YOUR EPISTEMIC UNCERTAINTY IS SECRETLY A DENSITY ESTIMATION AND YOU SHOULD TREAT IT LIKE ONE

Pitfalls of Uncertainty Quantification @ WUML 2024

DLR

Nis Meinert, Institute of Communications and Navigation, February 2024

Deep Evidential Regression (DER)

Deep evidential regression [PDF] neurips.cc

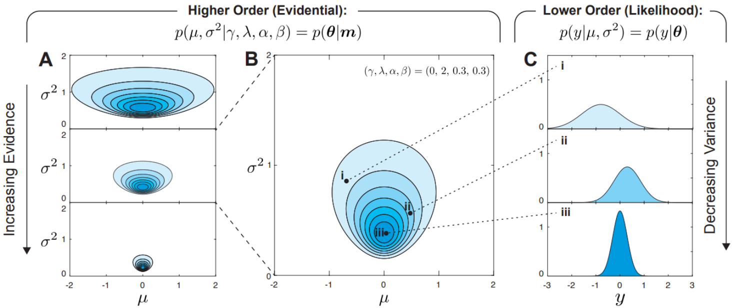
A Amini, W Schwarting… - Advances in Neural …, 2020 - proceedings.neurips.cc

Deterministic neural networks (NNs) are increasingly being deployed in safety critical domains, where calibrated, robust, and efficient measures of uncertainty are crucial. In this paper, we propose a novel method for training non-Bayesian NNs to estimate a continuous target as well as its associated evidence in order to learn both aleatoric and epistemic uncertainty. We accomplish this by placing evidential priors over the original Gaussian likelihood function and training the NN to infer the hyperparameters of the evidential …

Save Cite Cited by 315 Related articles All 9 versions

DOI:10.1007/s10618-021-00782-4 (NeurIPS 2020)

DER in a Nutshell



(taken from Amini et al., *Deep Evidential Regression*, NeurIPS 2020)

Train a NN $\boldsymbol{m} : \mathcal{X} \times \Omega \to \mathbb{R}^4$ s.t. $\mathbb{E}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta}|\boldsymbol{m}(\boldsymbol{x_i}, \boldsymbol{\omega}))}[p(y|\boldsymbol{\theta})] \to y_i$

DER in a Nutshell

- Minimize

$$\mathcal{L}_i(\boldsymbol{\omega}) = \underbrace{-\log L_i^{\mathrm{NIG}}(\boldsymbol{\omega})}_{\mathcal{L}_i^{\mathrm{NLL}}} + \lambda \underbrace{|y_i - \gamma_i|(2\nu_i + \alpha_i)}_{\mathcal{L}_i^{\mathrm{R}}(\boldsymbol{\omega})}$$

where

$$L_i^{\mathrm{NIG}}(\boldsymbol{\omega}) = p(y_i | \underbrace{\gamma_i, \nu_i, \alpha_i, \beta_i}_{\boldsymbol{m}(\boldsymbol{x}_i;\boldsymbol{\omega})}) = \mathrm{St}_{2\alpha_i}\left(y_i \middle| \gamma_i, \frac{\beta_i(1+\nu_i)}{\nu_i \alpha_i}\right)$$

- Find uncertainties with

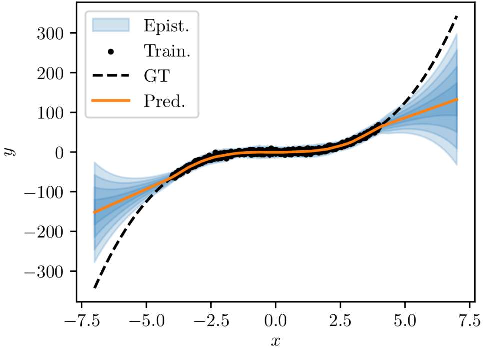
$$\underbrace{\mathbb{E}\left[\sigma_i^2\right] = \frac{\beta_i}{\alpha_i - 1}}_{\text{aleatoric}} \qquad \underbrace{\mathrm{var}\left[\mu_i\right] = \frac{\mathbb{E}\left[\sigma_i^2\right]}{\nu_i}}_{\text{epistemic}}$$

DER in a Nutshell

Minimize

$$-\log \mathrm{St}(y_i|\boldsymbol{m}_i) + \lambda \mathcal{L}_i^{\mathrm{R}}$$

aka **fitting** $\mathrm{St}(y_i|x_i)$ **point-wise** to data $(\boldsymbol{x}_i, y_i)$ with *regularization*.

