Combining model and parameter uncertainty in BNNs

Hubin A.A., Storvik G.O.

SAMBA, Norwegian Computing Center

aliaksandr.hubin@nr.no

Department of Mathematics, University of Oslo

geirs@math.uio.no

Nordic Probabilistic AI School,
Trondheim, Norway

06.06.2019

Sources of uncertainty in Bayesian modelling

Consider a regression with input x, output y and measurement noise ε:

$$y = g(x;\theta) + \varepsilon, \quad g \in \mathcal{G}, \theta \in \Theta_g, \varepsilon \sim \mathfrak{f};$$

The four sources of uncertainty are here present.

Sources of uncertainty in Bayesian modelling

Consider a regression with input x, output y and measurement noise ε:

$$y = g(x;\theta) + \varepsilon, \quad g \in \mathcal{G}, \theta \in \Theta_g, \varepsilon \sim \mathfrak{f};$$

The four sources of uncertainty are here present.

1. In ε responsible for measurement error [Hubin et al., 2018, addressed in DBRM], where mixtures of latent Gaussian variables are selected;

Sources of uncertainty in Bayesian modelling

Consider a regression with input x, output y and measurement noise ε:

$$y = g(x;\theta) + \varepsilon, \quad g \in \mathcal{G}, \theta \in \Theta_g, \varepsilon \sim \mathfrak{f};$$

The four sources of uncertainty are here present.

1. In ε responsible for measurement error [Hubin et al., 2018, addressed in DBRM], where mixtures of latent Gaussian variables are selected;

2. In specification of g conditional on ε responsible for interpretability [Hubin et al., 2018, addressed in DBRM], where physical models in the closed form are recovered from raw data with large power;

Sources of uncertainty in Bayesian modelling

Consider a regression with input x, output y and measurement noise ε:

$$y = g(x;\theta) + \varepsilon, \quad g \in \mathcal{G}, \theta \in \Theta_g, \varepsilon \sim \mathfrak{f};$$

The four sources of uncertainty are here present.

1. In ε responsible for measurement error [Hubin et al., 2018, addressed in DBRM], where mixtures of latent Gaussian variables are selected;
2. In specification of g conditional on ε responsible for interpretability [Hubin et al., 2018, addressed in DBRM], where physical models in the closed form are recovered from raw data with large power;
3. In selecting subsets of θ conditional on g and ε responsible for sparsity (common, addressed in this talk);

Sources of uncertainty in Bayesian modelling

Consider a regression with input x, output y and measurement noise ε:

$$y = g(x;\theta) + \varepsilon, \quad g \in \mathcal{G}, \theta \in \Theta_g, \varepsilon \sim \mathfrak{f};$$

The four sources of uncertainty are here present.

1. In ε responsible for measurement error [Hubin et al., 2018, addressed in DBRM], where mixtures of latent Gaussian variables are selected;
2. In specification of g conditional on ε responsible for interpretability [Hubin et al., 2018, addressed in DBRM], where physical models in the closed form are recovered from raw data with large power;
3. In selecting subsets of θ conditional on g and ε responsible for sparsity (common, addressed in this talk);
4. In the values of θ conditional on the rest (most common, addressed so far at ProbAI).

The first three items here are addressed to as **model uncertainty** and only the last one is the **parameter uncertainty**.

Introduction. Issues with existing ANNs

- Neural networks (NN) allow to model flexible parametric distributions;
- At the same time, frequentist neural networks are exposed to overfit;
- Bayesian neural networks (BNNs) are robust to overfitting (like DBRM [Hubin et al., 2018]);
- Scalable methods for BNNs exist (unlike DBRM);
- BNNs are still heavily over-parameterized (unlike DBRM);
- No scalable methods for modeling model uncertainty and performing Bayesian model selection and averaging formally (like in DBRM) exist (however there are some ad-hoc based approaches).

The model

$$\boldsymbol{y}_i \sim \mathfrak{f}\left(\boldsymbol{\mu}_i, \phi\right), \quad i \in \{1, ..., n\} \tag{1}$$

$$\boldsymbol{\mu}_i = \{z_{i1}^{(L)}, ..., z_{ir}^{(L)}\}, \tag{2}$$

$$z_{ij}^{(l+1)} = \sigma_j^{(l)} \left(\gamma_{0j}^{(l)} \beta_{0j}^{(l)} + \sum_{k=1}^{p^{(l)}} \gamma_{kj}^{(l)} \beta_{kj}^{(l)} z_{ik}^{(l)} \right), \tag{3}$$

- $f(\cdot|\boldsymbol{\mu}, \phi)$ is a density/distribution with expectation μ and dispersion parameter ϕ;
- $\beta_{kj}^{(l)} \in \mathcal{R}$ are the weights (slope coefficients) for the inputs $\boldsymbol{z}_{ik}^{(l)}$; $\gamma_{kj}^{(l)} \in \{0, 1\}$ are latent binary indicators switching the corresponding weights on and off;
- $p^{(l)}$ is the number of neurons at layer l;
- L is the total number of layers.

Inference on the model

Let:

- $\boldsymbol{\gamma} = \cup_{l,j,k} \gamma_{kj}^{(l)}$ define a model itself, i.e. which weights are switched on and which are switched off;
- $\boldsymbol{\theta}|\boldsymbol{\gamma} = \{\boldsymbol{\beta}, \phi|\boldsymbol{\gamma}\}$, where $\boldsymbol{\beta}|\boldsymbol{\gamma} = \cup_{l,j,k:\gamma_{kj}^{(l)}=1} \beta_{kj}^{(l)}$, define parameters of $\boldsymbol{\gamma}$.

Inference on the model

Let:

- $\boldsymbol{\gamma} = \cup_{l,j,k} \gamma_{kj}^{(l)}$ define a model itself, i.e. which weights are switched on and which are switched off;
- $\boldsymbol{\theta}|\boldsymbol{\gamma} = \{\boldsymbol{\beta}, \phi|\boldsymbol{\gamma}\}$, where $\boldsymbol{\beta}|\boldsymbol{\gamma} = \cup_{l,j,k:\gamma_{kj}^{(l)}=1} \beta_{kj}^{(l)}$, define parameters of $\boldsymbol{\gamma}$.

Goals:

- $p(\boldsymbol{\gamma}, \boldsymbol{\theta}|\mathbb{D})$ posterior distribution of parameters and models;
- $p(\boldsymbol{\gamma}|\mathbb{D})$ marginal posterior probabilities of the models;
- $p(\Delta|\mathbb{D})$ marginal posterior probabilities of the parameter of interest Δ.

Inference on the model

Let:

- $\boldsymbol{\gamma} = \cup_{l,j,k} \gamma_{kj}^{(l)}$ define a model itself, i.e. which weights are switched on and which are switched off;
- $\boldsymbol{\theta}|\boldsymbol{\gamma} = \{\boldsymbol{\beta}, \phi|\boldsymbol{\gamma}\}$, where $\boldsymbol{\beta}|\boldsymbol{\gamma} = \cup_{l,j,k:\gamma_{kj}^{(l)}=1} \beta_{kj}^{(l)}$, define parameters of $\boldsymbol{\gamma}$.

Goals:

- $p(\boldsymbol{\gamma}, \boldsymbol{\theta}|\mathbb{D})$ posterior distribution of parameters and models;
- $p(\boldsymbol{\gamma}|\mathbb{D})$ marginal posterior probabilities of the models;
- $p(\Delta|\mathbb{D})$ marginal posterior probabilities of the parameter of interest Δ.

But:

- $\exists 2^q$ different models in Ω_γ;
- q is the number of weights in the BNN, which is huge;
- Ω_γ is not feasible to even specify.

Model priors

$$p(\gamma) \propto \prod_{l=1}^{L-1} \prod_{j=1}^{p^{(l+1)}} \prod_{k=0}^{p^{(l)}} a^{\gamma_{kj}^{(l)}} \ . \tag{4}$$

- $a \in (0, 1)$ is the penalty for including weight $\beta_{kj}^{(l)}$ into the model;
- $a = \exp(-2)$ corresponds to AIC penalization of the weights;
- $a = \exp(-2 \log n)$, where n is the full training sample size, - to BIC;
- Of course, more advanced model priors like Dirichlet or dilusion priors can be considered.

Parameter priors

$$p(\beta_{kj}^{(l)}|\gamma_{kj}^{(l)} = 1) = N(0, \sigma_\beta^2), \tag{5}$$

$$p(\phi|\boldsymbol{\gamma}) = \phi^{-1}. \tag{6}$$

- σ_β^2 is the prior variance of the weights;
- Of course, more advanced weight priors like mixtures of g-priors, horseshoe prior or Jeffrey's prior can be alternatively considered;
- In many distributions from the exponential family ϕ is known and fixed.

