI)

Divergences

Definition 1: A statistical divergence $D(p||q)$ is a measure of discrepancy between two probability densities p and q on the same parameter space Θ, or more generally on the same statistical manifold, with the following properties:

1. $D(p||q) \geq 0 \;\; \forall q \in P(\Theta)$
2. $D(p||q) = 0 \iff p = q$

Definition 2: For a convex function f such that $f(1) = 0$, the **f-divergence** of P from Q is defined as:

$$D_f(P||Q) = \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right] = \int_\Omega f\left(\frac{dP}{dQ}\right) dQ$$

Divergences

Kullback-Leibler Divergence

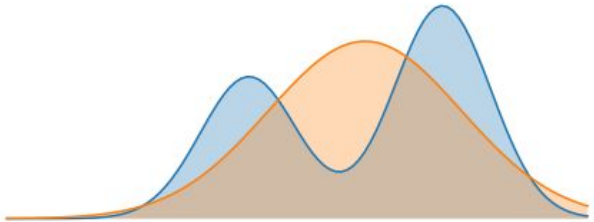
For $f(x) = x \log(x)$,

$$KL(p||q) = \mathbb{E}_Q[\log p(x) - \log q(x)] = \int_{-\infty}^{+\infty} p(x) log\Big(\frac{p(x)}{q(x)}\Big) dx$$

Reverse Kullback-Leibler Divergence

For $f(x) = -\log(x)$,

$$RKL(p||q) = KL(q||p) = \int_{-\infty}^{+\infty} q(x) \log\Big(\frac{q(x)}{p(x)}\Big) dx$$



Jensen-Shannon Divergence

For $f(x) = x \log(\frac{2x}{x+1}) + \log(\frac{2}{x+1})$,

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

Total Variation Distance

For $f(x) = \frac{1}{2}|x-1|$,

$$TV(p,q) = \frac{1}{2}\,\mathbb{E}_Q\left[\left|\frac{dp}{dq} - 1\right|\right] = \frac{1}{2}\int |dp - dq|$$

Divergences

α-Divergence

For $f(x) = -\log_{\alpha}(x)$,

$$D_A^{(\alpha)}(p||q) = \mathbb{E}_q[\log_{\alpha}(p(x) - log_{\alpha}q(x)]$$

where $\log_{\alpha}(x)$ is a generalised log function with $\log_{\alpha}(x) = \frac{1}{\alpha(1-\alpha)}(x^{\alpha-1} - 1)$

For certain values of α in the α-Divergence we can retrieve other members of the f-divergences family. Specifically,

- For $\alpha = \frac{1}{2}$ we get the squared **Hellinger distance**:

$$D_A^{(1/2)}(p||q) = 2D_H(p||q) = 2\int(\sqrt{p(x)} - \sqrt{q(x)})^2 dx$$

- For $\alpha = 2$ we get the **Pearson χ^2-square divergence**:

$$D_A^{(2)}(p||q) = 2\chi^2(p||q) = \frac{1}{2}\int\frac{(p(x) - q(x))^2}{q(x)}dx$$

- We can easily see that when the limit is evaluated (using the L'Hopital's rule) for $\alpha \to 1$, we obtain the Kullback–Leibler divergence. In other words,

$$\lim_{\alpha\to 1} D_A^{(\alpha)}(p||q) = KL(p||q)$$

Divergences

α-Rényi Divergence

We define a divergence called α–Rényi divergence as follows:

$$D_{AR}^{(\alpha)}(p||q) = \frac{1}{(1-\alpha)} \log \left(\mathbb{E} \left[\left(\frac{p(x)}{q(x)} \right)^{\alpha-1} \right] \right)$$

$$\implies D_{AR}^{(\alpha)}(p||q) = \frac{1}{(1-\alpha)} \log \int p(x) q(x)^{1-\alpha} dx$$

Surprisingly, the Renyi divergence is closely related to the α-divergence:

$$D_{AR}^{(\alpha)}(p||q) = \frac{1}{\alpha(1-\alpha)} \log \left(1 + \alpha(\alpha-1) D_{A}^{(\alpha)}(p||q)\right)$$

$$\implies D_{AR}^{(\alpha)}(p||q) = \frac{1}{\alpha(1-\alpha)} \log \left(\int (p^{\alpha} q^{1-\alpha} - \alpha p + (\alpha-1) q) dx + 1 \right), \alpha \in \mathbb{R} \backslash \{0, 1\}$$

An interesting observation here is that for $\alpha = 1$ and $\alpha = 0$ the α-Renyi divergence simplifies to Kullback–Leibler and Reverse Kullback–Leibler divergences respectively.

$$KL(p||q) = \lim_{\alpha \to 1} D_{AR}^{(\alpha)}(p||q)$$

$$RKL(p||q) = KL(q||p) = \lim_{\alpha \to 0} D_{AR}^{(\alpha)}(p||q)$$

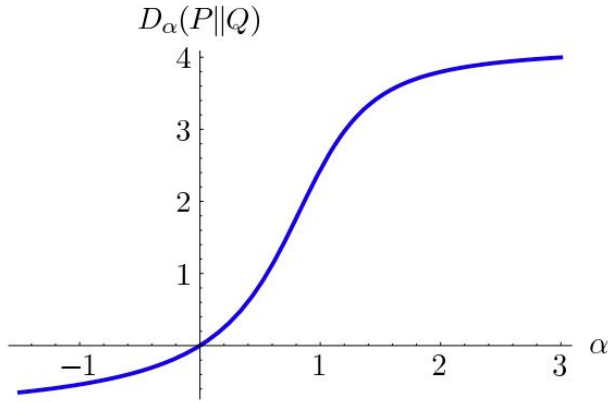
Also,

- For $\alpha = \frac{1}{2}$, we can retrive a function of the square Hellinger distance:

$$-2 \log \left(1 - D_H[p||q]\right)$$

- For $\alpha = \frac{1}{2}$, we can retrive a function of the χ^2-divergence:

$$-\log \left(1 - \chi^2[p||q]\right)$$



. **Rényi divergence as a function of its order for fixed distributions**

Notation:

We are going to follow the notation below for the rest of the talk:

- Θ is the **parameter space** and $\theta \in \Theta$ the parameter value.
- $P(\Theta)$ is the set of all probability measures on Θ.
- $\mathcal{Q}$ is a **variational family** (i.e. parametrised subset of $P(\Theta)$).
- $q : \Theta \to \mathbb{R}_+$ is the **posterior** density (i.e., known after data are seen).
- $\pi : \Theta \to \mathbb{R}_+$ is the **prior** density (i.e., known before data are seen).

Rule of Three: A new Bayesian paradigm

The traditional Bayesian inference paradigm suffers from some crucial problems which make it harder to tackle complex real-world problems. In particular, the main issues that modern machine learning models suffer from under the standard Bayesian inference framework are the following:

1. Prior misspecification.
2. Likelihood misspecification.
3. Computational limitations.

Rule of Three: A new Bayesian paradigm

We recall the ELBO formula from before:

$$ELBO(q) = \mathbb{E}[\log p(D|Z)] - KL(q(Z)||p(Z))$$

We can rewrite it in the negative form:

$$-ELBO(q) = -\mathbb{E}[\log p(D|Z)] + KL(q(Z)||p(Z))$$

Zellner et. al (1984) proposed that the Bayesian posterior can be computed from the following equation:

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

Rule of Three: A new Bayesian paradigm

Bissiri et. al extensively discussed a generalised Bayesian solution of Zellner's equation inspired by the Fenchel's conjugate of KL divergence and restated the problem as:

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

Rule of Three: A new Bayesian paradigm

Definition 3: 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:

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

- $\Pi \subseteq P(\Theta)$ is the set of the admissible posterior beliefs. It answers the question *"Which beliefs are allowed?"*. If Π is a variational family then we write $\Pi = \mathcal{Q}$
- The divergence measure $D : P(\Theta) \times P(\Theta) \to \mathbb{R}_+$ is acting as the uncertainty quantifier/ regularizer and gives us information about how q^* looks
- The loss $l : \Theta \times \mathcal{X} \to \mathbb{R}$ tells us which parameter θ we care about.