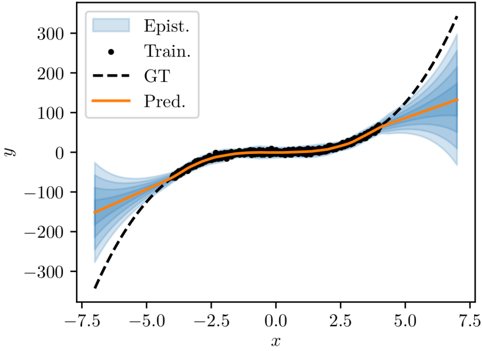


DER in a Nutshell

Minimize

$$-\log \mathrm{St}_{2\alpha_i}\left(y_i \middle| \gamma_i, \frac{\beta_i(1+\nu_i)}{\nu_i\alpha_i}\right) + \lambda \left|y_i - \gamma_i\right| (2\nu_i + \alpha_i)$$

aka **fitting** $\mathrm{St}(y_i|x_i)$ **point-wise** to data $(\boldsymbol{x}_i, y_i)$ with *regularization*.



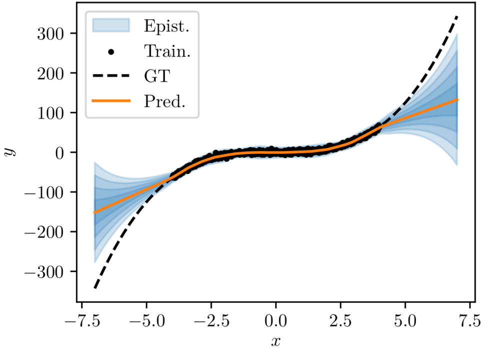
DER in a Nutshell

Minimize

$$-\log \mathrm{St}_{2\alpha_i}\left(y_i \middle| \gamma_i, \frac{\beta_i(1+\nu_i)}{\nu_i\alpha_i}\right) + \lambda \left|y_i - \gamma_i\right| (2\nu_i + \alpha_i)$$

aka **fitting** $\mathrm{St}(y_i|x_i)$ **point-wise** to data $(\boldsymbol{x}_i, y_i)$ with *regularization*.

…but there is no unique minimum!



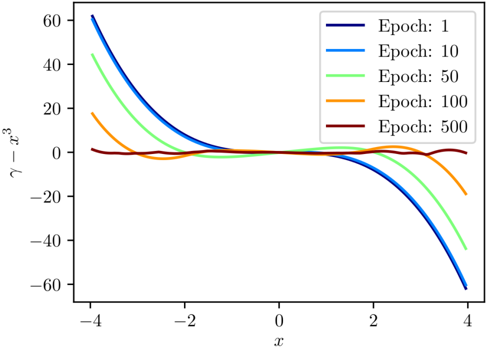
DER in a Nutshell

Minimize

$$-\log \mathrm{St}_{2\alpha_i}\left(y_i \middle| \gamma_i, \frac{\beta_i(1+\nu_i)}{\nu_i\alpha_i}\right) + \lambda\,|y_i - \gamma_i|\,(2\nu_i + \alpha_i)$$

aka **fitting** $\mathrm{St}(y_i|x_i)$ **point-wise** to data $(\boldsymbol{x}_i, y_i)$ with *regularization*.

...but there is no unique minimum!



