universität innsbruck AppliedMathematics

Bayesian Learning & Uncertainty Quantification in Computer Vision

SE Research Seminar: Applied Mathematics

Christoph Angermann

https://applied-math.uibk.ac.at/

I. MLE/MAP vs. Bayesian Inference

II. Epistemic Uncertainty

III. Aleatoric Uncertainty

IV. Results

Point Estimators

- Let $\mathcal{X} = \{x_1, \ldots, x_N\}$ and $\mathcal{Y} = \{y_1, \ldots, y_N\}$ denote observed data and corresponding response, respectively.
- Let $\theta \in \Theta$ denote the model parameters (deterministic quantities).
- **Maximum Likelihood Estimation (MLE)**

$$\hat{\theta} = \arg\max_{\theta}\{p(\mathcal{Y} \mid \mathcal{X}, \theta)\}$$

- **Maximum A Posteriori Estimation (MAP)**

$$\hat{\theta} = \arg\max_{\theta}\{p(\theta \mid \mathcal{X}, \mathcal{Y})\} = \arg\max_{\theta}\{p(\mathcal{Y} \mid \mathcal{X}, \theta) \cdot p(\theta)\}$$

Bayesian View

- Parameters are described in a probabilistic way.
- Model parameter θ is considered a RV, following the posterior probability distribution $p(\theta \mid \mathcal{X}, \mathcal{Y})$.
- Bayesian inference returns a probability density on model parameters $\Rightarrow$ implicit regularization, uncertainty estimates and robustness through model averaging.
- Instead of using the best fitting model $\Rightarrow$ obtain a predictive distribution by using different parameter settings that have significant posterior probability .

Uncertainty Quantification

- Two main types of uncertainty one can model
- **Aleatoric uncertainty** - captures noise inherent in the observations; categorized into homoscedastic and heteroscedastic uncertainty
- **Epistemic uncertainty** - uncertainty in the model parameters (systematic/model uncertainty); can be "explained away" given enough data

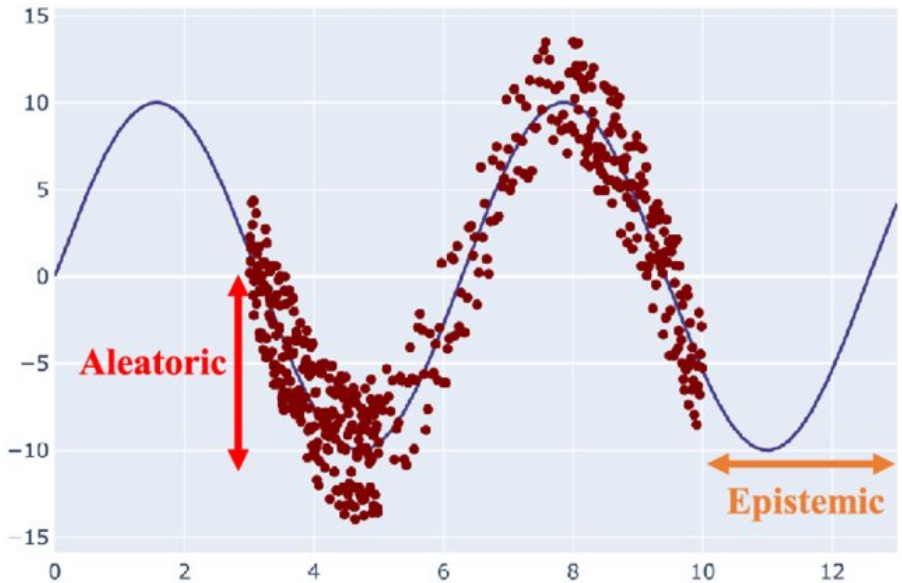

Figure: M. Abdar et al. “A review of uncertainty quantification in deep learning: Techniques, applications and challenges”. In: *Information Fusion* 76 (2021), pp. 243–297