Inference possibilities

- Markov chain Monte Carlo (exact inference);
- Laplace approximations;
- *Integrated nested Laplace approximations* ([Rue et al., 2009, became famous here at NTNU]);
- **Variational inference** (addressed at ProbAI);
- Approximate Bayesian computation.

Doubly stochastic variational inference

Posterior joint distribution $p(\boldsymbol{\theta}, \boldsymbol{\gamma}|\mathbb{D})$ is approximated by combining:

- Scalable variational inference for BNN proposed by [Graves, 2011]:

$$\mathrm{KL}(q_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \boldsymbol{\gamma})||p(\boldsymbol{\theta}, \boldsymbol{\gamma}|\mathbb{D})) = \sum_{\boldsymbol{\gamma} \in \Gamma} \int_{\Theta} q_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \boldsymbol{\gamma}) \log \frac{q_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{p(\boldsymbol{\theta}, \boldsymbol{\gamma}|\mathbb{D})} \mathrm{d}\boldsymbol{\theta} \to \min_{\eta}; \quad (7)$$

Doubly stochastic variational inference

Posterior joint distribution $p(\boldsymbol{\theta}, \boldsymbol{\gamma}|\mathbb{D})$ is approximated by combining:

- Scalable variational inference for BNN proposed by [Graves, 2011]:

$$\mathrm{KL}(q_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \boldsymbol{\gamma})||p(\boldsymbol{\theta}, \boldsymbol{\gamma}|\mathbb{D})) = \sum_{\boldsymbol{\gamma} \in \Gamma} \int_{\Theta} q_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \boldsymbol{\gamma}) \log \frac{q_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{p(\boldsymbol{\theta}, \boldsymbol{\gamma}|\mathbb{D})} \mathrm{d}\boldsymbol{\theta} \to \min_{\eta}; \quad (7)$$

- Variational distributions for the joint parameter-model settings for linear models introduced by [Carbonetto et al., 2012]:

$$q_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \boldsymbol{\gamma}) = q_{\boldsymbol{\eta}_0}(\phi) \prod_{l=1}^{L-1} \prod_{j=1}^{p^{(l+1)}} \prod_{k=0}^{p^{(l)}} q_{\boldsymbol{\eta}_{kj}^{(l)}}(\beta_{kj}^{(l)}, \gamma_{kj}^{(l)}), \quad (8)$$

$$q_{\boldsymbol{\eta}_{kj}^{(l)}}\left(\beta_{kj}^{(l)}, \gamma_{kj}^{(l)}\right) = \begin{cases} \alpha_{kj}^{(l)} N\left(\mu_{kj}^{(l)}, \sigma^{2}{}_{kj}^{(l)}\right), & \text{if } \gamma_{kj}^{(l)} = 1, \\ \left(1 - \alpha_{kj}^{(l)}\right) \delta_0(\beta_{kj}^{(l)}), & \text{if } \gamma_{kj}^{(l)} = 0. \end{cases} \quad (9)$$

Evidence lower bound

Proposition

Minimization of $KL\left(q_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \boldsymbol{\gamma}) || p(\boldsymbol{\theta}, \boldsymbol{\gamma} | \mathbb{D})\right)$ *and maximization of the evidence (log marginal likelihood) lower bound (ELBO) are equivalent.*

$$\mathcal{L}_{VI}(\boldsymbol{\eta}) := \sum_{\boldsymbol{\gamma} \in \Gamma} \int_{\Theta} q_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \boldsymbol{\gamma}) \log p(\mathbb{D} | \boldsymbol{\theta}, \boldsymbol{\gamma}) d\boldsymbol{\theta} - KL\left(q_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \boldsymbol{\gamma}) || p(\boldsymbol{\theta}, \boldsymbol{\gamma})\right)$$

Evidence lower bound

Proposition

Minimization of $KL\left(q_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \gamma) || p(\boldsymbol{\theta}, \gamma | \mathbb{D})\right)$ *and maximization of the evidence (log marginal likelihood) lower bound (ELBO) are equivalent.*

$$\mathcal{L}_{VI}(\boldsymbol{\eta}) := \sum_{\gamma \in \Gamma} \int_{\Theta} q_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \gamma) \log p(\mathbb{D} | \boldsymbol{\theta}, \gamma) d\boldsymbol{\theta} - KL\left(q_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \gamma) || p(\boldsymbol{\theta}, \gamma)\right)$$

Proof.

$$\mathrm{KL}(q_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \gamma) || p(\boldsymbol{\theta}, \gamma | \mathbb{D})) = \sum_{\gamma \in \Gamma} \int_{\Theta} q_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \gamma) \log \frac{q_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \gamma) p(\mathbb{D})}{p(\mathbb{D} | \boldsymbol{\theta}, \gamma) p(\boldsymbol{\theta}, \gamma)} \mathrm{d}\boldsymbol{\theta}$$

$$= \log p(\mathbb{D}) + \sum_{\gamma \in \Gamma} \int_{\Theta} q_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \gamma) \log \frac{q_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \gamma)}{p(\boldsymbol{\theta}, \gamma)} \mathrm{d}\boldsymbol{\theta} - \sum_{\gamma \in \Gamma} \int_{\Theta} q_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \gamma) \log p(\mathbb{D} | \boldsymbol{\theta}, \gamma) \mathrm{d}\boldsymbol{\theta}$$

$$= \log p(\mathbb{D}) - \mathcal{L}_{VI}(\boldsymbol{\eta}).$$

from which the result follows. □

ELBO estimation with sub-sampling

1. Assuming conditional independence of the observations:

$$\sum_{\gamma \in \Gamma} \int_{\Theta} q_{\eta}(\theta, \gamma) \log p(\mathbb{D}|\theta, \gamma) \mathrm{d}\theta = \sum_{i=1}^{n} \sum_{\gamma \in \Gamma} \int_{\Theta} q_{\eta}(\theta, \gamma) \log p(\boldsymbol{y_i}|\boldsymbol{x_i}, \theta, \gamma) \mathrm{d}\theta.$$

ELBO estimation with sub-sampling

1. Assuming conditional independence of the observations:

$$\sum_{\gamma \in \Gamma} \int_{\Theta} q_{\eta}(\boldsymbol{\theta}, \gamma) \log p(\mathbb{D}|\boldsymbol{\theta}, \gamma) \mathrm{d}\boldsymbol{\theta} = \sum_{i=1}^{n} \sum_{\gamma \in \Gamma} \int_{\Theta} q_{\eta}(\boldsymbol{\theta}, \gamma) \log p(\boldsymbol{y_i}|\boldsymbol{x_i}, \boldsymbol{\theta}, \gamma) \mathrm{d}\boldsymbol{\theta}.$$

2. Now sample mini-batches S of size N from the full data, yielding:

$$\widehat{\mathcal{L}}_{VI}(\boldsymbol{\eta}) = \frac{n}{N} \sum_{i \in S} \sum_{\gamma \in \Gamma} \int_{\Theta} q_{\eta}(\boldsymbol{\theta}, \gamma) \log p(\boldsymbol{y_i}|\boldsymbol{x_i}, \boldsymbol{\theta}, \gamma) \mathrm{d}\boldsymbol{\theta} -$$

$$\mathrm{KL}\left(q_{\eta}(\boldsymbol{\theta}, \gamma) || p(\boldsymbol{\theta}|\gamma) p(\gamma)\right).$$

ELBO estimation with sub-sampling

1. Assuming conditional independence of the observations:

$$\sum_{\gamma\in\Gamma}\int_{\Theta} q_{\eta}(\boldsymbol{\theta},\gamma)\log p(\mathbb{D}|\boldsymbol{\theta},\gamma)\mathrm{d}\boldsymbol{\theta} = \sum_{i=1}^{n}\sum_{\gamma\in\Gamma}\int_{\Theta} q_{\eta}(\boldsymbol{\theta},\gamma)\log p(\boldsymbol{y_i}|\boldsymbol{x_i},\boldsymbol{\theta},\gamma)\mathrm{d}\boldsymbol{\theta}.$$

2. Now sample mini-batches S of size N from the full data, yielding:

$$\widehat{\mathcal{L}}_{VI}(\boldsymbol{\eta}) = \frac{n}{N}\sum_{i\in S}\sum_{\gamma\in\Gamma}\int_{\Theta} q_{\eta}(\boldsymbol{\theta},\gamma)\log p(\boldsymbol{y_i}|\boldsymbol{x_i},\boldsymbol{\theta},\gamma)\mathrm{d}\boldsymbol{\theta} -$$

$$\mathrm{KL}\left(q_{\eta}(\boldsymbol{\theta},\gamma)||p(\boldsymbol{\theta}|\gamma)p(\gamma)\right).$$