Rule of Three for existing methods

Method	$\ell(\boldsymbol{\theta}, x_i)$	D	Π
Standard Bayes	$-\log p(x_i\|\boldsymbol{\theta})$	KLD	$\mathcal{P}(\boldsymbol{\Theta})$
Power Likelihood Bayes[1]	$-\log p(x_i\|\boldsymbol{\theta})$	$\frac{1}{w}$KLD, $w < 1$	$\mathcal{P}(\boldsymbol{\Theta})$
Composite Likelihood Bayes[2]	$-w_i \log p(x_i\|\boldsymbol{\theta})$	KLD	$\mathcal{P}(\boldsymbol{\Theta})$
Divergence-based Bayes[3]	divergence-based ℓ	KLD	$\mathcal{P}(\boldsymbol{\Theta})$
PAC/Gibbs Bayes[4]	any ℓ	KLD	$\mathcal{P}(\boldsymbol{\Theta})$
VAE[5,†]	$-\log p_{\boldsymbol{\zeta}}(x_i\|\boldsymbol{\theta})$	KLD	$\mathcal{Q}$
β-VAE[6,†]	$-\log p_{\boldsymbol{\zeta}}(x_i\|\boldsymbol{\theta})$	$\beta \cdot$ KLD, $\beta > 1$	$\mathcal{Q}$
Bernoulli-VAE[7,†]	continuous Bernoulli	KLD	$\mathcal{Q}$
Standard VI	$-\log p(x_i\|\boldsymbol{\theta})$	KLD	$\mathcal{Q}$
Power VI[8]	$-\log p(x_i\|\boldsymbol{\theta})$	$\frac{1}{w}$KLD, $w < 1$	$\mathcal{Q}$
Utility VI[9]	$-\log p(x_i\|\boldsymbol{\theta}) + \log u(h, x_i)$	KLD	$\mathcal{Q}$
Regularized Bayes[10]	$-\log p(x_i\|\boldsymbol{\theta}) + \phi(\boldsymbol{\theta}, x_i)$	KLD	$\mathcal{Q}$
Gibbs VI[11]	any ℓ	KLD	$\mathcal{Q}$
Generalized VI	any ℓ	any D	$\mathcal{Q}$

Table 1: Relationship of $P(\ell, D, Q)$ to a selection of existing methods. [1](e.g. Grünwald, 2011, 2012; Holmes and Walker, 2017; Grünwald and Van Ommen, 2017; Miller and Dunson, 2019), [2](e.g. Varin et al., 2011; Pauli et al., 2011; Ribatet et al., 2012; Hamelijnck et al., 2019), [3](e.g. Hooker and Vidyashankar, 2014; Ghosh and Basu, 2016; Futami et al., 2018; Jewson et al., 2018; Chérief-Abdellatif and Alquier, 2019), [4](Bissiri et al., 2016; Germain et al., 2016; Guedj, 2019), [5](Kingma and Welling, 2013), [6](Higgins et al., 2017), [7](Loaiza-Ganem and Cunningham, 2019) [8](e.g. Yang et al., 2017; Huang et al., 2018) [9](e.g. Kuśmierczyk et al., 2019; Lacoste-Julien et al., 2011) [10](Ganchev et al. (2010), but only if the regularizer can be written as $\mathbb{E}_{q(\boldsymbol{\theta})}[\phi(\boldsymbol{\theta}, \boldsymbol{x})]$ as in Zhu et al. (2014)), [11](e.g. Alquier et al., 2016) [†]For the VAE entries in the table, we abuse notation by denoting the local latent variable for x_i as $\boldsymbol{\theta}$. Further, $\boldsymbol{\zeta}$ denote the generative parameters.

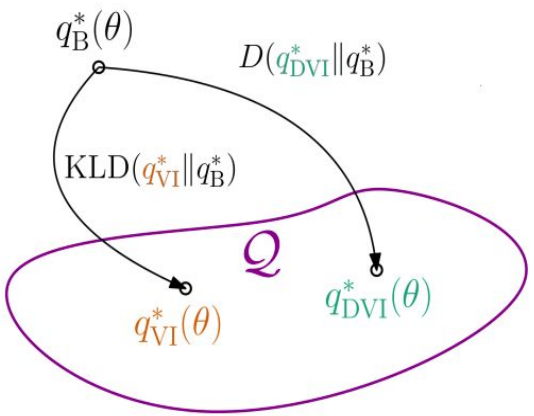
Different views of Variational Inference

1. ELBO view

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

2. Discrepancy-minimisation view (DVI)

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



Different views of Variational Inference

3. Constrained optimisation view (GVI):

The third view of VI is its generalisation approach and which was fist presented at the paper of Knoblauch, Damoulas et al. and treats the VI solution as the best Q- constrained solution. We recall:

$$q^*_{\text{VI}}(\theta) = \underset{q \in \mathcal{Q}}{\text{argmax}} ELBO(q) = \underset{q \in \mathcal{Q}}{\text{argmin}} - ELBO(q) = \underset{q \in P(\Theta)}{\text{argmin}} \left\{ \mathbb{E}_{q(\theta)} \left[-\sum_{i=1}^{n} \log p(x_i|\theta) \right] + KL(q||\pi) \right\}$$

Hence, VI solves a problem specified by the Rule of Three $P(l, D, \Pi)$ for $l = \log p(x_i|\theta)$, $D = KL(q||\pi)$ and $\Pi = \mathcal{Q}$.

$$\implies q^*_{\text{VI}}(\theta) = P(\log p(x_i|\theta), KL(q||\pi), \mathcal{Q})$$

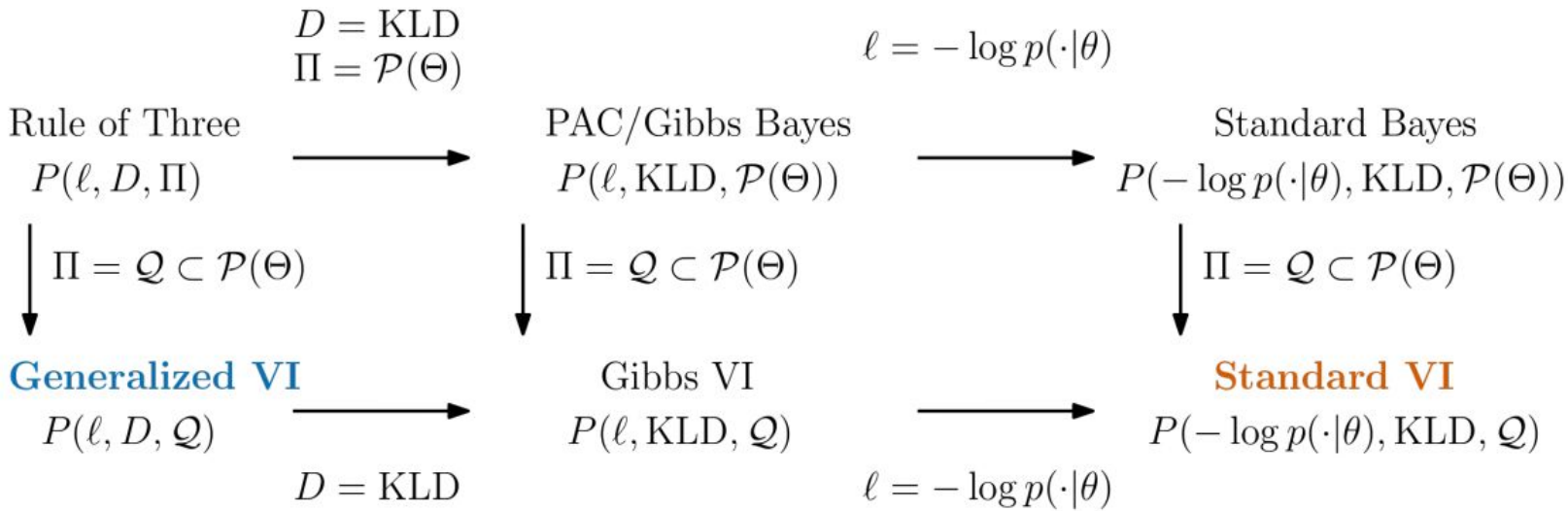
Generalised Variational Inference

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).**

The main areas that GVI has proved to be really effIcient are the following:

1. Robust alternatives to $l(\theta, x_i) = -\log p(x_i|\theta)$.
2. Prior-robust uncertainty quantification via D.
3. Adjusting marginal variances via D.