Epistemic Uncertainty

- Let $f_\theta(\cdot) = W_L \sigma_L(W_{L-1}\sigma_{L-1}(\dots W_1(\cdot)))$ denote a Bayesian neural network with parameter $\theta = [W_1, \dots, W_L]$ which is a set of RVs.
- Bayes theorem:

$$p(\theta \mid \mathcal{X}, \mathcal{Y}) = \frac{p(\mathcal{Y} \mid \mathcal{X}, \theta) \cdot p(\theta)}{p(\mathcal{Y} \mid \mathcal{X})} \tag{1}$$

- The model evidence (denominator) in (1) is intractable, which makes analytical inference hardly possible.
- **Variational inference** - a family Q of parameterized distributions is searched to locate a distribution which is complex enough to approximate $p(\theta \mid \mathcal{X}, \mathcal{Y})$ and still is tractable.
- The inference problem is now an optimization problem:

$$q^*_\phi(\theta) = \arg\min_{q_\phi \in Q} \mathrm{KL}(q_\phi(\theta) \,||\, p(\theta \mid \mathcal{X}, \mathcal{Y})).$$

Epistemic Uncertainty

- Minimizing the KL divergence is equivalent to maximizing the evidence lower bound (ELBO), which yields the following loss functional for the parameter ϕ of the approximating distribution $q_\phi(\theta)$:

$$L_{VI}(\phi) = -\sum_{i=1}^{N} \int q_\phi(\theta) \log p(y_i \mid f_\theta(x_i)) d\theta + \mathrm{KL}(q_\phi(\theta) \,||\, p(\theta)). \quad (2)$$

- The expectation over the likelihood function is approximated via Monte Carlo integration, i.e., $q_\phi(\theta)$ is replaced by stochastic samples $\hat{\theta}$.
- Likelihood-term in (2) becomes independent of variational parameter ϕ during optimization $\Rightarrow$ re-parametrization trick:

$$\theta = g(\phi, \epsilon) \Rightarrow \hat{\theta} = g(\phi, \hat{\epsilon}) \quad \text{for a stochastic quantity } \epsilon.$$

- Example: **Monte Carlo dropout inference**

Monte Carlo dropout inference

- Recall: $f_\theta(\cdot) = W_L \sigma_L(W_{L-1}\sigma_{L-1}(\dots W_1(\cdot)))$ with $\theta = [W_1, \dots, W_L]$ a set of RVs.
- Let $\phi = [M_1, \dots, M_L]$ be a set of deterministic weights and $W_l = \text{diag}(\epsilon_l) \cdot M_l$ for $l = 1, \dots, L$ and $\epsilon_l \sim$ Bernoulli.
- Equivalent to applying Dropout during train and test phase.
- Transforming the stochasticity from the weight parameters to $\boldsymbol{\epsilon} = [\epsilon_1, \dots, \epsilon_L]$ yields

$$\theta = [\text{diag}(\epsilon_1) \cdot M_1, \dots, \text{diag}(\epsilon_L) \cdot M_L] = g(\phi, \boldsymbol{\epsilon})$$

and

$$L_{VI}(\phi) = -\sum_{i=1}^{N} \log p(y_i \mid f_{g(\phi,\hat{\epsilon})}(x_i)) + \text{KL}(q_\phi(\theta) \,||\, p(\theta)).$$

- Final prediction for test sample x^* and realizations $\boldsymbol{\epsilon}^1, \dots, \boldsymbol{\epsilon}^T$:

  Predictive mean: $\frac{1}{T}\sum_{t=1}^{T} f_{g(\phi,\boldsymbol{\epsilon}^t)}(x^*)$

  Predictive var.: $\frac{1}{T}\sum_{t=1}^{T} f_{g(\phi,\boldsymbol{\epsilon}^t)}(x^*)^2 - \left(\frac{1}{T}\sum_{t=1}^{T} f_{g(\phi,\boldsymbol{\epsilon}^t)}(x^*)\right)^2$

Heteroscedastic Aleatoric Uncertainty

- Consider the negative log-likelihood objective

$$L_{\text{MLE}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \log p(y_i \mid f_\theta(x_i)).$$

- In regression, a Gaussian likelihood with the observations noise parameter σ can be assumed:

$$L_{\text{MLE}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{1}{2\sigma^2} \left\| y_i - f_\theta(x_i) \right\|^2 \right] + \frac{1}{2} \log \sigma^2.$$

- Due to heteroscedasticity, observation noise σ vary with input x_i and therefore is learned as a function of the data:

$$L_{\text{alea}}(\theta, \rho) = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{1}{2f_\rho(x_i)^2} \left\| y_i - f_\theta(x_i) \right\|^2 + \frac{1}{2} \log f_\rho(x_i)^2 \right].$$

- Neural networks f_θ and f_ρ nearly share all the parameters.
- Learned loss attenuation makes the model more robust to noisy data.

