On estimating epistemic uncertainty

Yingzhen Li

Microsoft Research Cambridge, UK

NeurIPS 2019 Bayesian deep learning tutorial on Monday was jammed with curious heads

Type of uncertainty

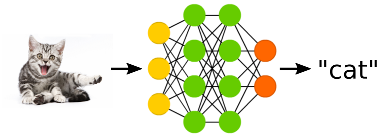
Imagine flipping a coin:

- Epistemic uncertainty: "How much do I believe the coin is fair?"
  - Model's belief after seeing the population
  - Reduces when having more data
- Aleatoric uncertainty: "What's the next coin flip outcome?"
  - Individual experiment outcome
  - Non-reducible
- Distribution shift: "Am I still flipping the same coin?"
  - Indicating changes of the underlying quantity of interest

Bayesian neural networks 101

Let's say we want to classify different types of cats

- $\boldsymbol{x}$: input images; $\boldsymbol{y}$: output label
- build a neural network (with param. W): $p(\boldsymbol{y}|\boldsymbol{x}, W) = \mathrm{softmax}(f_W(\boldsymbol{x}))$



A Bayesian solution:

Put a prior distribution $p(W)$ over W

- compute posterior $p(W|\mathcal{D})$ given a dataset $\mathcal{D} = \{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^N$:

$$p(W|\mathcal{D}) \propto p(W) \prod_{n=1}^{N} p(\boldsymbol{y}_n|\boldsymbol{x}_n, W)$$

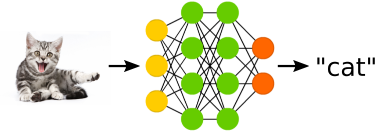
- Bayesian predictive inference:

$$p(\boldsymbol{y}^*|\boldsymbol{x}^*, \mathcal{D}) = \mathbb{E}_{p(W|\mathcal{D})}[p(\boldsymbol{y}^*|\boldsymbol{x}^*, W)]$$

Bayesian neural networks 101

Let's say we want to classify different types of cats

- $\boldsymbol{x}$: input images; $\boldsymbol{y}$: output label
- build a neural network (with param. W):
  $p(\boldsymbol{y}|\boldsymbol{x}, W) = \mathrm{softmax}(f_W(\boldsymbol{x}))$



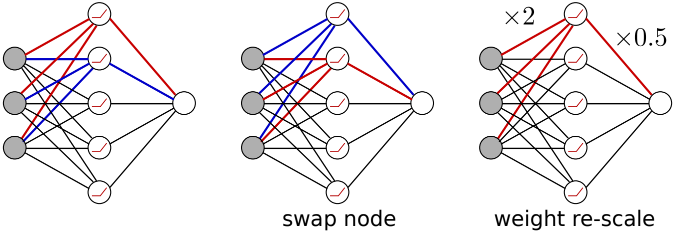
In practice: $p(W|\mathcal{D})$ is intractable

- First find approximation $q(W) \approx p(W|\mathcal{D})$ (e.g. via VI or MCMC)
- In prediction, do Monte Carlo sampling:

$$p(\boldsymbol{y}^*|\boldsymbol{x}^*, \mathcal{D}) \approx \frac{1}{K}\sum_{k=1}^{K} p(\boldsymbol{y}^*|\boldsymbol{x}^*, W^k), \quad W^k \sim q(W)$$

Our qualitative description on epistemic uncertainty is vague...

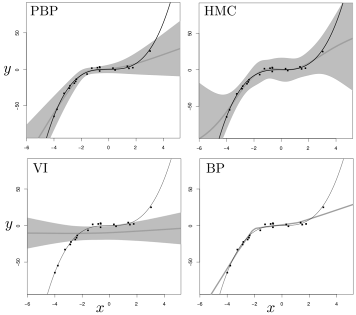
- Weight-space uncertainty is less interesting
  - in many cases neural network weights are NOT scientific parameters
  - symmetries/invariances in parameterisation



Our qualitative description on epistemic uncertainty is vague...

- sample $W \sim q(W) \Leftrightarrow$ sample $f(\cdot) \sim q_{\text{BNN}}(f) \approx q_{\text{BNN}}(f|\mathcal{D})$
- Folklore belief for function-space (or output-space) uncertainty:

"Epistemic uncertainty should be high when new input is less similar to observed inputs"



What do "high uncertainty" and "less similar" mean quantitatively?

Hernández-Lobato and Adams ICML 2015

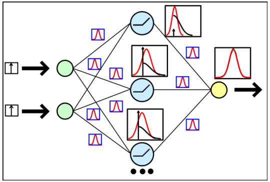
Evaluation by comparing to a reference

BNN performance relies on the approximate posterior:

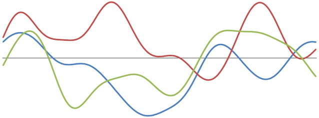
$$q(W) \approx p(W|\mathcal{D}) \propto p(W) \prod_{(x,y)\in\mathcal{D}} p(y|x, W)$$

- Evaluating inference:
  compute some distance metric between $q(W)$ and $p(W|\mathcal{D})$
- Problem: intractable exact posterior $p(W|\mathcal{D})$!
  (even we have no robust way to estimate moments of $p(W|\mathcal{D})$)

Evaluation by comparing to a reference



(a) weight space view



(b) function space view

Function space "reference posterior"[1] for BNN regression:

- wide BNN has GP limit (under certain conditions)
- for regression problems $p_{\text{GP}}(f|\mathcal{D})$ is tractable

$\Rightarrow$ Compare with $p_{\text{GP}}(f|\mathcal{D})$ of the wide-limit GP:

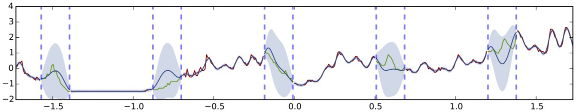
- Is $q_{\text{BNN}}(f)$ close to $p_{\text{GP}}(f|\mathcal{D})$ (at least in the first 2 moments)?