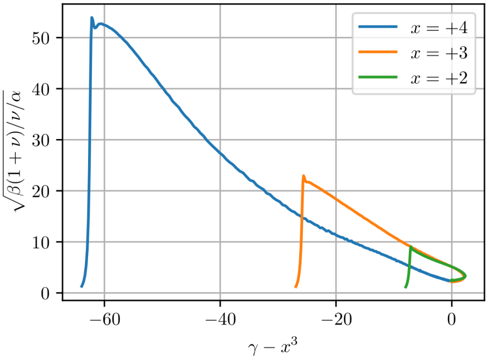
DER in a Nutshell

Minimize

$$-\log \mathrm{St}_{2\alpha_i}\left(y_i \middle| \gamma_i, \frac{\beta_i(1+\nu_i)}{\nu_i\alpha_i}\right) + \lambda\,|y_i-\gamma_i|\,(2\nu_i+\alpha_i)$$

aka **fitting** $\mathrm{St}(y_i|x_i)$ **point-wise** to data $(\boldsymbol{x}_i, y_i)$ with *regularization*.

...but there is no unique minimum!



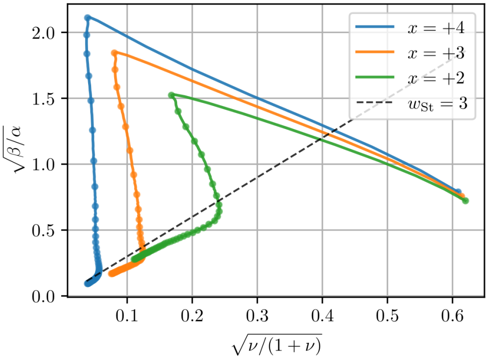
DER in a Nutshell

Minimize

$$-\log \mathrm{St}_{2\alpha_i}\left(y_i \middle| \gamma_i, \frac{\beta_i(1+\nu_i)}{\nu_i\alpha_i}\right) + \lambda\,|y_i - \gamma_i|\,(2\nu_i + \alpha_i)$$

aka **fitting** $\mathrm{St}(y_i|x_i)$ **point-wise** to data $(\boldsymbol{x}_i, y_i)$ with *regularization*.

…but there is no unique minimum!



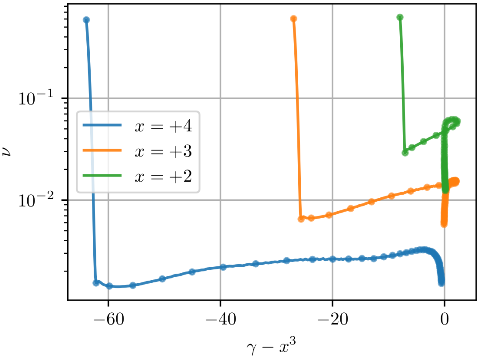
DER in a Nutshell

Minimize

$$-\log \mathrm{St}_{2\alpha_i}\left(y_i \middle| \gamma_i, \frac{\beta_i(1+\nu_i)}{\nu_i\alpha_i}\right) + \lambda\,|y_i - \gamma_i|\,(2\nu_i + \alpha_i)$$

aka **fitting** $\mathrm{St}(y_i|x_i)$ **point-wise** to data $(\boldsymbol{x}_i, y_i)$ with *regularization*.

...but there is no unique minimum!

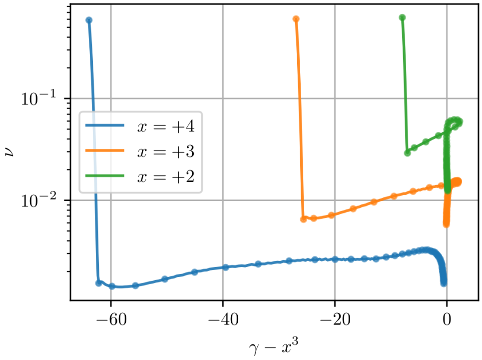


Measuring Epistemic Uncertainty by Convergence Speed

Epistemic uncertainty:

$$\mathrm{var}\left[\mu_i\right] \propto \nu_i^{-1}$$

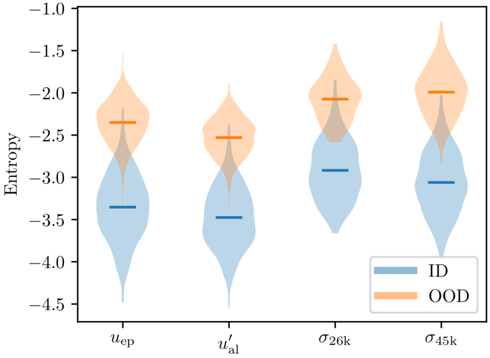
...is measured by point-wise convergence speed!?