3. Still infeasible - use another unbiased Monte-Carlo approximation:

$$\widetilde{\mathcal{L}}_{VI}(\boldsymbol{\eta}) = \frac{1}{M}\sum_{m=1}^{M}\frac{n}{N}\sum_{i\in S}\log p(\boldsymbol{y_i}|\boldsymbol{x_i},\boldsymbol{\theta}^{(m)},\boldsymbol{\gamma}^{(m)}) - \frac{1}{M}\sum_{m=1}^{M}\log\frac{q_{\eta}(\boldsymbol{\theta}^{(m)},\boldsymbol{\gamma}^{(m)})}{p(\boldsymbol{\theta}^{(m)}|\boldsymbol{\gamma}^{(m)})p(\boldsymbol{\gamma}^{(m)})}.$$

Unbiased gradient estimator

Proposition

Assume $\left(\boldsymbol{\theta}^{(m)}, \boldsymbol{\gamma}^{(m)}\right) \sim q_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \boldsymbol{\gamma})$ *for* $m \in \{1, ..., M\}$ *and* S *is a random subset of* $\{1, ..., n\}$ *of size* N*. Then an unbiased estimator for the gradient of* $\mathcal{L}_{VI}(\boldsymbol{\eta})$ *is given by:*

$$\widetilde{\nabla}\mathcal{L}_{VI}(\boldsymbol{\eta}) = \frac{1}{M}\sum_{m=1}^{M}\frac{n}{N}\sum_{i\in S}\nabla \log p(\boldsymbol{y_i}|\boldsymbol{x_i}, \boldsymbol{\theta}, \boldsymbol{\gamma}) - \frac{1}{M}\sum_{m=1}^{M}\nabla \log \frac{q_{\boldsymbol{\eta}}(\boldsymbol{\theta}, \boldsymbol{\gamma})}{p(\boldsymbol{\theta}, \boldsymbol{\gamma})}. \tag{10}$$

Remark

Without inducing bias, a term $l_{\boldsymbol{\eta}}(\boldsymbol{\psi}, \boldsymbol{x}) := \sum_{m=1}^{M}\nabla \log q_{\boldsymbol{\eta}}(\boldsymbol{\theta}^{(m)}, \boldsymbol{\gamma}^{(m)})C_{\boldsymbol{\psi}}(\boldsymbol{x})$ *can be added to* $\widetilde{\nabla}\mathcal{L}_{VI}(\boldsymbol{\eta})$*.*

Doubly stochastic variational inference step

sample N indices uniformly from $\{1, ..., n\}$ defining S;
for m **in** $\{1, ..., M\}$ **do**
  for $(k, j, l) \in \mathcal{B}$ **do**
    set $\alpha_{kj}^{(l)} = \frac{1}{1+\exp(-\omega_{kj}^{(l)})}$;
    set $\sigma_{kj}^{(l)} = \log(1 + \exp(\rho_{kj}^{(l)}))$;
    sample $\gamma_{kj}^{(l)} \sim \text{Bernoulli}(\alpha_{kj}^{(l)})$;
    sample $\beta_{kj}^{(l)} \sim N(\mu_{kj}^{(l)}, \sigma^{2}{}_{kj}^{(l)}) \mathbb{I}(\gamma_{kj}^{(l)} = 1)$;
  end for
end for
compute $\widetilde{\nabla}\mathcal{L}_{VI}(\eta)$ by (10);
set $\eta \leftarrow \eta + a\widetilde{\nabla}\mathcal{L}_{VI}(\eta)$;

Model averaging/ aka marginalized inference/ aka ensembling/ aka Rao-Blackwerization on parameter Δ

Algorithm:

for r **in** $1, ..., R$ **do**
  for $(k, j, l) \in \mathcal{B}$ **do**
    sample $\boldsymbol{\gamma}^{(i)}$ as $\gamma_{kj}^{(l)} \sim \text{Bernoulli}(\alpha_{kj}^{(l)})$;
    sample $\beta_{kj}^{(l)} \sim N(\mu_{kj}^{(l)}, \sigma^{2}{}_{kj}^{(l)}) \mathbb{I}(\gamma_{kj}^{(l)} = 1)$;
  end for
  calculate $p^{(r)}(\Delta) = p(\Delta | \boldsymbol{\beta}^{(r)}, \boldsymbol{\gamma}^{(r)}, \mathbb{D})$;
end for
set $\hat{p}(\Delta|\mathbb{D}) = \frac{1}{R} \sum_{r=1}^{R} p^{(r)}(\Delta)$.

Model averaging/ aka marginalized inference/ aka ensembling/ aka Rao-Blackwerization on parameter Δ

Algorithm:

for r **in** $1, ..., R$ **do**
  for $(k, j, l) \in \mathcal{B}$ **do**
    sample $\boldsymbol{\gamma}^{(i)}$ as $\gamma_{kj}^{(l)} \sim \text{Bernoulli}(\alpha_{kj}^{(l)})$;
    sample $\beta_{kj}^{(l)} \sim N(\mu_{kj}^{(l)}, \sigma^{2}{}_{kj}^{(l)}) \mathsf{I}(\gamma_{kj}^{(l)} = 1)$;
  end for
  calculate $p^{(r)}(\Delta) = p(\Delta | \boldsymbol{\beta}^{(r)}, \boldsymbol{\gamma}^{(r)}, \mathbb{D})$;
end for
set $\hat{p}(\Delta|\mathbb{D}) = \frac{1}{R} \sum_{r=1}^{R} p^{(r)}(\Delta)$.

Examples of Δ:

1. Posterior predictive distribution $\Delta = \mathsf{I}(\hat{\boldsymbol{y}} = \boldsymbol{y_{i+1}})$;
2. Credible interval (regression) $\Delta = \mathsf{I}(y_{lb,95\%} \leq \boldsymbol{y_{i+1}} \leq y_{ub,95\%})$;
3. Credible interval (classification) $\Delta = \mathsf{I}(\max_{j \in J}\{p(y_{i+1,j} = 1)\} > p_{0.95})$;
4. Avoid prediction (classification) $\Delta = \mathsf{I}(\max_{j \in J}\{p(y_{i+1,j} = 1)\} \leq p_{0.95})$.

Inference with confidence

a) In-domain difficult cases

Model selection/ aka pruning

1. Median probability model (**MED**): select all $\beta_{kj}^{(l)} : p(\gamma_{kj}^{(l)} = 1) \geq 0.5$;
2. Posterior mean based model (**MN**): fix $\beta_{kj}^{(l)} = \hat{E}\{\beta_{kj}^{(l)} | \mathbb{D}\} = \alpha_{kj}^{(l)} \mu_{kj}^{(l)}$;
3. Combination of the two (**MED+MN**): fix $\beta_{kj}^{(l)} = \mathsf{I}(\alpha_{kj}^{(l)} \geq 0.5) \mu_{kj}^{(l)}$;
4. Mode probability model, WAIC, DIC, FIC (not addressed yet);
5. Ad-hoc pruning (**SEL**).

Alternative approaches

The only approach (known to me) that addresses properly jointly model and parameter uncertainty in deep learning is:

- The **deep Bayesian regression model** [Hubin et al., 2018];
- Unfortunately in its current form the approach is not really scalable;

Alternative approaches

The only approach (known to me) that addresses properly jointly model and parameter uncertainty in deep learning is:

- The **deep Bayesian regression model** [Hubin et al., 2018];
- Unfortunately in its current form the approach is not really scalable;

At the same time there are Bayesian approaches for inference on a single BNN, which are only addressing parameter related uncertainty:

- BNNs with a **Gaussian parameter prior**[Graves, 2011];
- BNNs with a **mixture of Gaussians prior**[Blundell et al., 2015];
- BNNs with a **horseshoe prior**[Louizos et al., 2017];
- BNNs with a **log-uniform parameter prior** [Molchanov et al., 2017].

Alternative approaches

The only approach (known to me) that addresses properly jointly model and parameter uncertainty in deep learning is:

- The **deep Bayesian regression model** [Hubin et al., 2018];
- Unfortunately in its current form the approach is not really scalable;

At the same time there are Bayesian approaches for inference on a single BNN, which are only addressing parameter related uncertainty:

- BNNs with a **Gaussian parameter prior**[Graves, 2011];
- BNNs with a **mixture of Gaussians prior**[Blundell et al., 2015];
- BNNs with a **horseshoe prior**[Louizos et al., 2017];
- BNNs with a **log-uniform parameter prior** [Molchanov et al., 2017].

Finally, notice that BNNs with **concrete dropout** *[Gal et al., 2017] can be seen as an approach lying in between:*

- In principle it allows for restricted parameter uncertainty with a strict assumption of having the same inclusion probabilities layer-wisely;
- Its properties related to structural uncertainty are not studied;
- Allows model averaging but not model selection.

Experimental design

Table 1: Inference possibilities for the mentioned. **SM** is a single sample, **MA** - model (sample) averaging, **MN** - posterior mean based inference, **MODE** - selecting the mode probability model, **MED** - selecting the median probability model, **WAIC**, **DIC**, **FIC** - selecting with the corresponding criteria, **SEL** - add hoc model selection, **PT** - post-training of the parameters having the model probabilities fixed, **PE** - point estimates of the predictions, **CI** - credible intervals for the predictions.