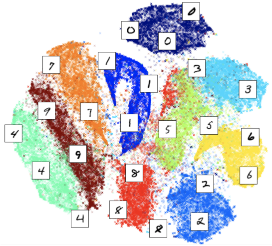
[1] only as reference for inference, no objective Bayesian here

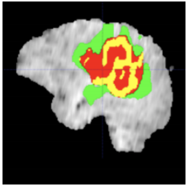
"In-between" uncertainty

"In-between" uncertainty:
uncertainty estimates in regions between data clusters

- Missing values (especially in time series)
- Ambiguous inputs

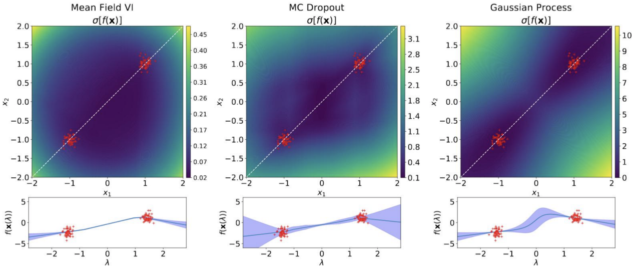






Foong et al. NeurIPS 2019 Bayesian deep learning workshop

"In-between" uncertainty

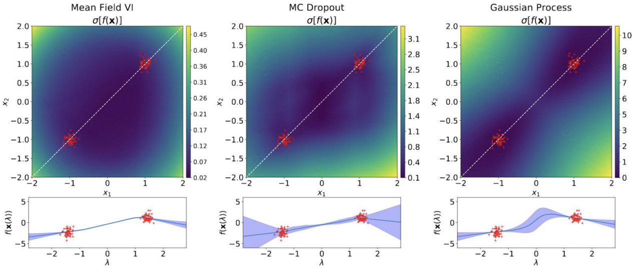


On mean-field Gaussian approximation for BNN regression:

- 1 hidden-layer: bad news for any approximate inference method
  - approximate inference require expressiveness of the q family
  - mean-field has theoretical limitations in representing in-between uncertainty

Foong et al. NeurIPS 2019 Bayesian deep learning workshop

“In-between” uncertainty



On mean-field Gaussian approximation for BNN regression:

- 2+ hidden-layers: mixed news:
  - expressiveness (theory): can represent any mean & variance function
  - algorithm (practice): weight-space VI + optimisation is to be blamed
  - increasing depth does not seem to help to close the gap between MFVI and GP-limit reference

Foong et al. NeurIPS 2019 Bayesian deep learning workshop

"Your GP-posterior reference is also subjective..."

Model selection for BNN in practice:

- Select model + inference together
  (we almost never try testing the same model with multiple inference checks)
- Criteria based on statistics of total uncertainty
  (or balancing between aleatory uncertainty and epistemic uncertainty)
- We often look at averaged metrics only
  (even when test examples can be different from training ones in very different ways)

"Your GP-posterior reference is also subjective..."

Good practical performance can come from

- A good model paired with (close-to) exact inference
- A bad model with a bad approximate inference
  (e.g. VI can return good results when the model with exact inference is under-confident)

Selecting the second pipeline:
do we expect to inherent benefits from Bayesian inference?

An online learning example

- Start from a bad model $p(W)p(y|x, W)$
- Observe the first task $\mathcal{D}_1 = \{(x, y)\}$, perform bad inference to obtain

$$q_1(W) \approx p(W|\mathcal{D}_1)$$

- $q_1(W)$ somehow returns good practical performance even when $p(W|\mathcal{D}_1)$ is bad
- then observe another task $\mathcal{D}_2$ that is similar to $\mathcal{D}_1$
    - Following online Bayesian learning, should compute

$$q_2(W) \approx \tilde{p}(W|\mathcal{D}_2) \propto p(\mathcal{D}_2|W)q_1(W)$$

    - do we still expect good pratical performance for $q_2(W)$?

What I'd love to see in future research...

- Scalable & accurate function space inference methods for BNNs
  (or improve GP/kernel methods?)
- Understand better the gap between exact/approx. inference
  (and potentially fix it)
- Better descriptions on what we really want from modelling uncertainty
  (e.g. evaluate statistics of uncertainty within data subgroups)



Thank you!

References

Neal 1994. Bayesian Learning for Neural Networks. PhD thesis

Hernández-Lobato and Adams. Probabilistic Backpropagation for Scalable Learning of Bayesian Neural Networks. ICML 2015

Matthews et al. 2018. Gaussian Process Behaviour in Wide Deep Neural Networks. ICLR 2018

Foong et al. 2019. Pathologies of Factorised Gaussian and MC Dropout Posteriors in Bayesian Neural Networks. arXiv:1909.00719

`http://mlg.eng.cam.ac.uk/yarin/blog_images/Solar_GP_SE.jpg`

`https://bigsnarf.wordpress.com/2016/11/17/t-sne-attack-data/`