



UPPSALA  
UNIVERSITET

# Towards Accurate and Reliable Deep Regression Models

---

Fredrik K. Gustafsson

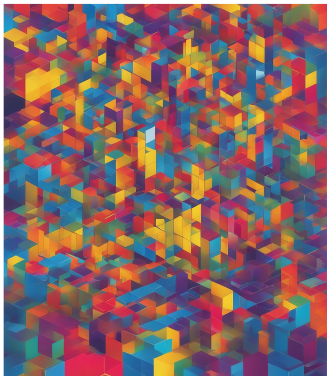
Uppsala University

[www.fregu856.com](http://www.fregu856.com)

PhD Defense

November 30, 2023

Towards Accurate and Reliable  
Deep Regression Models



Fredrik K. Gustafsson



Supervised machine learning problems.

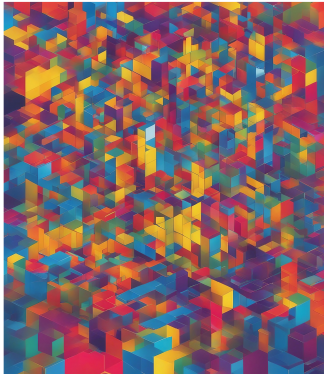
“Deep” parametric models, i.e. deep neural networks (DNNs).

Regression (not classification).

- Predict *continuous* targets  $y \in \mathcal{Y} = \mathbb{R}^K$  for given inputs  $x \in \mathcal{X}$ .
- Simple examples: Predict house prices, power consumption, arrival times, product sales, 3D object positions, volume/area measurements.

Most studied applications are taken from the computer vision domain.

Towards Accurate and Reliable  
Deep Regression Models



Fredrik K. Gustafsson



8 included papers, divided into two different tracks.  
Each track constitutes one main contribution.

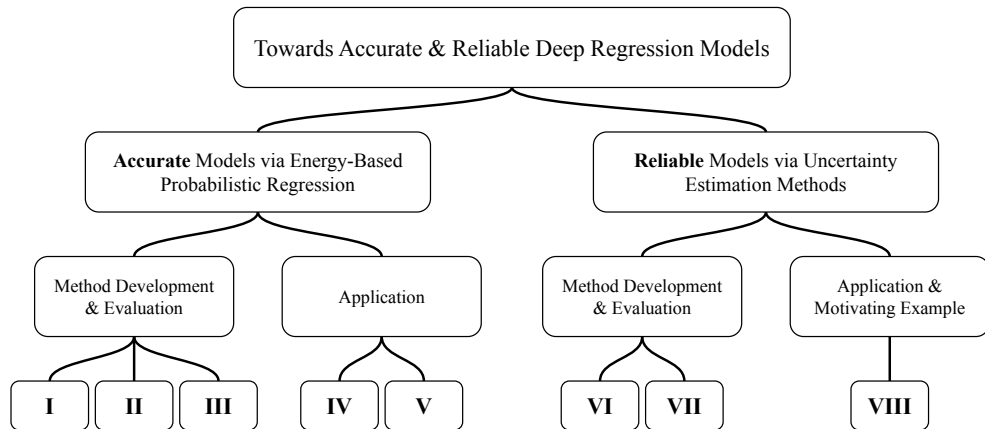
Overarching ultimate goal: Develop deep regression models which are *accurate* and *reliable* enough for real-world deployment within safety-critical domains.

Main contribution 1: Formulation and development of energy-based probabilistic regression.

- Paper I, II & III.

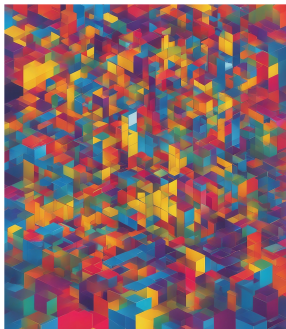
Main contribution 2: A critical evaluation of various uncertainty estimation methods.

- Paper VI & VII.





Towards Accurate and Reliable  
Deep Regression Models



Fredrik K. Gustafsson



Although only my name is on the cover, writing the 8 included papers was of course a big collaborative effort.

Supervisors: *Thomas B. Schön* and *Martin Danelljan*.