How is DER Used in Practice?

In practice: OOD

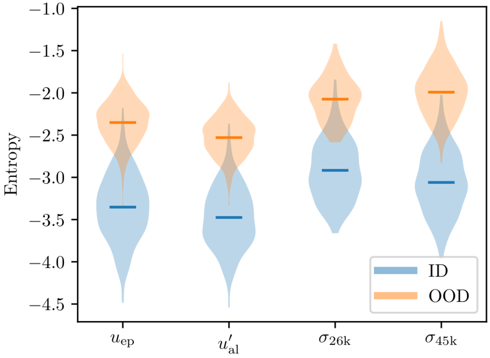
- low epistemic uncertainty?
  - → great! Trust model prediction (**within aleatoric bounds**)
- high epistemic uncertainty?
  - → OOD
  - → don't trust model (**ignore aleatoric uncertainty**)
  - → resample data in this region



How is DER Used in Practice?

In practice: OOD

- low epistemic uncertainty?
  - → great! Trust model prediction (**within aleatoric bounds**)
- high epistemic uncertainty?
  - → OOD
  - → don't trust model (**ignore aleatoric uncertainty**)
  - → resample data in this region

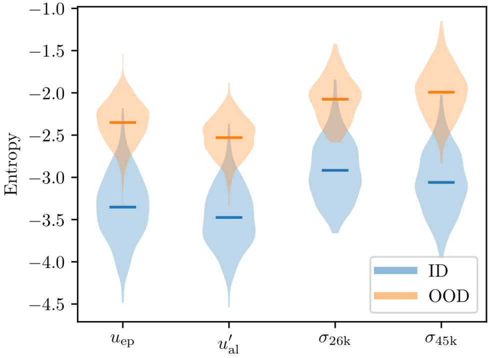


Do we actually need to disentangle types of uncertainties?

How is DER Used in Practice?

In practice: OOD

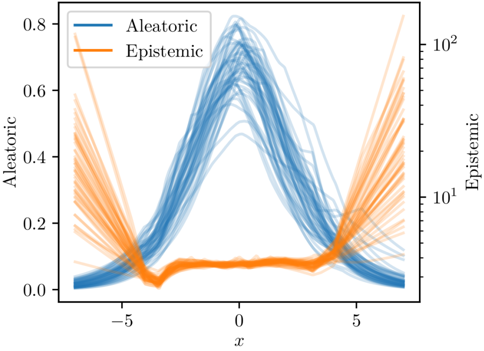
- low epistemic uncertainty?
  - → great! Trust model prediction (**within aleatoric bounds**)
- high epistemic uncertainty?
  - → OOD
  - → don't trust model (**ignore aleatoric uncertainty**)
  - → resample data in this region



Epistemic uncertainty $\leftrightarrow f(\text{convergence speed}) \leftrightarrow f(\text{density})$!?

DER: Nota Bene

- Predicted epistemic uncertainty *somehow* looks *reasonable*
- Aleatoric uncertainty is definitely wrong



find more details in *The Unreasonable Effectiveness of Deep Evidential Regression*, `DOI:10.1609/aaai.v37i8.26096`

Neural Optimization-based Model Uncertainty (NOMU)

Nomu: Neural optimization-based model uncertainty [PDF] arxiv.org

J Heiss, J Weissteiner, H Wutte, S Seuken… - arXiv preprint arXiv …, 2021 - arxiv.org

… Due to its modular architecture, **NOMU** can provide model … **NOMU** in various regressions tasks and noiseless Bayesian optimization (BO) with costly evaluations. In regression, **NOMU** …

Save Cite Cited by 18 Related articles All 8 versions

`arXiv:2102.13640` (ICML 2022)

NOMU: Desiderata

NOMU predicts $\hat{y}$ and epistemic uncertainty as $(\underline{UB}, \overline{UB})$ s.t.