Generalised Variational Inference



4. Experiments and Analysis

Notation:

Divergence	Notation
KL Divergence	KL
Reverse KL Divergence	RKL
α-Divergence	A
α-Renyi Divergence	AR
param. α-Renyi Divergence	αAR
Jensen–Shannon Divergence	JS
Total Variation Distance	TV
Fisher Distance	F

I. Regression on UCI data sets - Practicalities

Idea:

1. To compare different divergence measures and different neural network depths on the GVI setting.
2. Grid-search for the hyperparameter dependent divergences like α-Divergence and α-Renyi divergence.

How? We just need to focus on trying different discrepancy measures while keeping the loss function fixed and compare their RMSE and NLL values across different data sets but also across different number of hidden layers.

Data: Four different data sets from the UCI Machine Learning repository (Boston, Concrete, Energy and Yacht)

I. Regression on UCI data sets - Practicalities

Our model:

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.

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).

I. Regression on UCI data sets - Practicalities

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 $\boldsymbol{\mu(x)}$ and the log-standard deviation $\boldsymbol{\log\sigma(x)}$.

Also, that helped us of having all of our divergences in a closed Gaussian form. For example:

For a prior $\pi \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and approximate posterior $q \sim \mathcal{N}(\mu_2, \sigma_2^2)$

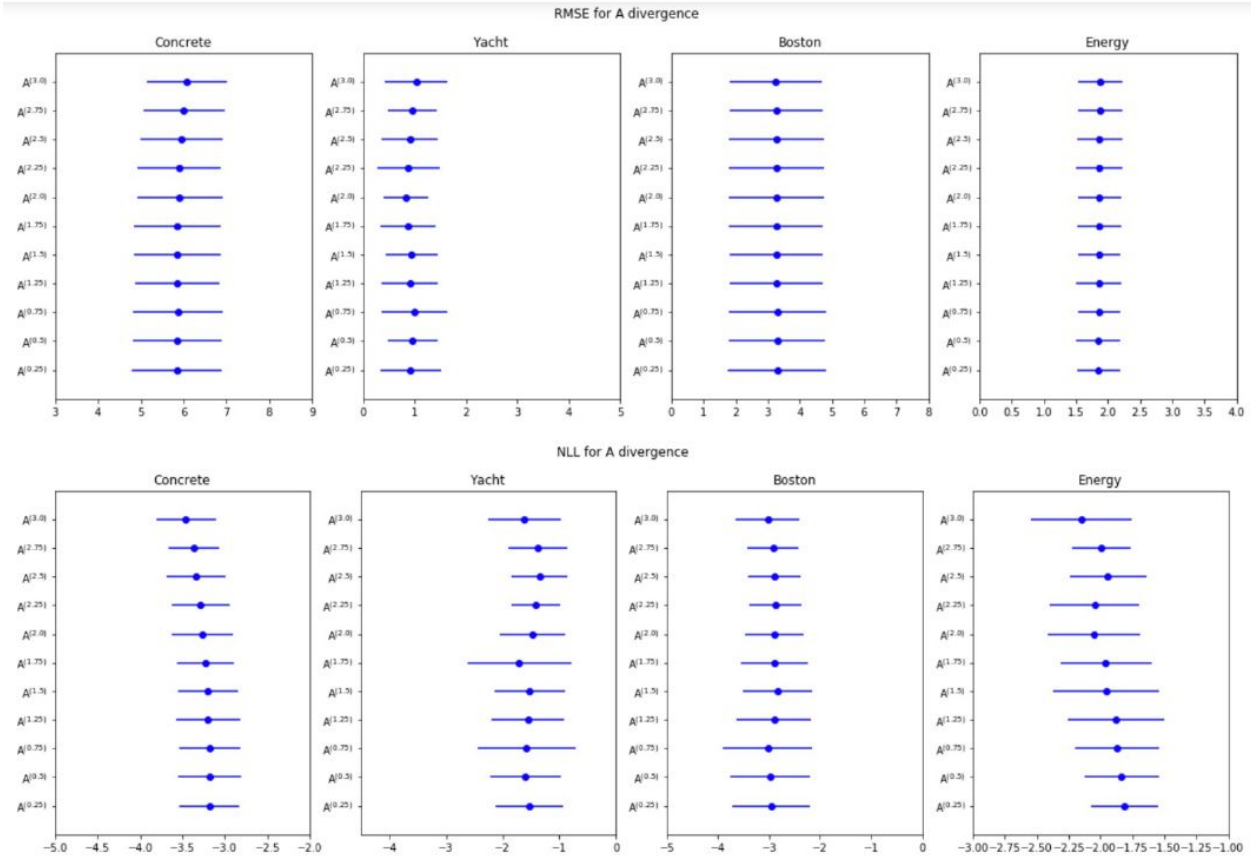
Kullback-Leibler Divergence:

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

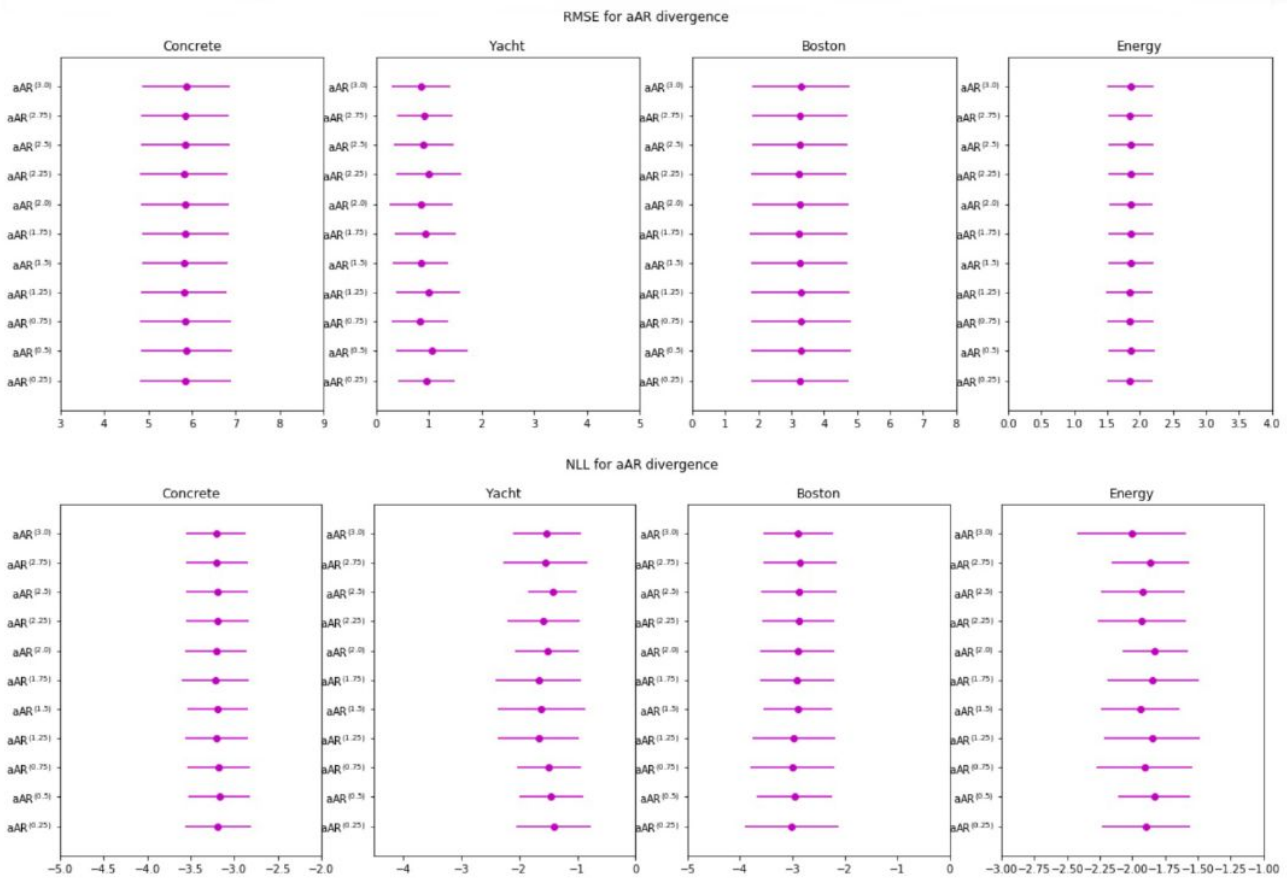
α-Divergence:

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

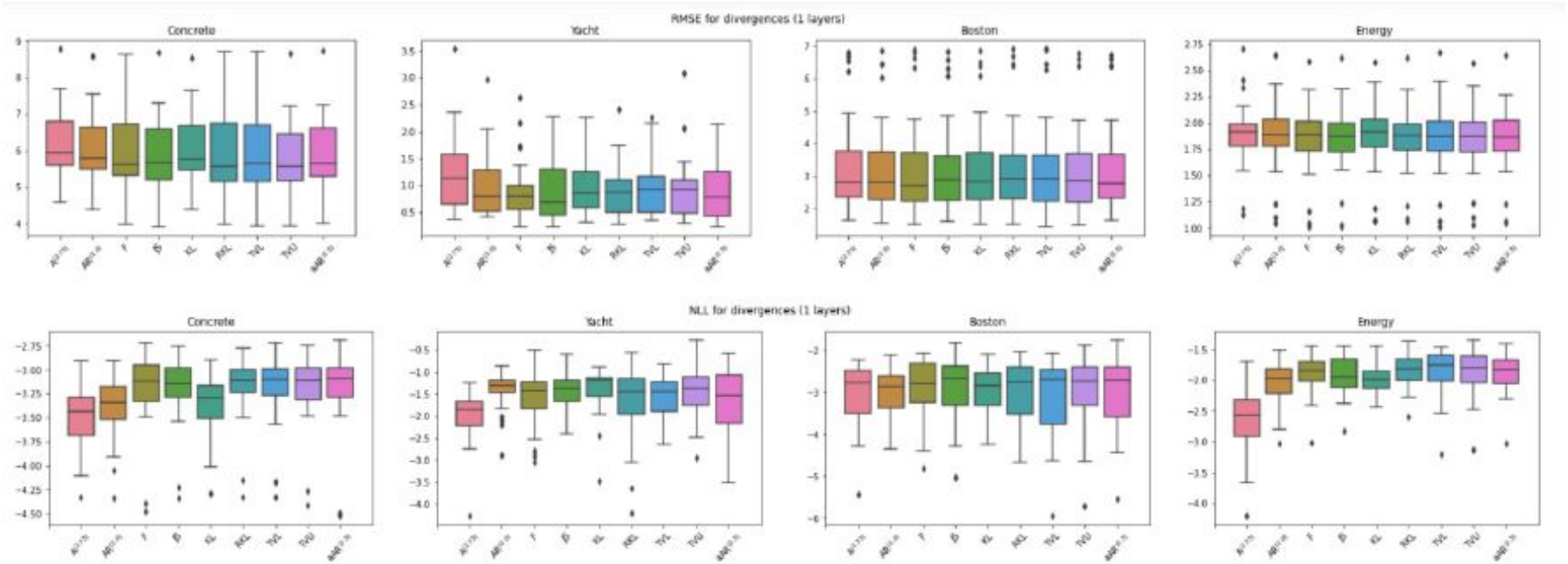
I.i The α grid-search



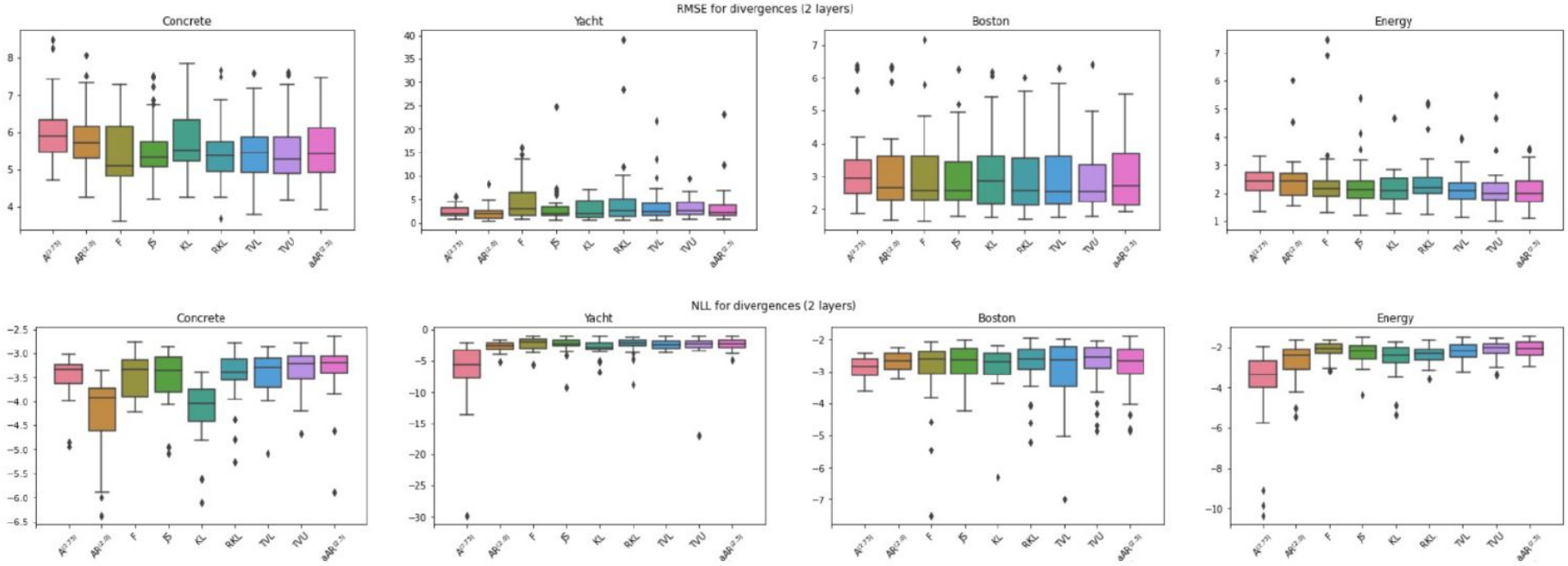
I.i The α grid-search



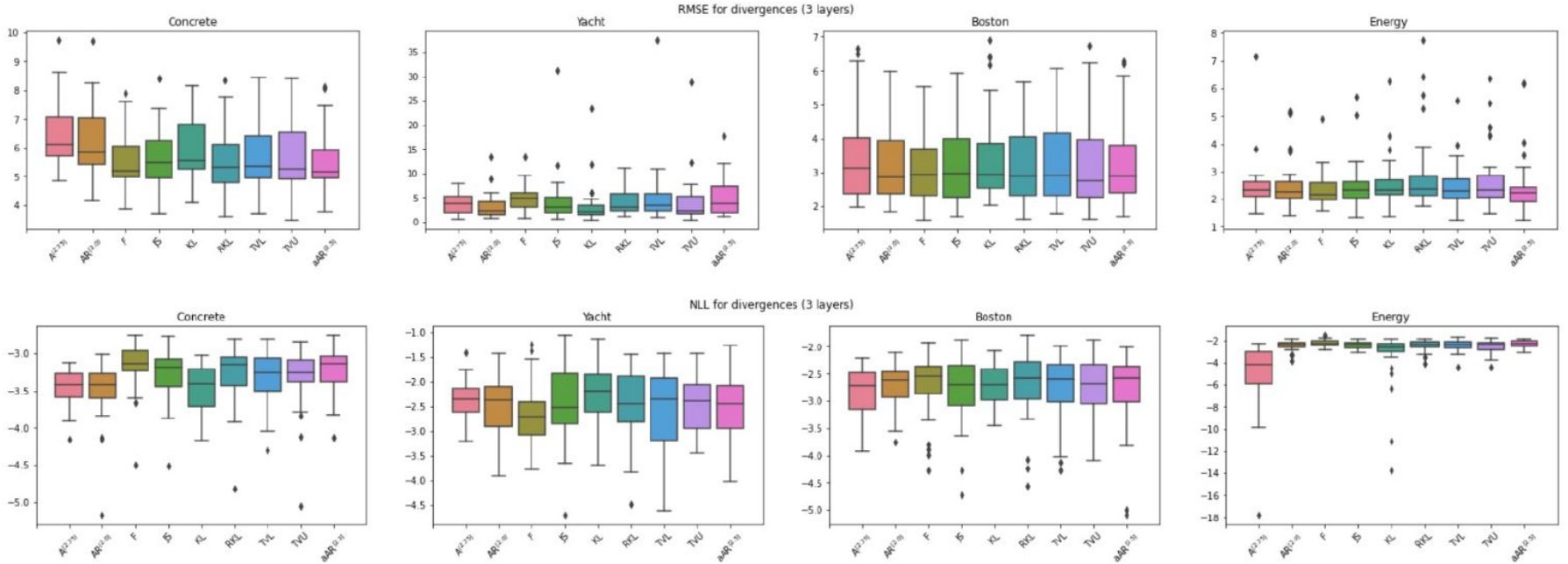
I.ii GVI Vs VI


RMSE for divergences (1 layers)
Concrete
Yacht
Boston
Energy
NLL for divergences (1 layers)
Concrete
Yacht
Boston
Energy

I.ii GVI Vs VI


RMSE for divergences (2 layers)
Concrete
Yacht
Boston
Energy
NLL for divergences (2 layers)
Concrete
Yacht
Boston
Energy
F
JS
KL
RKL
TVL
TVU

I.ii GVI Vs VI


RMSE for divergences (3 layers)
Concrete
Yacht
Boston
Energy
NLL for divergences (3 layers)
Concrete
Yacht
Boston
Energy

I.iii GVI Vs Approximate Inference methods