Method	Our approach	Gaussian prior	Mixture prior	Concrete dropout	Log-uniform/ Horseshoe prior	Studied
			Dense			
SM	Joint	Par	Par	Joint?	Par	**Yes**
MA	Joint	Par	Par	Joint?	Par	**Yes**
MN	Joint	Par	Par	Joint?	Par	**Yes**
			Model selection			
MODE	Joint	Par	Par	Par	Par	**No**
MED	+	-	-	-	-	**Yes**
WAIC, DIC, FIC	+	-	-	-	-	**No**
SEL	+	?	?	?	+	**Yes**
			Post training			
PT	MA/MS	-	-	MA	MS	**Yes**
			Inference			
PE	Joint	Par	Par	Joint?	Par	**Yes**
CI	Joint	Par	Par	Joint?	Par	**Yes**

The model for the experiments

Dense neural network with:

- ReLU activation function;
- multinomially distributed observations with 10 classes and 784 input explanatory variables (pixels);
- 3 hidden layers with 400, 600 and 600 neurons correspondingly;
- Priors as follows:

$$\beta_{kj}^{(l)} \overset{iid}{\sim} \mathsf{I}(\gamma_{kj}^{(l)} = 1) N(0, 1);$$

$$p(\gamma_{kj}^{(l)} = 1) \overset{iid}{\propto} \exp(-2);$$

- ADAM optimizer, 250 epochs, 100 batch size;
- Post-training - another 50 epochs.

Results. MNIST test data

Method	Acc. All Med (Min,Max)	Acc. 95% CI Med (Min,Max)	Clsf. 95% CI Med	Density Med
Full BNN with Gaussian priors				
SM	0.958 (0.954,0.960)	-	-	0.056
SM+PT	0.971 (0.969,0.973)	-	-	0.056
MA	0.967 (0.966,0.971)	0.999 (0.999,0.999)	7064	0.084
MA+PT	0.978 (0.976,0.980)	0.999 (0.999,1.000)	8366	0.084
MN	0.969 (0.967,0.970)	-	-	1.000
MN+PT	0.979 (0.978,0.980)	-	-	1.000
MED+SM	0.961 (0.957,0.964)	-	-	0.051
MED+MA	0.964 (0.962,0.967)	0.998 (0.997,0.999)	7441	0.051
MED+MN	0.965 (0.963,0.968)	-	-	0.051
MED+SM+PT	0.973 (0.971,0.977)	-	-	0.051
MED+MA+PT	0.977 (0.976,0.979)	0.999 (0.998,0.999)	8645	0.051
MED+MN+PT	0.978 (0.976,0.979)	-	-	0.051

Results. MNIST test data

Method	Acc. All Med (Min,Max)	Acc. 95% CI Med (Min,Max)	Clsf. 95% CI Med	Density Med
Full BNN with Gaussian priors				
SM	0.958 (0.954,0.960)	-	-	0.056
SM+PT	0.971 (0.969,0.973)	-	-	0.056
MA	0.967 (0.966,0.971)	0.999 (0.999,0.999)	7064	0.084
MA+PT	0.978 (0.976,0.980)	0.999 (0.999,1.000)	8366	0.084
MN	0.969 (0.967,0.970)	-	-	1.000
MN+PT	0.979 (0.978,0.980)	-	-	1.000
MED+SM	0.961 (0.957,0.964)	-	-	0.051
MED+MA	0.964 (0.962,0.967)	0.998 (0.997,0.999)	7441	0.051
MED+MN	0.965 (0.963,0.968)	-	-	0.051
MED+SM+PT	0.973 (0.971,0.977)	-	-	0.051
MED+MA+PT	0.977 (0.976,0.979)	0.999 (0.998,0.999)	8645	0.051
MED+MN+PT	0.978 (0.976,0.979)	-	-	0.051
Dense BNN with Gaussian priors				
SM	0.965 (0.965,0.966)	-	-	1.000
MA	0.984 (0.982,0.985)	0.999 (0.999,1.000)	8477	1.000
MN	0.984 (0.982,0.985)	-	-	1.000

Results. MNIST test data

Method	Acc. All Med (Min,Max)	Acc. 95% CI Med (Min,Max)	Clsf. 95% CI Med	Density Med
Full BNN with Gaussian priors				
SM	0.958 (0.954,0.960)	-	-	0.056
SM+PT	0.971 (0.969,0.973)	-	-	0.056
MA	0.967 (0.966,0.971)	0.999 (0.999,0.999)	7064	0.084
MA+PT	0.978 (0.976,0.980)	0.999 (0.999,1.000)	8366	0.084
MN	0.969 (0.967,0.970)	-	-	1.000
MN+PT	0.979 (0.978,0.980)	-	-	1.000
MED+SM	0.961 (0.957,0.964)	-	-	0.051
MED+MA	0.964 (0.962,0.967)	0.998 (0.997,0.999)	7441	0.051
MED+MN	0.965 (0.963,0.968)	-	-	0.051
MED+SM+PT	0.973 (0.971,0.977)	-	-	0.051
MED+MA+PT	0.977 (0.976,0.979)	0.999 (0.998,0.999)	8645	0.051
MED+MN+PT	0.978 (0.976,0.979)	-	-	0.051
Dense BNN with Gaussian priors				
SM	0.965 (0.965,0.966)	-	-	1.000
MA	0.984 (0.982,0.985)	0.999 (0.999,1.000)	8477	1.000
MN	0.984 (0.982,0.985)	-	-	1.000
Dense BNN with mixture priors				
SM	0.965 (0.964,0.967)	-	-	1.000
MA	0.982 (0.981,0.983)	0.999 (0.999,1.000)	8329	1.000
MN	0.983 (0.981,0.984)	-	-	1.000

Results. MNIST test data

Method	Acc. All Med (Min,Max)	Acc. 95% CI Med (Min,Max)	Clsf. 95% CI Med	Density Med
Full BNN with Gaussian priors				
SM	0.958 (0.954,0.960)	-	-	0.056
SM+PT	0.971 (0.969,0.973)	-	-	0.056
MA	0.967 (0.966,0.971)	0.999 (0.999,0.999)	7064	0.084
MA+PT	0.978 (0.976,0.980)	0.999 (0.999,1.000)	8366	0.084
MN	0.969 (0.967,0.970)	-	-	1.000
MN+PT	0.979 (0.978,0.980)	-	-	1.000
MED+SM	0.961 (0.957,0.964)	-	-	0.051
MED+MA	0.964 (0.962,0.967)	0.998 (0.997,0.999)	7441	0.051
MED+MN	0.965 (0.963,0.968)	-	-	0.051
MED+SM+PT	0.973 (0.971,0.977)	-	-	0.051
MED+MA+PT	0.977 (0.976,0.979)	0.999 (0.998,0.999)	8645	0.051
MED+MN+PT	0.978 (0.976,0.979)	-	-	0.051
Dense BNN with Gaussian priors				
SM	0.965 (0.965,0.966)	-	-	1.000
MA	0.984 (0.982,0.985)	0.999 (0.999,1.000)	8477	1.000
MN	0.984 (0.982,0.985)	-	-	1.000
Dense BNN with mixture priors				
SM	0.965 (0.964,0.967)	-	-	1.000
MA	0.982 (0.981,0.983)	0.999 (0.999,1.000)	8329	1.000
MN	0.983 (0.981,0.984)	-	-	1.000
BNN with Concrete dropout				
SM	0.982 (0.894,0.984)	-	-	0.226
MA	0.984 (0.896,0.986)	0.995 (0.994,0.996)	9581	0.820
MN	0.983 (0.896,0.984)	-	-	1.000
SM+PT	0.982 (0.894,0.984)	-	-	0.226
MA+PT	0.984 (0.896,0.986)	0.995 (0.994,0.996)	9586	0.820
MN+PT	0.983 (0.894,0.984)	-	-	1.000