1. **Non-Negativity:** $\underline{UB}(x) \leq \hat{y}(x) \leq \overline{UB}$
2. **In-Sample:** $\underline{UB}(x_{\text{train}}) = \overline{UB}(x_{\text{train}})$
3. **Out-Sample:** $\overline{UB}(x_{\text{train}}) - \underline{UB}(x_{\text{train}})$ grows if $||x - x_{\text{train}}||_{\mathcal{M}}$ gets large
4. **Metric Learning:** $||x - x_{\text{train}}||_{\mathcal{M}}$ strongly depends on features that have high predictive power
5. **Vanishing:** $\overline{UB}(x) - \underline{UB}(x) \to 0$ for $n_{\text{train}} \to 0$

NOMU: Desiderata

NOMU predicts $\hat{y}$ and epistemic uncertainty as $(\underline{UB}, \overline{UB})$ s.t.

1. **Non-Negativity:** $\underline{UB}(x) \leq \hat{y}(x) \leq \overline{UB}$
2. **In-Sample:** $\underline{UB}(x_{\text{train}}) = \overline{UB}(x_{\text{train}})$
3. **Out-Sample:** $\overline{UB}(x_{\text{train}}) - \underline{UB}(x_{\text{train}})$ grows if $||x - x_{\text{train}}||_{\mathcal{M}}$ gets large
4. **Metric Learning:** $||x - x_{\text{train}}||_{\mathcal{M}}$ strongly depends on features that have high predictive power
5. **Vanishing:** $\overline{UB}(x) - \underline{UB}(x) \to 0$ for $n_{\text{train}} \to 0$

$\hookrightarrow$ $((\underline{UB}(x), \overline{UB}(x)) = (f(x) \mp c\,\varphi(r_f(x))$ with NNs f and r_f

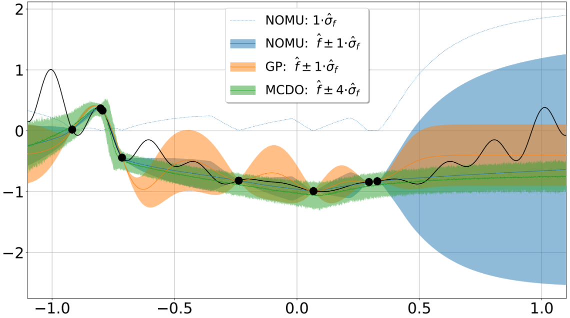
NOMU: Desiderata

NOMU predicts $\hat{y}$ and epistemic uncertainty as $(\underline{UB}, \overline{UB})$ s.t.

1. **Non-Negativity:** $\underline{UB}(x) \leq \hat{y}(x) \leq \overline{UB}$
2. **In-Sample:** $\underline{UB}(x_{\text{train}}) = \overline{UB}(x_{\text{train}})$
3. **Out-Sample:** $\overline{UB}(x_{\text{train}}) - \underline{UB}(x_{\text{train}})$ grows if $||x - x_{\text{train}}||_{\mathcal{M}}$ gets large
4. **Metric Learning:** $||x - x_{\text{train}}||_{\mathcal{M}}$ strongly depends on features that have high predictive power
5. **Vanishing:** $\overline{UB}(x) - \underline{UB}(x) \to 0$ for $n_{\text{train}} \to 0$

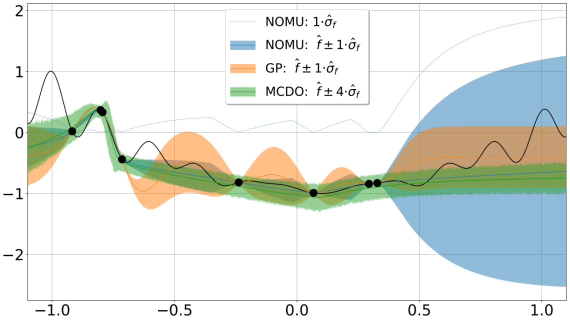
$$\hookrightarrow \arg\min_{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2} \sum_i \left(f(x_i|\boldsymbol{\omega}_1) - y_i\right)^2 + \lambda \sum_i r_f^2(x_i|\boldsymbol{\omega}_2) + \lambda' \int_X \mathrm{d}x \, \exp\{-c\, r_f(x|\boldsymbol{\omega}_2)\}$$