Other co-authors: *Goutam Bhat*, *Radu Timofte*, *Johannes Hendriks*, *Antônio H. Ribeiro*, *Adrian Wills*, *Philipp Von Bachmann*, *Daniel Gedon*, *Erik Lampa*, *Stefan Gustafsson* and *Johan Sundström*.

A number of anonymous reviewers have also provided (mostly) constructive feedback on different versions of all the included papers.

Thesis Overview

General Setting

Deep Regression Approaches

Track 1: Energy-Based Probabilistic Regression

Track 2: Uncertainty Estimation Methods

Conclusion

Thesis Overview

General Setting

Deep Regression Approaches

Track 1: Energy-Based Probabilistic Regression

Track 2: Uncertainty Estimation Methods

Conclusion

Supervised machine learning problems:

1. Collect examples of how an input  $x$  relates to some target  $y$ .
2. Fit a model to the collected data  $\{(x_i, y_i)\}_{i=1}^N$ .
3. Use this model to output predicted targets for other, previously unseen inputs  $x$ .

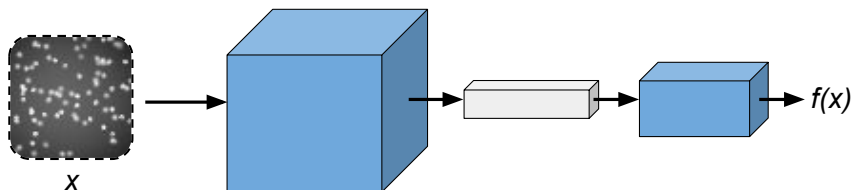
In a **supervised regression problem**, the task is to predict a *continuous* target value  $y^* \in \mathcal{Y} = \mathbb{R}^K$  for any given input  $x^* \in \mathcal{X}$ . To solve this, we are also given a training set of i.i.d. input-target pairs,  $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ ,  $(x_i, y_i) \sim p(x, y)$ .

The focus is often on the 1D case, i.e. when  $\mathcal{Y} = \mathbb{R}$ .

The input space  $\mathcal{X}$  typically corresponds to the space of images.

I view a **Deep Neural Network (DNN)** simply as a function  $f_\theta : \mathcal{X} \rightarrow \mathcal{O}$ , parameterized by  $\theta \in \mathbb{R}^P$ . This function maps inputs  $x \in \mathcal{X}$  to outputs  $f_\theta(x) \in \mathcal{O}$  in some output space  $\mathcal{O}$ .

I also divide the DNN  $f_\theta$  into a *backbone feature extractor*, and one or more smaller *network heads*. The feature extractor takes  $x$  as input and outputs a feature vector  $g(x)$ , which is then fed into the network heads, producing the final output  $f_\theta(x) \in \mathcal{O}$ .



Thesis Overview

General Setting

Deep Regression Approaches

Track 1: Energy-Based Probabilistic Regression

Track 2: Uncertainty Estimation Methods

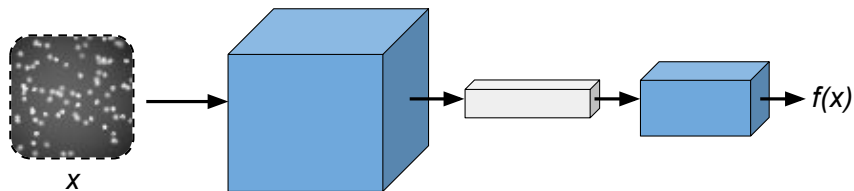
Conclusion

Regression is a fundamental machine learning task, but remains somewhat understudied compared to classification.

While classification problems generally are addressed using standardized target representations and loss functions, these are not directly applicable to regression.

Therefore, a wide variety of regression approaches have been explored. There is no broad consensus on how to construct deep regression models for best possible accuracy, or how to represent and estimate the uncertainty in their predictions.

The most common and straightforward approach, called *direct regression*, is to let the DNN  $f_\theta$  directly output predicted targets,  $\hat{y}(x) = f_\theta(x)$ .



The DNN  $f_\theta$  is trained by minimizing e.g. the L2 loss over the training data,

$$J(\theta) = \sum_{i=1}^N (y_i - f_\theta(x_i))^2.$$



From a probabilistic perspective, the choice of loss function corresponds to minimizing the negative log-likelihood  $\mathcal{L}(\theta) = \sum_{i=1}^N -\log p(y_i|x_i; \theta)$  for a specific model  $p(y|x; \theta)$  of the conditional target distribution  $p(y|x)$ .