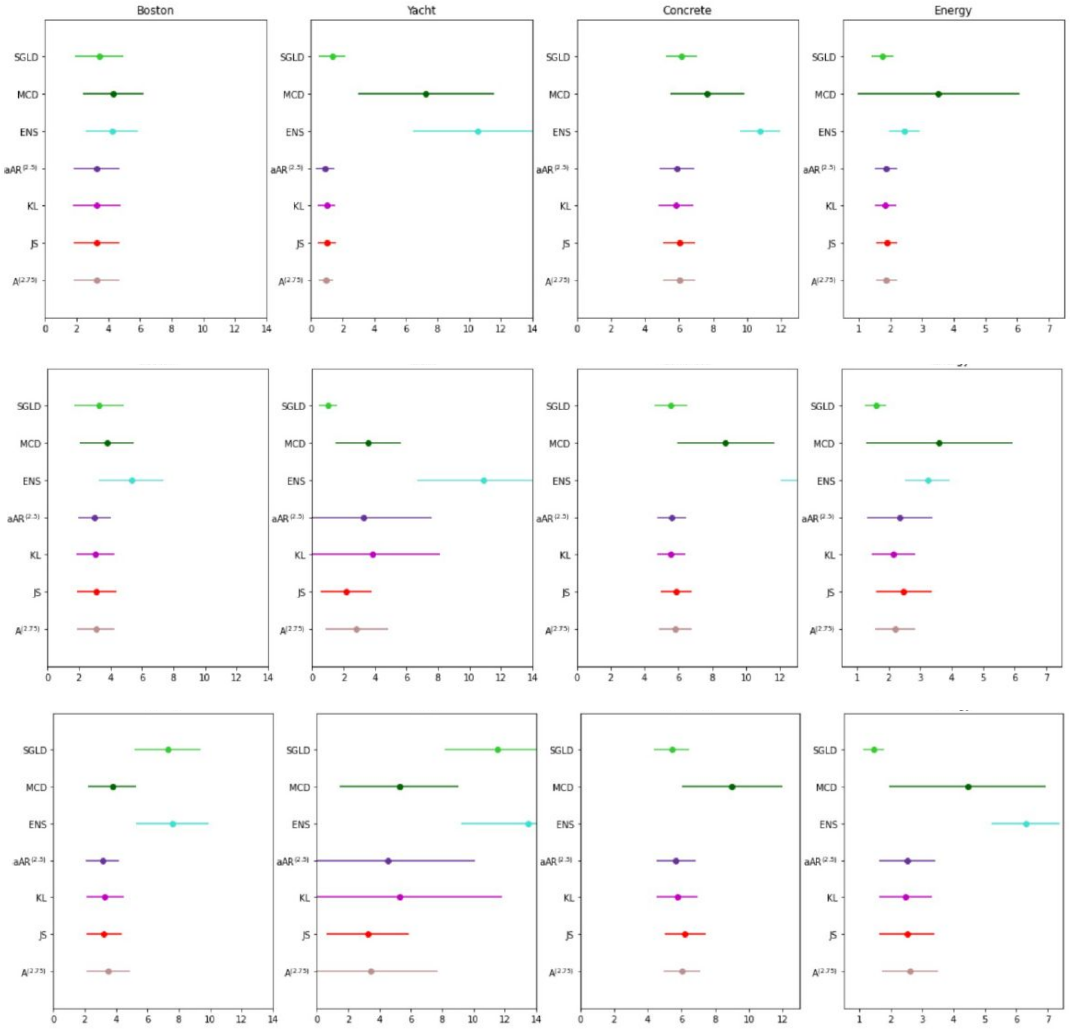
We are going to make a straight comparison of some of the "best" performing

divergences of the GVI setting with:

(a) The KL divergence (Standard Variational Inference) and

(b) Three state-of-the-art approximate inference methods

- Monte-Carlo Dropout
- Deep Ensemble
- Stochastic Gradient Langevin Dynamics



1 hidden layer

2 hidden layers

3 hidden layers

II. Regression on Gaussian Process ground truth

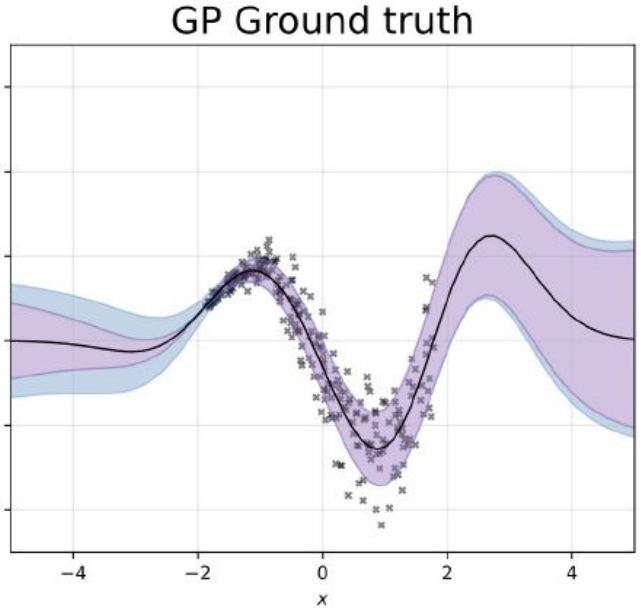
Idea: We want to evaluate aleatoric and epistemic uncertainty for Bayes by backprop algorithm while performing GVI across different discrepancy measures and neural network depths.