Results. MNIST test data

Method	Acc. All Med (Min,Max)	Acc. 95% CI Med (Min,Max)	Clsf. 95% CI Med	Density Med
Full BNN with Gaussian priors				
SM	0.958 (0.954,0.960)	-	-	0.056
SM+PT	0.971 (0.969,0.973)	-	-	0.056
MA	0.967 (0.966,0.971)	0.999 (0.999,0.999)	7064	0.084
MA+PT	0.978 (0.976,0.980)	0.999 (0.999,1.000)	8366	0.084
MN	0.969 (0.967,0.970)	-	-	1.000
MN+PT	0.979 (0.978,0.980)	-	-	1.000
MED+SM	0.961 (0.957,0.964)	-	-	0.051
MED+MA	0.964 (0.962,0.967)	0.998 (0.997,0.999)	7441	0.051
MED+MN	0.965 (0.963,0.968)	-	-	0.051
MED+SM+PT	0.973 (0.971,0.977)	-	-	0.051
MED+MA+PT	0.977 (0.976,0.979)	0.999 (0.998,0.999)	8645	0.051
MED+MN+PT	0.978 (0.976,0.979)	-	-	0.051
Dense BNN with Gaussian priors				
SM	0.965 (0.965,0.966)	-	-	1.000
MA	0.984 (0.982,0.985)	0.999 (0.999,1.000)	8477	1.000
MN	0.984 (0.982,0.985)	-	-	1.000
Dense BNN with mixture priors				
SM	0.965 (0.964,0.967)	-	-	1.000
MA	0.982 (0.981,0.983)	0.999 (0.999,1.000)	8329	1.000
MN	0.983 (0.981,0.984)	-	-	1.000
BNN with Concrete dropout				
SM	0.982 (0.894,0.984)	-	-	0.226
MA	0.984 (0.896,0.986)	0.995 (0.994,0.996)	9581	0.820
MN	0.983 (0.896,0.984)	-	-	1.000
SM+PT	0.982 (0.894,0.984)	-	-	0.226
MA+PT	0.984 (0.896,0.986)	0.995 (0.994,0.996)	9586	0.820
MN+PT	0.983 (0.894,0.984)	-	-	1.000
Dense BNN with horseshoe priors				
SM	0.964 (0.962,0.967)	-	-	1.000
MA	0.982 (0.981,0.983)	1.000 (0.000,1.000)	0003	1.000
MN	0.966 (0.963,0.968)	-	-	1.000
Sparse BNN with horseshoe priors				
SM	0.965 (0.962,0.969)	-	-	0.194
MA	0.982 (0.981,0.983)	1.000 (0.000,1.000)	0002	0.194
MN	0.965 (0.963,0.968)	-	-	0.194
SEL+SM+PT	0.967 (0.965,0.968)	-	-	0.194
SEL+MA+PT	0.982 (0.981,0.983)	1.000 (1.000,1.000)	0007	0.194
SEL+MN+PT	0.966 (0.964,0.969)	-	-	0.194

Results. FMNIST test data

Method	Acc. All Med (Min,Max)	Acc. 95% Med (Min,Max)	Clsf. 95% Med	Density Med
Full BNN with Gaussian priors				
SM	0.854 (0.850,0.858)	-	-	0.066
SM+PT	0.868 (0.863,0.872)	-	-	0.066
MA	0.867 (0.863,0.870)	0.996 (0.994,0.997)	4097	0.083
MA+PT	0.880 (0.875,0.882)	0.994 (0.993,0.995)	4933	0.083
MN	0.866 (0.864,0.874)	-	-	1.000
MN+PT	0.880 (0.877,0.884)	-	-	1.000
MED+SM	0.858 (0.854,0.865)	-	-	0.065
MED+MA	0.863 (0.859,0.869)	0.993 (0.990,0.996)	4347	0.065
MED+MN	0.863 (0.859,0.870)	-	-	0.065
MED+SM+PT	0.872 (0.870,0.875)	-	-	0.065
MED+MA+PT	0.878 (0.876,0.881)	0.992 (0.990,0.993)	5223	0.065
MED+MN+PT	0.879 (0.876,0.882)	-	-	0.065
Dense BNN with Gaussian priors				
SM	0.864 (0.863,0.866)	-	-	1.000
MA	0.893 (0.890,0.894)	0.997 (0.995,0.997)	5089	1.000
MN	0.886 (0.882,0.888)	-	-	1.000
Dense BNN with mixture priors				
SM	0.867 (0.866,0.868)	-	-	1.000
MA	0.893 (0.892,0.897)	0.996 (0.995,0.997)	5151	1.000
MN	0.888 (0.885,0.890)	-	-	1.000
BNN with Concrete dropout				
SM	0.896 (0.820,0.902)	-	-	0.094
MA	0.897 (0.823,0.901)	0.942 (0.941,0.951)	8825	0.447
MN	0.896 (0.821,0.901)	-	-	1.000
SM+PT	0.897 (0.820,0.899)	-	-	0.094
MA+PT	0.897 (0.823,0.902)	0.943 (0.940,0.950)	8826	0.447
MN+PT	0.896 (0.820,0.901)	-	-	1.000
Dense BNN with horseshoe priors				
SM	0.864 (0.863,0.869)	-	-	1.000
MA	0.887 (0.886,0.889)	1.000 (1.000,1.000)	0181	1.000
MN	0.867 (0.861,0.868)	-	-	1.000
Sparse BNN with horseshoe priors				
SM	0.865 (0.860,0.868)	-	-	0.302
MA	0.887 (0.884,0.888)	1.000 (1.000,1.000)	0179	0.302
MN	0.865 (0.862,0.869)	-	-	0.302
SEL+SM+PT	0.867 (0.864,0.871)	-	-	0.302
SEL+MA+PT	0.888 (0.887,0.890)	1.000 (1.000,1.000)	0147	0.302
SEL+MN+PT	0.868 (0.864,0.869)	-	-	0.302

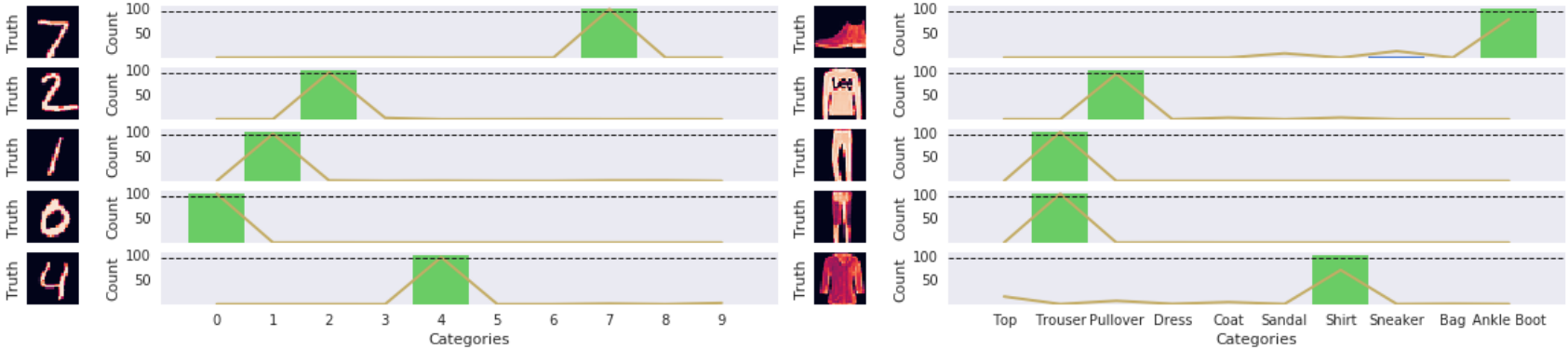
Marginal inclusion probabilities

Table 2: Average (per layer) marginal inclusion probabilities for the full BNN model for both MNIST and FMNIST experiments.

	MNIST data		**FMNIST data**	
Layer	Med.	SD.	Med.	SD.
$\rho(\gamma^{(1)}\|\mathbb{D})$	0.0520	0.0005	0.0665	0.0004
$\rho(\gamma^{(2)}\|\mathbb{D})$	0.0598	0.0003	0.0613	0.0005
$\rho(\gamma^{(3)}\|\mathbb{D})$	0.2217	0.0064	0.2013	0.0051