Combining Aleatoric & Epistemic Uncertainty

- **Goal**: infer the predictive distribution $p(y^* \mid x^*, \theta)$ for a Bayesian neural network (BNN) $f_\theta : \mathcal{X} \to \mathcal{Y}$ that gives also a measure of aleatoric uncertainty.
- Let $[\hat{y}_i, \hat{s}_i] = f_{g(\phi, \epsilon)}(x_i)$ and the entries of ϵ follow a Bernoulli distribution. Then:

$$L_{\text{BNN}}(\phi) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{2} \exp(-\hat{s}_i) \, \|y_i - \hat{y}_i\| + \frac{1}{2} \hat{s}_i.$$

- Final prediction for input x^* and T sampled outputs $\hat{y}_1^*, \ldots, \hat{y}_T^*$:

  Predictive mean: $\frac{1}{T} \sum_{t=1}^{T} \hat{y}_t^*$

  Predictive var.: $\frac{1}{T} \sum_{t=1}^{T} (\hat{y}_t^*)^2 - \left(\frac{1}{T} \sum_{t=1}^{T} \hat{y}_t^* \right)^2 + \frac{1}{T} \sum_{t=1}^{T} \exp(\hat{s}_t)$

Depth Regression - Qualitative Results

- Aleatoric uncertainty (AU) is greater for larger depths, reflective surfaces and occlusion boundaries.
- Larger epistemic uncertainty (EU) for objects which are rare in the training set (e.g. humans).