NOMU in Practice



(taken from Heiss et al., *NOMU: Neural Optimization-based Model Uncertainty*, ICML 2022)

NOMU in Practice

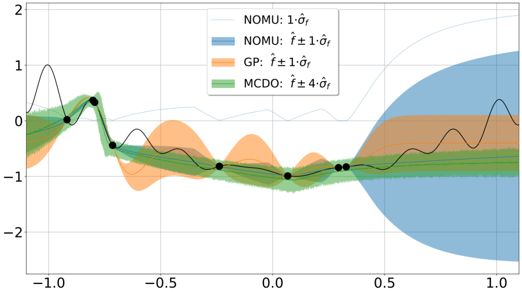


$$\arg\min_{\boldsymbol{\omega}_1,\boldsymbol{\omega}_2} \sum_i \left(f(x_i|\boldsymbol{\omega}_1) - y_i\right)^2 + \lambda \sum_i r_f^2(x_i|\boldsymbol{\omega}_2) + \lambda' \int_X \mathrm{d}x\ \exp\{-c\, r_f(x|\boldsymbol{\omega}_2)\}$$

NOMU in Practice

Do 1. – 3. describe a density estimation?

1. **Non-Negativity**
2. **In-Sample**
3. **Out-Sample**
4. Metric Learning
5. Vanishing



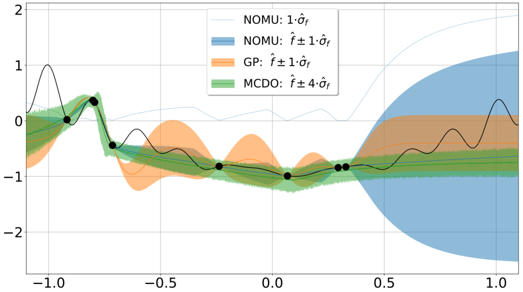
Proposed architecture:

$$f(x) = (f^m \circ \cdots \circ f^1)(x) \qquad r_f(x) = r_f^n\Big(f^{m-1}(x),\, (r_f^{n-1} \circ \cdots \circ r_f^1)(x)\Big)$$

NOMU in Practice

Do 1. – 3. describe a density estimation (in *latent* space)?

1. **Non-Negativity**
2. **In-Sample**
3. **Out-Sample**
4. Metric Learning
5. Vanishing



Claim: Better than GP because r_f incorporates model information from f^{m-1}

Natural Posterior Network (NatPN)

Natural Posterior Network: Deep Bayesian Uncertainty for Exponential Family Distributions [PDF] arxiv.org

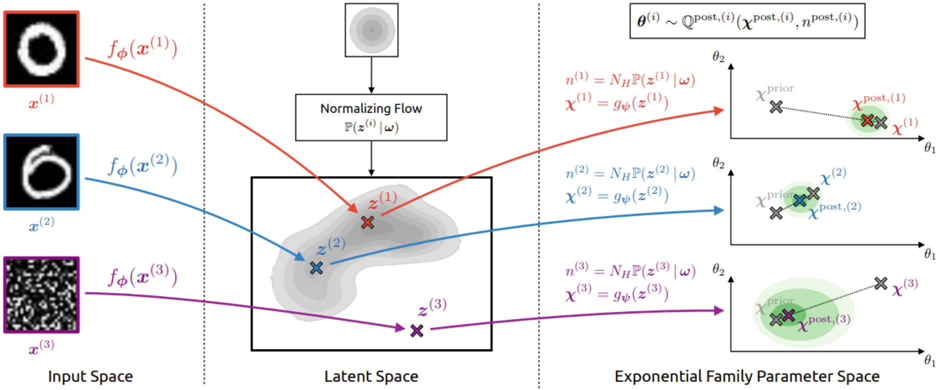
B Charpentier, O Borchert, D Zügner, S Geisler… - arXiv preprint arXiv …, 2021 - arxiv.org