Data: The heteroscedastic data was generated by a Gaussian Process with an RBF kernel

$$(l = 1, \sigma_n = 0.3|x + 2|)$$

Model: 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.

II. Regression on Gaussian Process ground truth



Uncertainty Decomposition

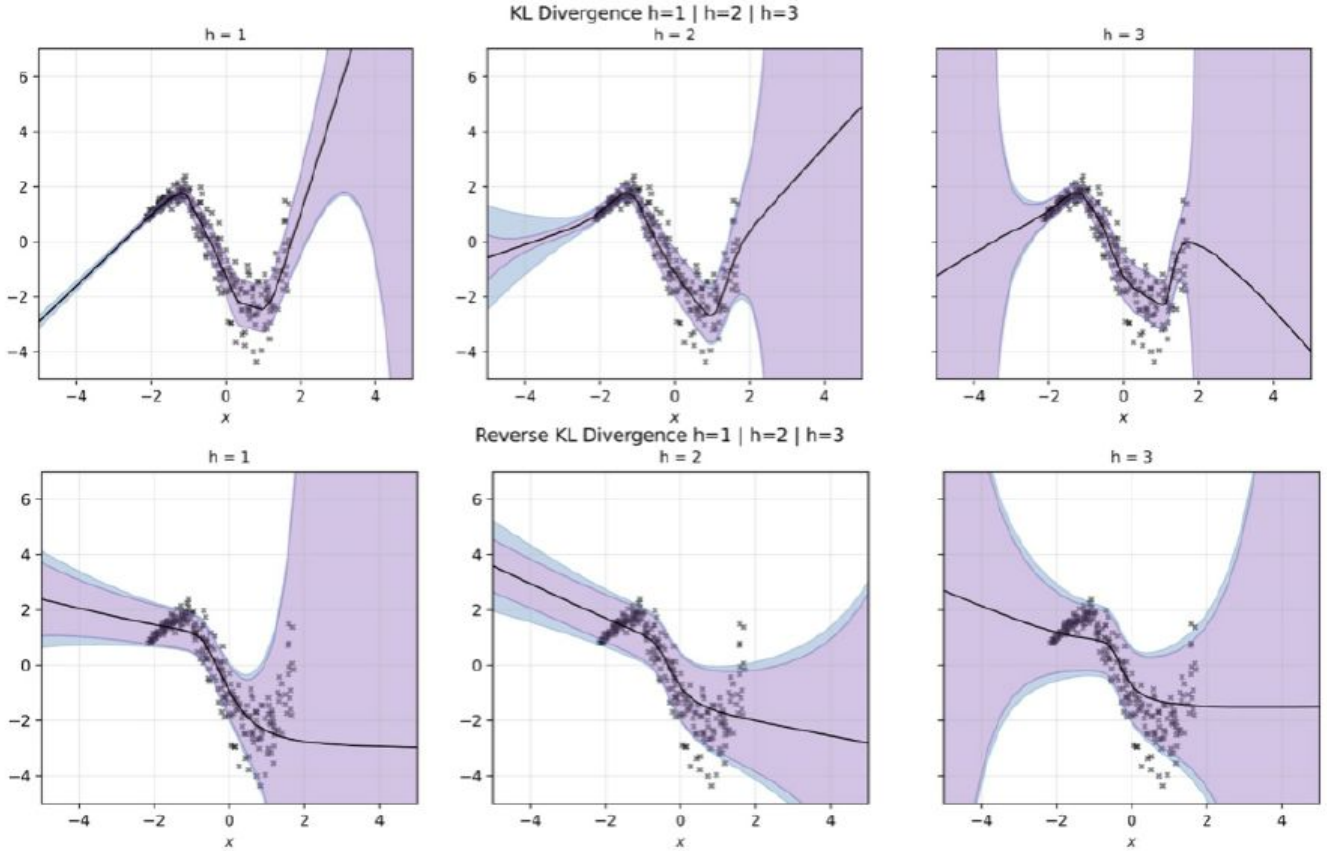
Aleatoric uncertainty (noise uncertainty) can be quantified as:

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

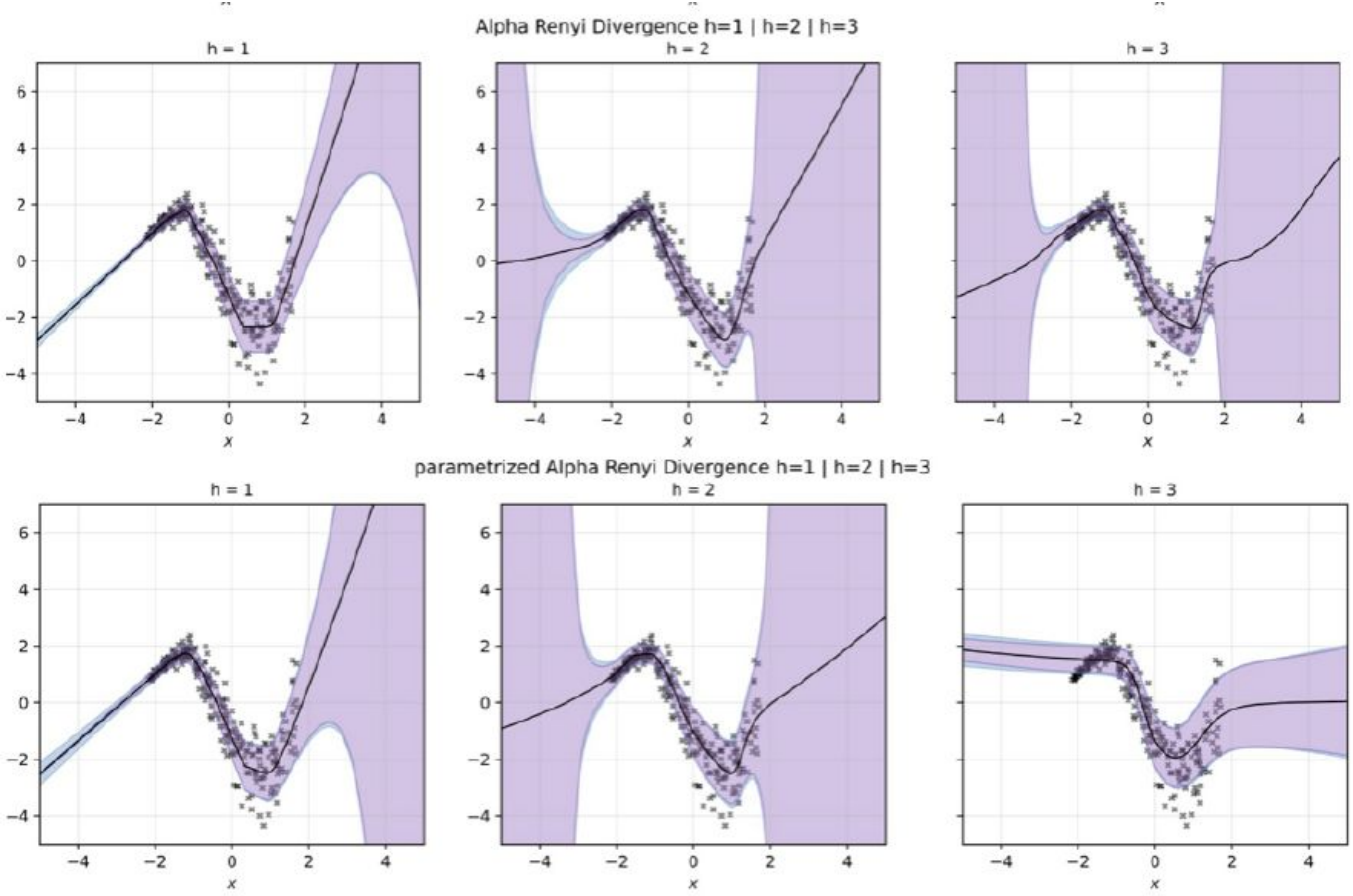
Epistemic uncertainty (model uncertainty) can be quantified as:

$$\mathcal{U}_E = Var_{q(w)}(\mu_{\text{pred}}) \quad \text{or} \quad \mathcal{U}_E = \mathcal{U}(y'|x') - \mathcal{U}_A$$

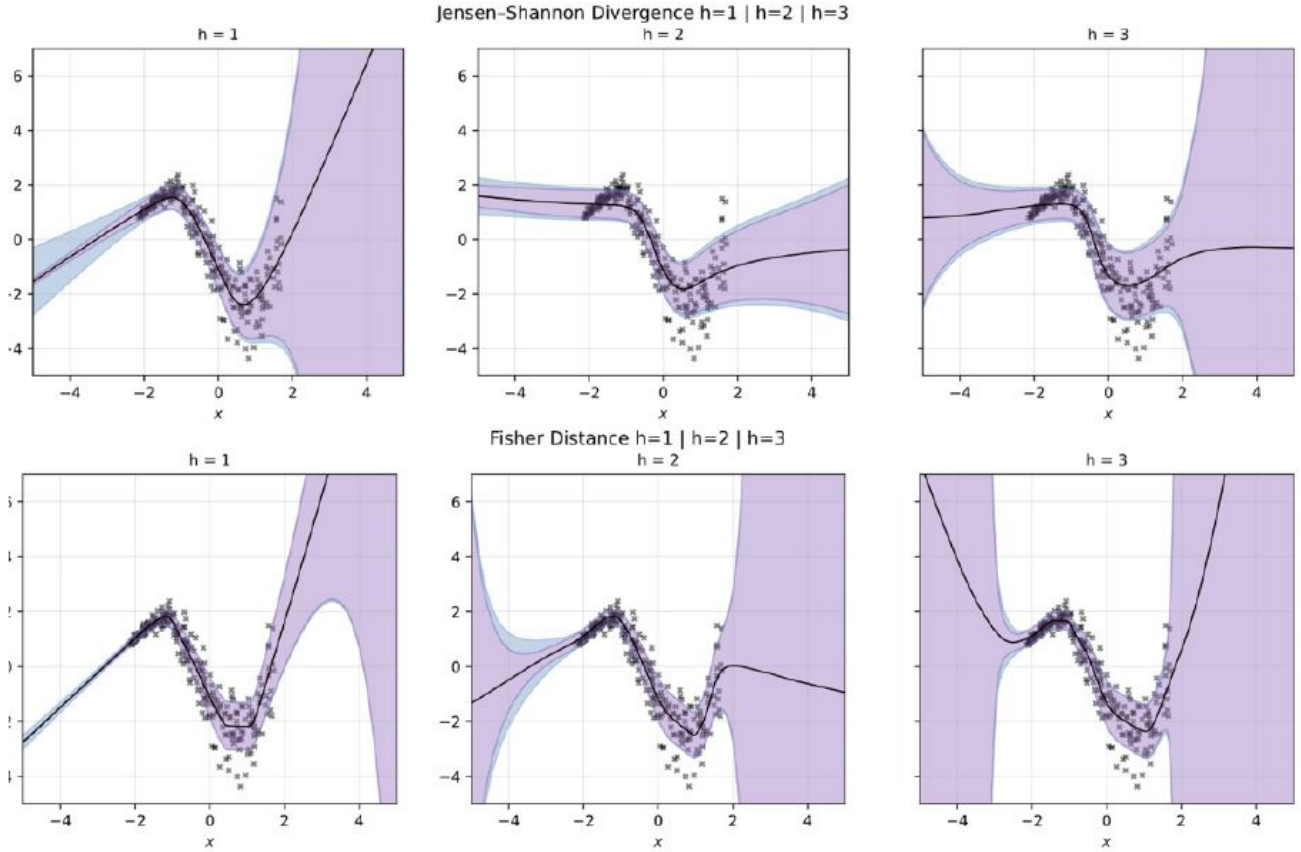
Results


KL Divergence h=1 | h=2 | h=3
h = 1
h = 2
h = 3
Reverse KL Divergence h=1 | h=2 | h=3
h = 1
h = 2
h = 3
x