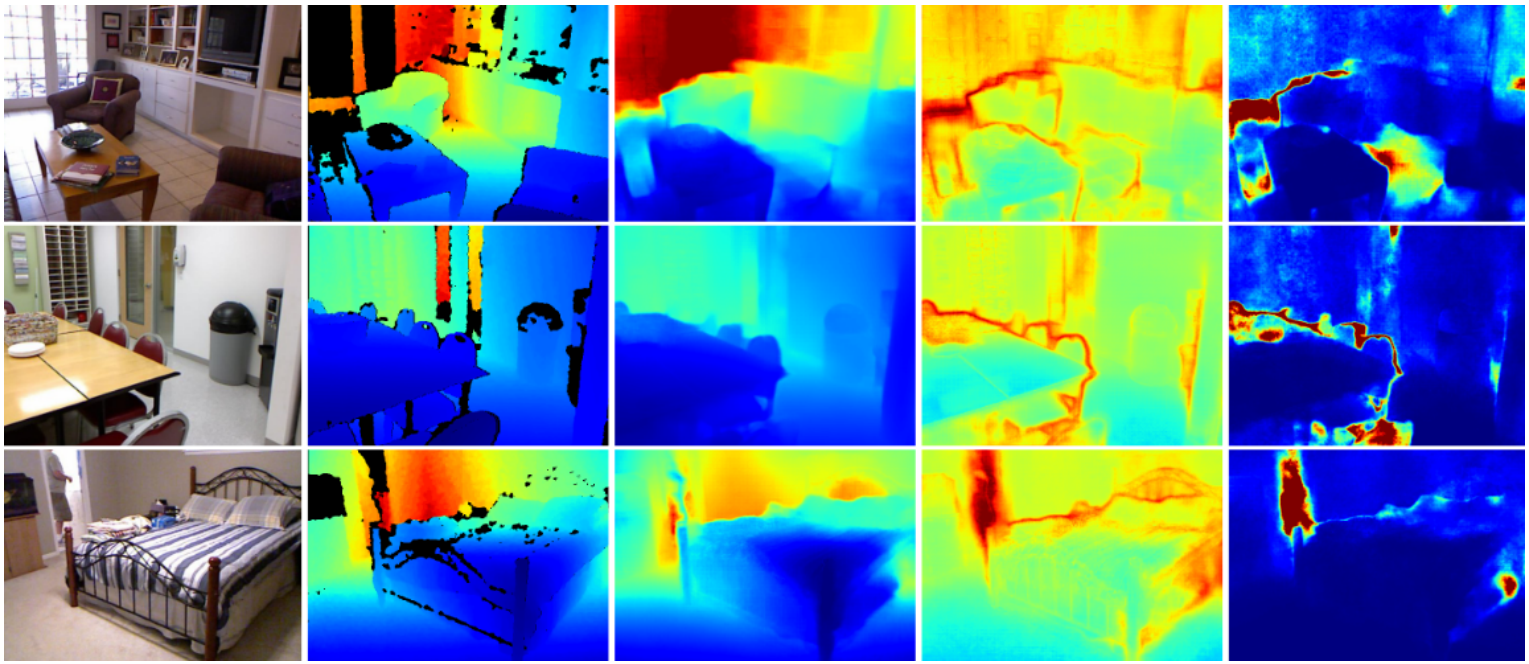

Figure: From left to right: input image, ground truth, depth regression, aleatoric uncertainty and epistemic uncertainty.

A. Kendall and Y. Gal. “What uncertainties do we need in bayesian deep learning for computer vision?” In: *Advances in neural information processing systems* 30 (2017)

Depth Regression - Quantitative Results

- RMSE decreases by removing pixels with uncertainty larger than percentile thresholds $\Rightarrow$ good correlation between model performance and uncertainty measurements.
- Curves for EU and AU models are quite similar $\Rightarrow$ each uncertainty ranks pixel confidence similarly to the other uncertainty (in the absence of the other uncertainty).

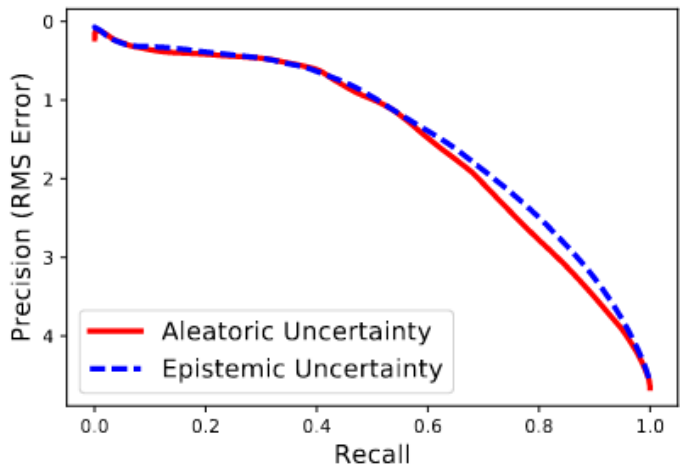

Figure: Accuracy (RMSE) vs. complementary percentage of removed pixels.

A. Kendall and Y. Gal. "What uncertainties do we need in bayesian deep learning for computer vision?" In: *Advances in neural information processing systems* 30 (2017)

Depth Regression - Quantitative Results

Train dataset	Test dataset	RMS	Aleatoric variance	Epistemic variance
Make3D / 4	Make3D	5.76	0.506	7.73
Make3D / 2	Make3D	4.62	0.521	4.38
Make3D	Make3D	3.87	0.485	2.78
Make3D / 4	NYUv2	-	0.388	15.0
Make3D	NYUv2	-	0.461	4.87

Figure: A. Kendall and Y. Gal. "What uncertainties do we need in bayesian deep learning for computer vision?" In: *Advances in neural information processing systems* 30 (2017)

- AU cannot be explained away with more data.
- AU does not increase for out-of-data examples, whereas EU does.
- AU for: large data situations, real-time applications.
- EU for: safety-critical applications, small datasets.

Thank you for your attention!

Christoph Angermann

https://applied-math.uibk.ac.at/