… In this work, we propose the Natural Posterior Network (**NatPN**… , **NatPN** finds application for both classification and general regression settings. Unlike many previous approaches, **NatPN** …

☆ Save 99 Cite Cited by 33 Related articles All 5 versions

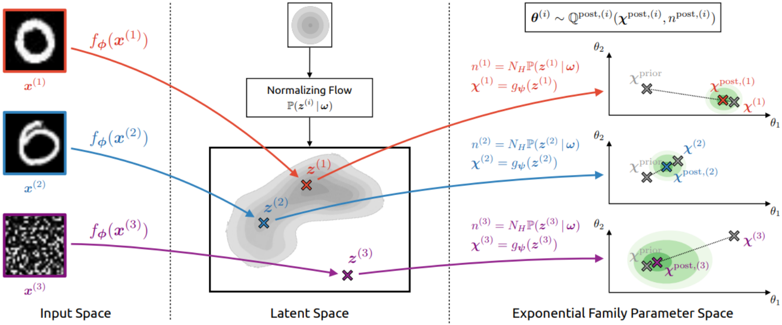
arXiv:2105.04471 (ICLR 2022)

NatPN



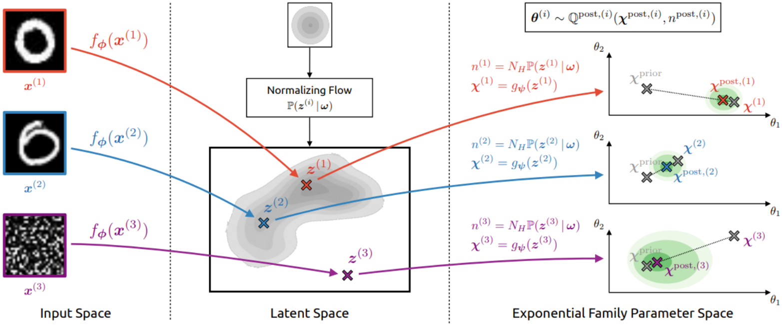
(taken from Charpentier et al., *Natural Posterior Network: Deep Bayesian Uncertainty for Exponential Family Distributions*, ICLR 2022)

NatPN



$$\underbrace{\mathcal{L}(y_i,\, \mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}|\boldsymbol{m}_i)}[p(y|\boldsymbol{\theta})])}_{\text{DER}} \quad \rightarrow \quad \underbrace{\mathbb{E}_{\boldsymbol{\theta}\sim p(\boldsymbol{\theta}|\boldsymbol{m}_i)}[\mathcal{L}(y_i, p(y|\boldsymbol{\theta}))]}_{\text{NatPN}}$$

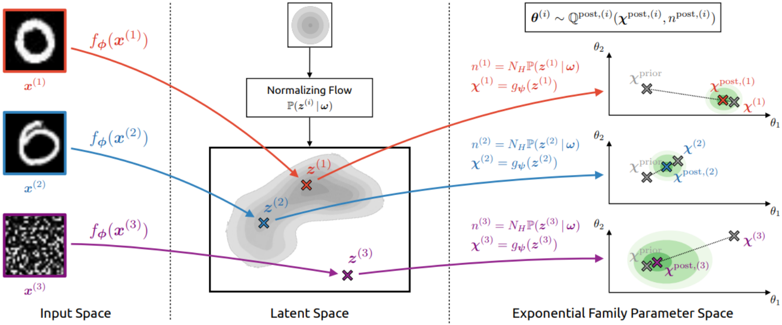
NatPN



$$\underbrace{\boldsymbol{\chi}_i^{\text{post}} = \frac{n^{\text{prior}}\boldsymbol{\chi}^{\text{prior}} + n_i \boldsymbol{\chi}_i}{n^{\text{prior}} + n_i}}_{\text{aleatoric}}$$

$$\underbrace{n_i^{\text{post}} = n^{\text{prior}} + n_i}_{\text{epistemic}}$$

NatPN



n_i comes from a Normalizing Flow $\rightarrow$ **density estimation in latent space**