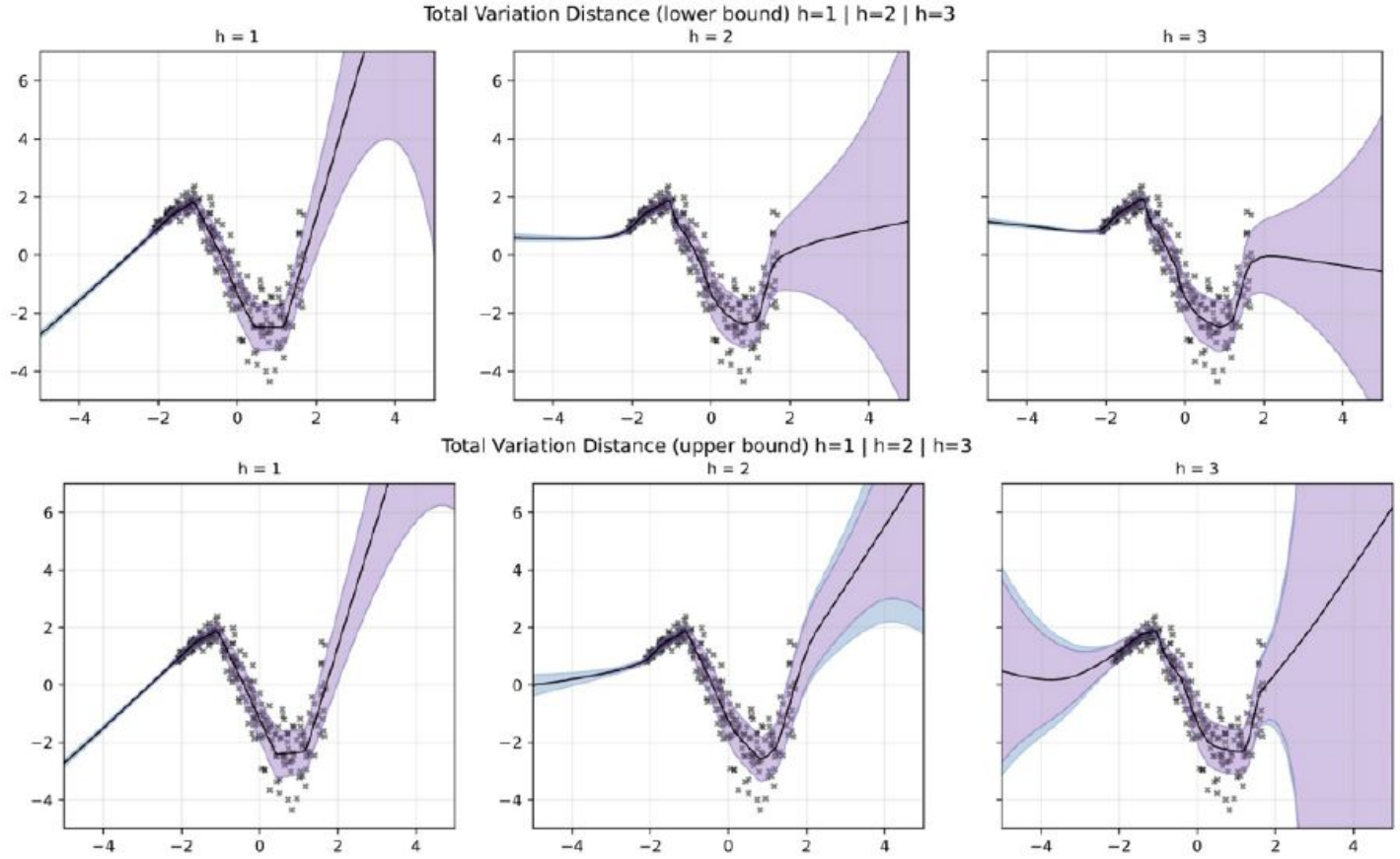
Results


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

Results


Jensen-Shannon Divergence h=1 | h=2 | h=3
h = 1
h = 2
h = 3
Fisher Distance h=1 | h=2 | h=3
h = 1
h = 2
h = 3
x

Results


Total Variation Distance (lower bound) h=1 | h=2 | h=3
h = 1
h = 2
h = 3
Total Variation Distance (upper bound) h=1 | h=2 | h=3
h = 1
h = 2
h = 3

Visual Inspection

Best models

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

Log-likelihood comparison

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**

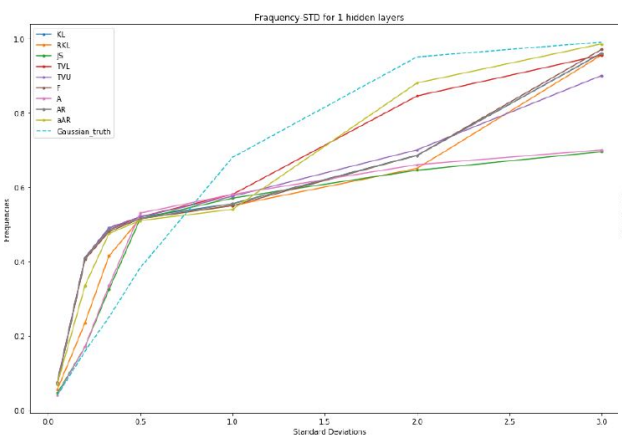
Model Calibration

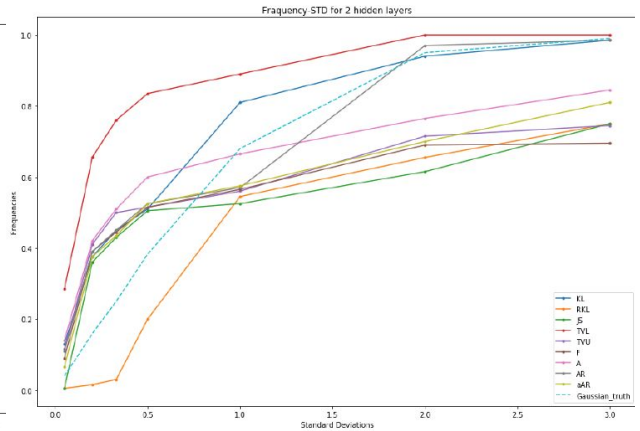
- In order to select the best calibrated model, we inspect their confidence at different intervals for all the divergence measures and neural network depths.
- We report the predicted probability and the difference with respect to the true probability at one standard deviation interval, which should be 0.68 under the Gaussian assumption, for all the different divergence measures and neural network depths.
- We define a well-calibrated model as the one whose predicted probability differs no more than 0.1 from the true probability.

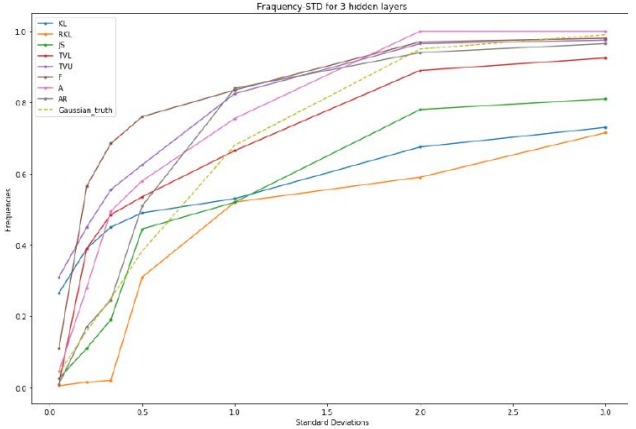
Divergence	No. of hidden layers	Predicted probability	Difference
KL	1	0.55	0.13
KL	2	0.81	0.13
KL	3	0.53	0.15
RKL	1	0.56	0.12
RKL	2	0.54	0.14
RKL	3	0.52	0.16
A	1	**0.58**	0.10
A	2	**0.66**	0.02
A	3	**0.75**	0.07
AR	1	0.55	0.13
AR	2	0.57	0.11
AR	3	0.84	0.16
αAR	1	0.56	0.12
αAR	2	**0.60**	0.10
αAR	3	**0.68**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

Model Calibration







We can make the following remarks:

1. Confidence increases together with the number of layers, except for the case of RKL, where the reverse can be observed.
2. There is a general overconfidence trend on lower sigma intervals i.e. (0; 0:7).

Model Calibration

Best models

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

5. Conclusion

Contributions

Through our experiments we made some really useful contributions:

- We have conducted optimisation over a vast divergence landscape on the GVI setting by trying different divergence functions and finding optimal hyperparameters for existing ones (A, AR, AR), extending the work of the original GVI paper.
- We have provided visualisations of uncertainty quantification and evaluation of epistemic and aleatoric uncertainties for a Gaussian Process ground truth.
- We have made an extensive discussion (visual inspection, information criteria, model calibration) about model selection for GVI posteriors.
- Overall, we have provided an empirical proof of the superiority of GVI over traditional approximate inference methods and mainly over standard VI, in certain cases in the framework of Bayesian Neural Networks.

Future Work

- Continue this exhaustive search in this direction by trying different model architectures, train for more epochs or try different splits, or sample more data points from the variational posterior.
- As far as the second experiment is concerned, as we mentioned before, it is interesting to see the behaviour of the different divergences for alternative kernels and especially for non stationary ones.
- As an extension of present work, it would be worthwhile exploring the performance for more robust divergence measures and loss functions, especially from the ones that have been proved to be robust in approximating posterior distributions. Specifically, new discrepancy measures that are worth investigating are: Wasserstein distance, β/γ-divergences, Kagan's divergence etc.

Future work

- Try a generalisation of Jensen-Shannon divergence: the α-Jensen-Shannon divergence:

$$JS_\alpha(p||q) = \frac{1}{2}(K_\alpha(p||q) + K_\alpha(q||p))$$

  where $K_\alpha(p||q) = KL(p\|(1-\alpha)p + \alpha q)$ and we can retrieve Jensen-Shannon divergence for $\boldsymbol{\alpha = 1/2}$ and Jeffrey's divergence for $\boldsymbol{\alpha = 1}$.

- Try different inference algorithm (e.g Probabilistic Backpropagation)

Thank you!

Questions?

Contact:

gio.felekis@gmail.com

https://www.linkedin.com/in/giorgos-felekis/

https://github.com/gfelekis *

*Thesis can be found here