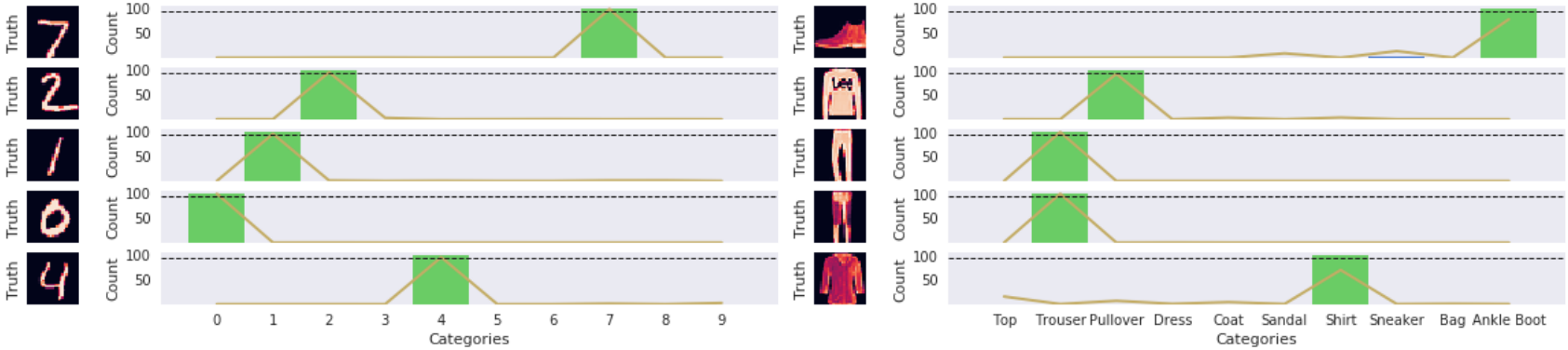
Out of sample classification

a) In-domain uncertainty

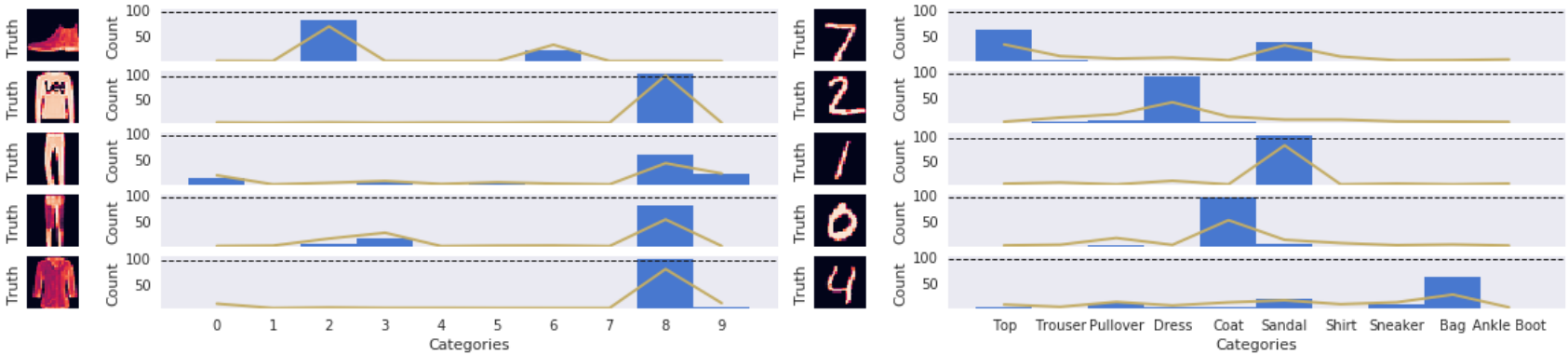


Out of sample classification

a) In-domain uncertainty

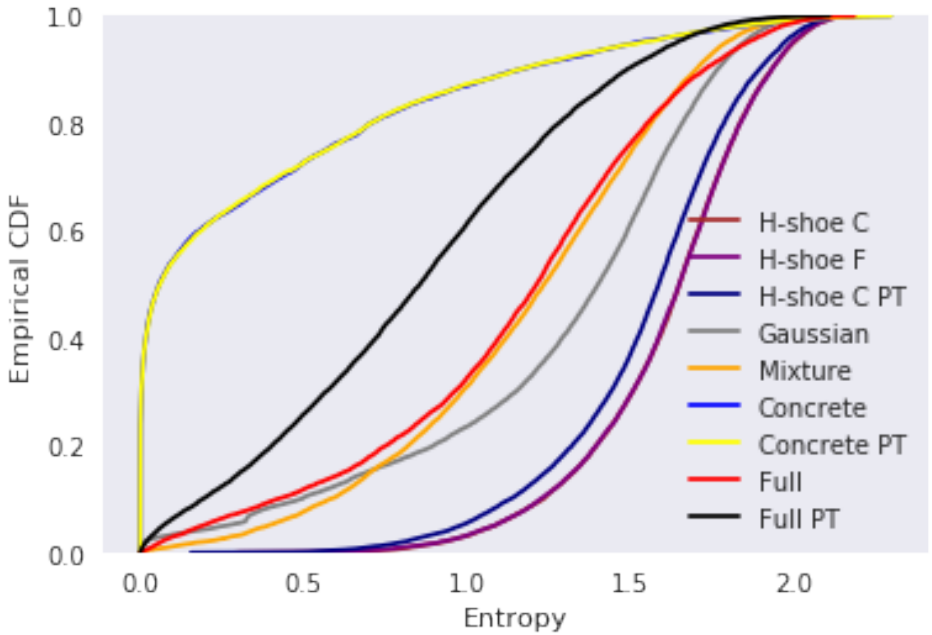


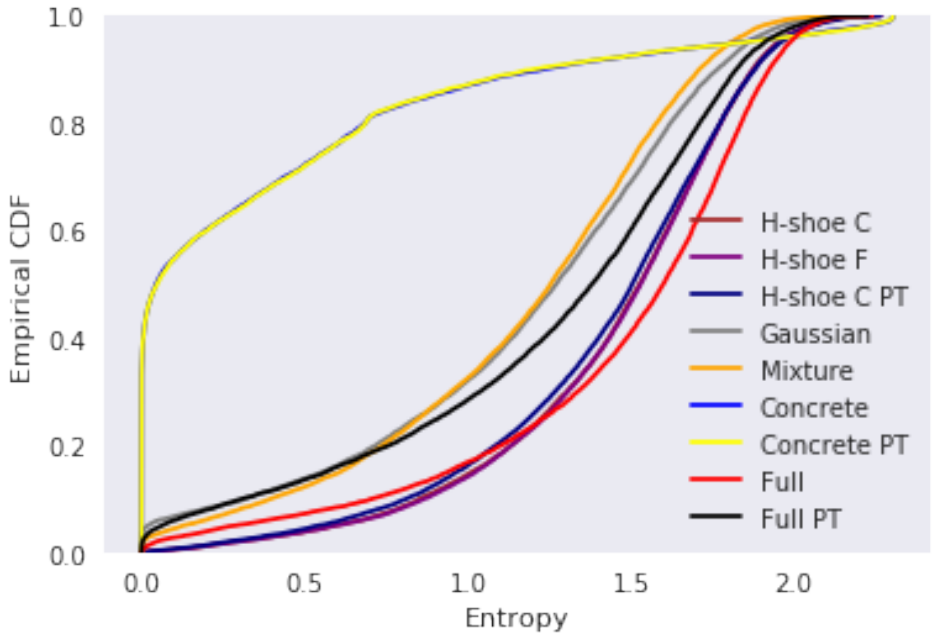
b) Out-of-domain uncertainty



Out of sample classification

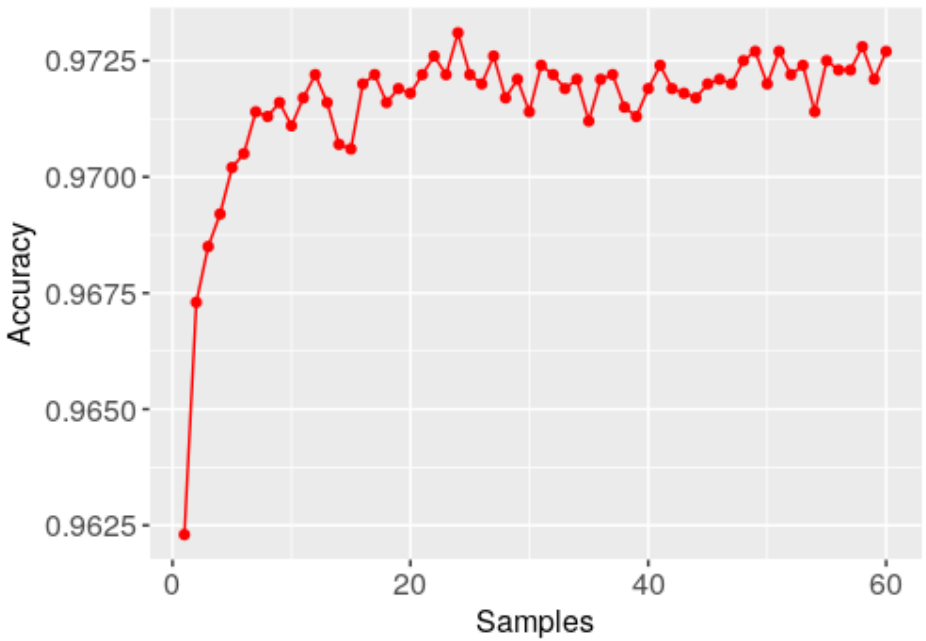
Entropies for the out of domain predictions (the closer to uniform: i.e. right bottom corner - the better) on the MNIST data (left) and FMNIST data (right) for simulation $s = 10$.