NatPN: Normal distribution

$$\mathcal{L}_i(\boldsymbol{\omega}) = \underbrace{-\frac{1}{2}\left(-\frac{n_i}{2\beta_i}(y_i-\gamma_i)^2 - \frac{1}{n_i} + \psi\left(\frac{n_i}{2}\right) - \log\beta_i\right)}_{\mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{Q}_i^{\mathrm{post}}}[\log\mathbb{P}(y_i|\boldsymbol{\theta})]}$$

$$- \lambda\underbrace{\frac{1}{2}\left(3\log\beta_i + 2\log\Gamma\left(\frac{n_i}{2}\right) - \log n_i + n_i - (n_i+3)\,\psi\left(\frac{n_i}{2}\right)\right)}_{\mathbb{H}\left[\mathbb{Q}_i^{\mathrm{post}}\right]}$$

***Peculiarities*:**

- $\mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{Q}_i^{\mathrm{post}}} \to \delta(y_i-\gamma_i)$, hence λ sets epistemic uncertainty budget (see `arXiv:2402.09056`)

NatPN: Normal distribution

$$\mathcal{L}_i(\boldsymbol{\omega}) = -\underbrace{\frac{1}{2}\left(-\frac{n_i}{2\beta_i}(y_i-\gamma_i)^2 - \frac{1}{n_i} + \psi\left(\frac{n_i}{2}\right) - \log\beta_i\right)}_{\mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{Q}_i^{\text{post}}}[\log\mathbb{P}(y_i|\boldsymbol{\theta})]}$$

$$-\lambda\underbrace{\frac{1}{2}\left(3\log\beta_i + 2\log\Gamma\left(\frac{n_i}{2}\right) - \log n_i + n_i - (n_i+3)\,\psi\left(\frac{n_i}{2}\right)\right)}_{\mathbb{H}\left[\mathbb{Q}_i^{\text{post}}\right]}$$

***Peculiarities*:**

- $\partial_n\mathcal{L}_i$ are propagated (does this break normalizing flow?)

NatPN: Normal distribution

$$\mathcal{L}_i(\boldsymbol{\omega}) = -\underbrace{\frac{1}{2}\left(-\frac{n_i}{2\beta_i}(y_i-\gamma_i)^2 - \frac{1}{n_i} + \psi\left(\frac{n_i}{2}\right) - \log\beta_i\right)}_{\mathbb{E}_{\boldsymbol{\theta}\sim\mathbb{Q}_i^{\text{post}}}[\log\mathbb{P}(y_i|\boldsymbol{\theta})]}$$

$$-\lambda\underbrace{\frac{1}{2}\left(3\log\beta_i + 2\log\Gamma\left(\frac{n_i}{2}\right) - \log n_i + n_i - (n_i+3)\,\psi\left(\frac{n_i}{2}\right)\right)}_{\mathbb{H}[\mathbb{Q}_i^{\text{post}}]}$$

Peculiarities:

- $\partial_\beta\mathcal{L}_i \propto -\beta_i^{-2} - \lambda'\beta_i^{-1}$ does not depend on data and induces $\beta_i \to \infty$ for $\lambda \geq 1/3$

Summary

- Don’t blindly trust magic loss functions / Don’t reinvent the wheel
- Do we really need **absolute** epistemic uncertainty estimations for downstream tasks in practice?
- . . .or should we threshold a density estimation (where?) for OOD?

Imprint

Topic:	**Your Epistemic Uncertainty is Secretly a Density Estimation and You Should Treat it Like One** Workshop on Uncertainty in Machine Learning (Munich, Germany)
Date:	February 2024
Author:	Nis Meinert
Institute:	Institute of Communications and Navigation
Credits:	All images „DLR (CC BY-NC-ND 3.0)“