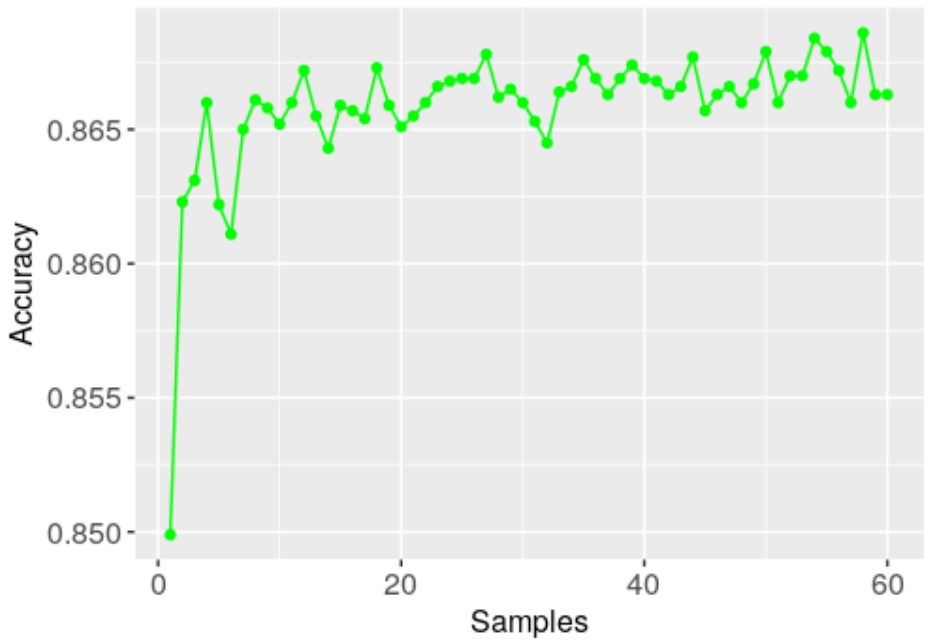




Improvements with larger samples

Accuracy of predictions versus the number of samples from the joint posterior of models and parameters R on the MNIST data (left) and FMNIST data (right) for simulation $s = 10$.





Further results

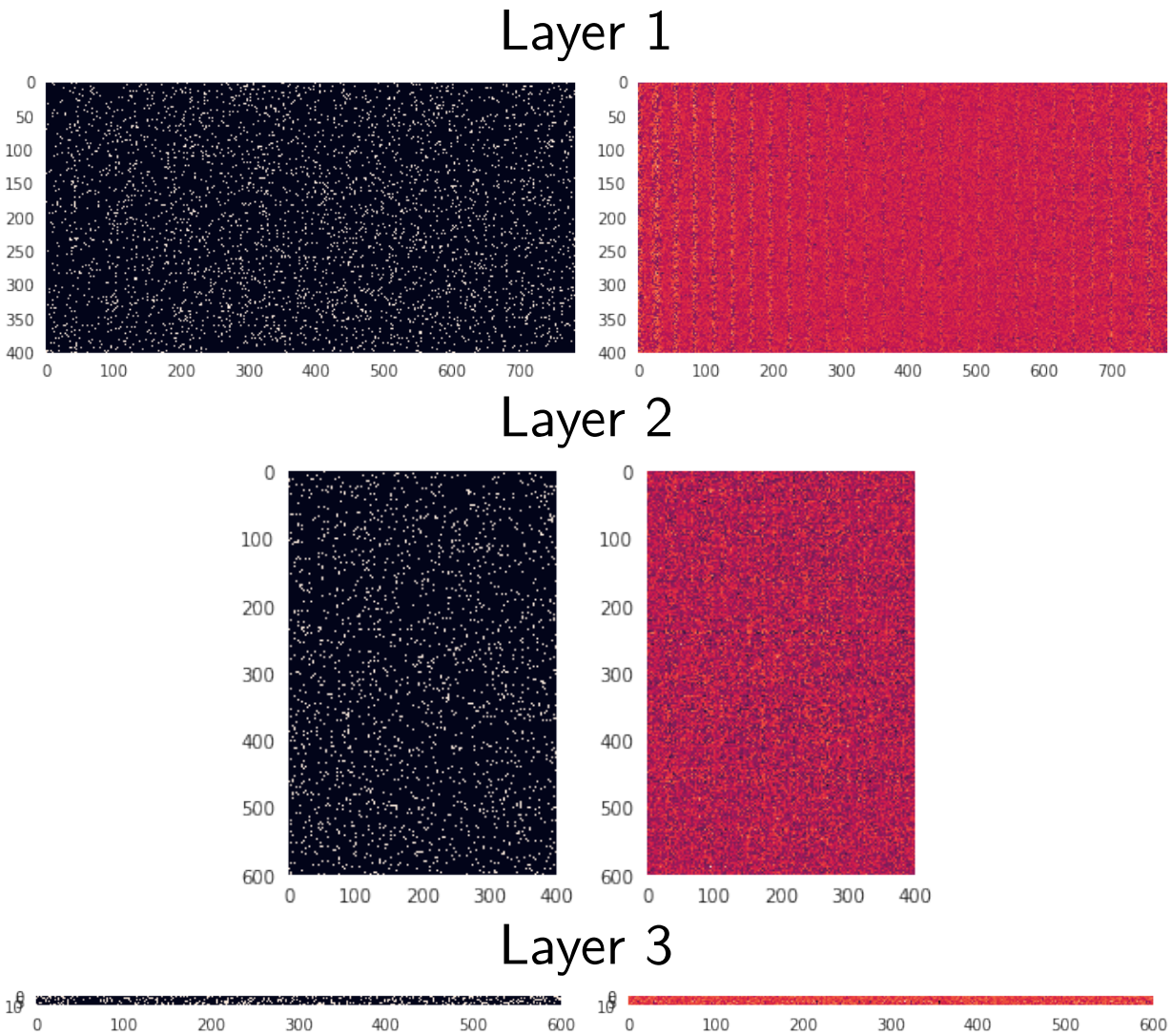


Figure 1: Random samples from the model space (left) and the weight matrices (right) for FMNIST data.

Concluding remarks

- We have developed a scalable joint model-parameter approximate inference approach in the class of BNNs;
- The approach allows to perform proper Bayesian model selection and model averaging;
- Both model selection and model averaging for our choice of priors lead to drastic sparsification of BNNs with no loss of predictive power;
- Furthermore, both model selection and model averaging within our approach allow for very accurate and robust handling of predictive uncertainty, as shown in our experiments;
- However, the VB approach generally is extremely biased and the ways to reduce the bias must be studied;
- Moreover, estimates of the gradient can be noisy, further variance reduction improvements should be addressed;
- The paper is on arXiv: [Hubin and Storvik, 2019];
- We would also like to extend the current approach DBRM.

References I

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015).

Weight uncertainty in neural networks.
arXiv preprint arXiv:1505.05424.

Carbonetto, P., Stephens, M., et al. (2012).

Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies.
Bayesian analysis, 7(1):73–108.

Gal, Y., Hron, J., and Kendall, A. (2017).

Concrete dropout.
In *Advances in Neural Information Processing Systems*, pages 3581–3590.

Graves, A. (2011).

Practical variational inference for neural networks.
In *Advances in Neural Information Processing Systems*, pages 2348–2356.

Hubin, A. and Storvik, G. (2019).

Combining model and parameter uncertainty in bayesian neural networks.
arXiv preprint arXiv:1903.07594.

Hubin, A., Storvik, G., and Frommlet, F. (2018).

Deep Bayesian regression models.
arXiv preprint arXiv:1806.02160.

Louizos, C., Ullrich, K., and Welling, M. (2017).

Bayesian compression for deep learning.
In *Advances in Neural Information Processing Systems*, pages 3288–3298.

Molchanov, D., Ashukha, A., and Vetrov, D. (2017).

Variational dropout sparsifies deep neural networks.
arXiv preprint arXiv:1701.05369.

References II

Rue, H., Martino, S., and Chopin, N. (2009).

Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations.
Journal of the Royal Statistical Sosciety, 71(2):319–392.

Time left? - Deep Bayesian regression models

$$Y_i|\mu_i, \phi \sim \mathfrak{f}(y|\mu_i; \phi), \quad i \in \{1, ..., n\} \tag{11}$$

$$\mu_i = h^{-1}\left(\beta_0 + \sum_{j=1}^{p} \gamma_j \beta_j F_j(\boldsymbol{x}) + \sum_{k=1}^{r} \gamma_{k+p} \delta_{ik}\right), \tag{12}$$

$$\boldsymbol{\delta_k} = (\delta_{1k}, ..., \delta_{nk}) \sim N_n(\mathbf{0}, \boldsymbol{\Sigma}_k). \tag{13}$$

- $f(\cdot|\mu, \phi)$ is a density/distribution with expectation μ and dispersion parameter ϕ;
- $F_j(\boldsymbol{x})$ are all features based on the input explanatory variables ordered w.r.t. complexity, p is the finite number of allowed features;
- $\beta_j \in \mathbb{R}, j \in \{1, ..., p\}$ are regression coefficients of the features;
- $h(\cdot)$ is a proper link function;
- $\gamma_j \in \{0, 1\}, j \in \{1, ..., q = r + p\}$ are latent indicators defining if a feature is included into the model ($\gamma_j = 1$) or not ($\gamma_j = 0$).

Hierarchy of the features

A feature $F_j(\boldsymbol{x})$ can be constructed recursively through:

$$F_j(\boldsymbol{x}) = \begin{cases} v(F_k(\boldsymbol{x})), & \text{for a modification;} \\ F_k(\boldsymbol{x}) * F_l(\boldsymbol{x}), & \text{for a crossover;} \\ v(\boldsymbol{\alpha}^T \mathbf{F}(\boldsymbol{x})), & \text{for a projection;} \end{cases}$$

- $F_k(\boldsymbol{x})$ and $F_l(\boldsymbol{x})$ are previously defined features $(k, l < j)$;
- $v \in \mathcal{G}$ is one of the allowed basic function from set $\mathcal{G}$;
- $\mathbf{F}(\boldsymbol{x})$ is a sub-vector of all possible features with indexes lower than j;
- A constraint on the complexity of feature $F_j(\boldsymbol{x})$ is defined by a finite number p of all possible features;
- Projections include modifications and crossovers as particular cases.

Types and meaning of functions in $\mathcal{G}$

- ANN: $\text{logit}(x)$, $\tanh(x)$, $\text{erf}(x)$, $\text{ReLU}(x)$;
- Polynomials: $F_k(\boldsymbol{x}) * F_l(\boldsymbol{x}) = \exp\left(\log(F_k(\boldsymbol{x})) + \log(F_l(\boldsymbol{x}))\right)$;
- Logical *AND* and *OR*: $L_k \wedge L_l = L_k * L_l$ and $L_k \vee L_l = L_k + L_l - L_k * L_l$;
- CART: $\text{I}(x \geq 1)$;
- Fourier series: $\sin(Ax)$ and $\cos(Bx)$;
- Fractional polynomials: $x^{\frac{1}{a}} = \exp\left(b \log(x)\right), b = \frac{1}{a}$;
- RNN and Lagged features: $\text{lag}^k(x)$.

The universal approximation theorem by *Hornik (1991)* is applicable if at least one of v functions is strictly monotonous and bounded.

Model priors

$$p(\boldsymbol{\gamma}) \propto \mathbb{I}\left(|\boldsymbol{\gamma}_{1:p}| \leq Q\right) \mathbb{I}\left(|\boldsymbol{\gamma}_{p+1:q}| \leq R\right) \prod_{j=1}^{p} a^{\gamma_j w_j c(F_j(\mathbf{x}))} \prod_{k=p+1}^{q} b^{\gamma_k \omega_k c(\delta_k)}. \quad (14)$$

- $a, b \in (0,1)$, $Q \leq p$, $R \leq r$;
- $|\boldsymbol{\gamma}_{1:K}| = \sum_{j=1}^{K} \gamma_j$ is the number of active features in subset $\{\gamma_1, ..., \gamma_K\}$;
- $c(F_j(\mathbf{x})) \geq 0$ is a measure of complexity for a feature $F_j(\mathbf{x})$;
- $c(\delta_k) \geq 0$ is a measure of complexity for a latent Gaussian variable δ_k;
- w_j are weights for complexities of the corresponding features;
- ω_k are weights for the complexities of the corresponding latent Gaussian variables.

Parameter priors

Particular choices are given in the applications to follow:

$$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim \pi_{\beta}(\boldsymbol{\beta}), \tag{15}$$

$$\boldsymbol{\psi}_k|\boldsymbol{\gamma} \sim \pi_k(\boldsymbol{\psi}_k), \tag{16}$$

$$\phi \sim \pi_{\phi}(\phi). \tag{17}$$

Prior distributions on $\boldsymbol{\beta}_j|\boldsymbol{\gamma}$, ϕ (if present) and $\boldsymbol{\psi}_k|\boldsymbol{\gamma}$ (if latent Gaussian variables are present) are usually selected in a way to efficiently compute marginal likelihoods of the models (by for example specifying conjugate priors) and should be carefully specified for the applications of interest.

Ground physical laws inference

From the **ground physical** laws:

$$m_p \propto R_p^3 \times \rho_p,$$

m_p is *PlanetaryMassJpt*, R_p^3 is *RadiusJpt*^{3}, and ρ_p is *PlanetaryDensJpt*. And:

$$a = \left(\frac{GP^2}{4\pi^2}(M_h + m_p)\right)^{\frac{1}{3}} \approx \left(\frac{GP^2}{4\pi^2}(M_s M_h^a)\right)^{\frac{1}{3}},$$

a is semi major axises of the ellipses of the orbits, M_h^a is *HostStarMassSlr-Mass*, a is *SemiMajorAxisAU*, and P is *PeriodDays*.

Generally the data has the **following other variables**:
TypeFlag, *RadiusJpt*, *PeriodDays*, *PlanetaryMassJpt*, *Eccentricity*, *Host-StarMassSlrMass*, *HostStarRadiusSlrRad*, *HostStarMetallicity*, *HostStarTempK*, *PlanetaryDensJpt* denoted as x_1-x_{10}

Planet mass inference. Results over 100 simulations

Table 3: Power, False Positives (FP) and FDP based on the decision rule that the posterior probability of a feature is larger than 0.25. The feature $PlanetaryRadiusJpt^3$ $PlanetaryDensJpt$ is counted as true positive, all other selected features as false positive.

	GMJMCMC			RGMJMCMC		
Threads	Power	FP	FDP	Power	FP	FDP
16	1.00	0.00	0.00	0.97	0.06	0.058
4	0.79	0.40	0.34	0.61	0.73	0.54
1	0.43	1.21	0.74	0.32	1.67	0.84

3rd Kepler's law inference. Results over 100 simulations

Table 4: Comparison of Results on Example 3 for GMJMCMC and RGMJMCMC using different number of threads. F_1, F_2 and F_3 refer to the number of times the specific features $(HostStarMassSlrMass \times PeriodDays^2)^{\frac{1}{3}}$, $(HostStarRadiusSlrRad \times PeriodDays^2)^{\frac{1}{3}}$ and $(HostStarTempK \times PeriodDays^2)^{\frac{1}{3}}$ had a posterior probability larger than 0.25. Power gives the percentage of runs where at least one of these three features was detected. FP counts the number of other features and FDP is the corresponding false discovery proportion.

	GMJMCMC						RGMJMCMC					
Th	F_1	F_2	F_3	Power	FP	FDP	F_1	F_2	F_3	Power	FP	FDP
64	81	71	1	1.00	0.02	0.013	78	75	2	0.99	0.03	0.019
32	63	58	11	0.99	0.14	0.11	55	57	9	0.95	0.12	0.09
16	34	41	32	0.84	0.46	0.30	31	38	18	0.79	0.68	0.44
4	15	10	16	0.38	1.05	0.62	8	14	8	0.29	1.47	0.83
1	6	5	3	0.13	1.46	0.82	6	4	2	0.12	1.81	0.94

What would have happened with simpler ANN

1. When $\mathcal{G} = \{\text{sigmoid}(x)\}$;
2. When $\mathcal{G} = \{\text{sigmoid}(x)\}$, $D_{max} = 300$, and $P_c = 0$;
3. When $\mathcal{G} = \{\text{sigmoid}(x)\}$, $D_{max} = 300$, and $P_c = 0$ and $p(\gamma_j) \propto 1$.

Table 5: 10 most frequent features detected under scenarios 1, 2 and 3

Fq	Feature	Fq	Feature	Fq	Feature
99	x_3	100	x_3	100	x_3
98	x_3*x_3	72	g_σ(-10.33+0.24x_4-8.83x_8)	54	x_2
93	x_3*x_{10}	64	x_{10}	21	g_σ(-16.91-4.94x_2)
4	x_3*x_3*x_{10}	62	x_2	19	x_9
1	x_9*x_3	16	g_σ(0.21+0.01x_3+0.20x_7)	16	x_5
1	x_9*x_3*x_3	9	x_4	14	x_{10}
1	x_{10}*x_{10}*x_3	7	g_σ(-13.11-7.76x_8-3.33x_2+0.40x_{10})	10	g_σ(6.88×10^9-3.92x_2+ 3.44×10^9 g_σ(-13.57-0.17x_4- 2.84x_2-7.66x_8+0.54x_{10}) -13.76×10^9 g_σ(g_σ(-13.57- 0.17x_4-2.84x_2-7.66x_8+ 0.54x_{10})))
1	x_7*x_3*x_3	5	g_σ(-3.36+2.83x_3+0.21x_3-3.36x_9)	9	x_4
1	x_6*x_3*x_3	3	g_σ(g_σ(-10.33+0.24x_4)-8.83x_8)	8	g_σ(-13.57-0.17x_4- 2.84x_2-7.66x_8+0.54x_{10})
1	x_3*x_3*x_3	3	g_σ(0.15+0.05x_4-0.01x_3+0.15x_7)	7	g_σ(0.21+0.21x_3)
0	Others	4	Others	> 300	Others

Trash features under the non regularized case

Figure 2: A snapshot of features detected under Scenario 3

Thank you!

Ideally models should remain as transparent and dense as possible, or quoting Einstein's famous *"It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience."*