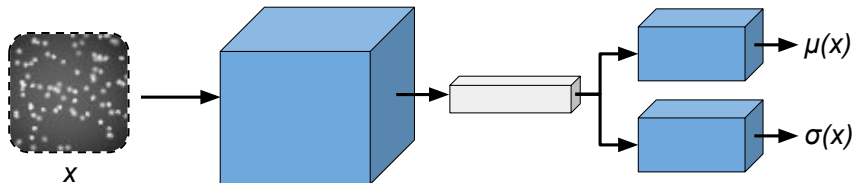
For example, the L2 loss  $\ell(f_\theta(x_i), y_i) = (y_i - f_\theta(x_i))^2$  is derived from a fixed-variance Gaussian model,  $p(y|x; \theta) = \mathcal{N}(y; f_\theta(x), \sigma^2 I)$ . Similarly, the L1 loss can be derived from a fixed-variance Laplace distribution.

By using the L2 or L1 loss, one is thus implicitly using quite restrictive models  $p(y|x; \theta)$ , which might fail to accurately represent the true target distribution  $p(y|x)$ .

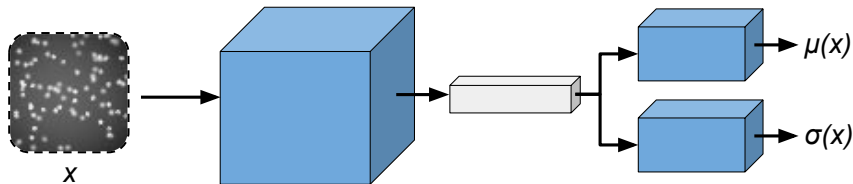
This probabilistic perspective can be extended to define a general regression approach:

**Probabilistic Regression:** Use a DNN  $f_\theta : \mathcal{X} \rightarrow \mathcal{O}$  to specify a model  $p(y|x; \theta)$  of the conditional target distribution, and minimize the corresponding negative log-likelihood (NLL)  $\mathcal{L}(\theta) = \sum_{i=1}^N -\log p(y_i|x_i; \theta)$  in order to train the DNN.

A general 1D Gaussian model can be realized as  $p(y|x; \theta) = \mathcal{N}(y; \mu_\theta(x), \sigma_\theta^2(x))$ , where the DNN outputs both a mean  $\mu_\theta(x)$  and variance  $\sigma_\theta^2(x)$  for each input  $x$ .



**Probabilistic Regression:** Use a DNN  $f_\theta : \mathcal{X} \rightarrow \mathcal{O}$  to specify a model  $p(y|x; \theta)$  of the conditional target distribution, and minimize the corresponding negative log-likelihood (NLL)  $\mathcal{L}(\theta) = \sum_{i=1}^N -\log p(y_i|x_i; \theta)$  in order to train the DNN.



For the Gaussian model, minimizing the NLL is equivalent to minimizing the loss,

$$J(\theta) = \sum_{i=1}^N \frac{(y_i - \mu_\theta(x_i))^2}{\sigma_\theta^2(x_i)} + \log \sigma_\theta^2(x_i).$$

**Probabilistic Regression:** Use a DNN  $f_\theta : \mathcal{X} \rightarrow \mathcal{O}$  to specify a model  $p(y|x; \theta)$  of the conditional target distribution, and minimize the corresponding negative log-likelihood (NLL)  $\mathcal{L}(\theta) = \sum_{i=1}^N -\log p(y_i|x_i; \theta)$  in order to train the DNN.

The Gaussian model  $p(y|x; \theta) = \mathcal{N}(y; \mu_\theta(x), \sigma_\theta^2(x))$  is however still quite restrictive, as it is unable to capture e.g. multi-modal or asymmetric true distributions  $p(y|x)$ .

To address this, mixture density networks (MDNs) or conditional VAEs (cVAEs) could potentially be used, creating mixtures of a certain base distribution.

Or, *energy-based models (EBMs)* could be used to specify  $p(y|x; \theta)$ . EBMs are not restricted to the functional form of any specific distribution (e.g. Gaussian) and, in contrast to MDNs and cVAEs, are not limited to distributions which are easy to sample.

Thesis Overview

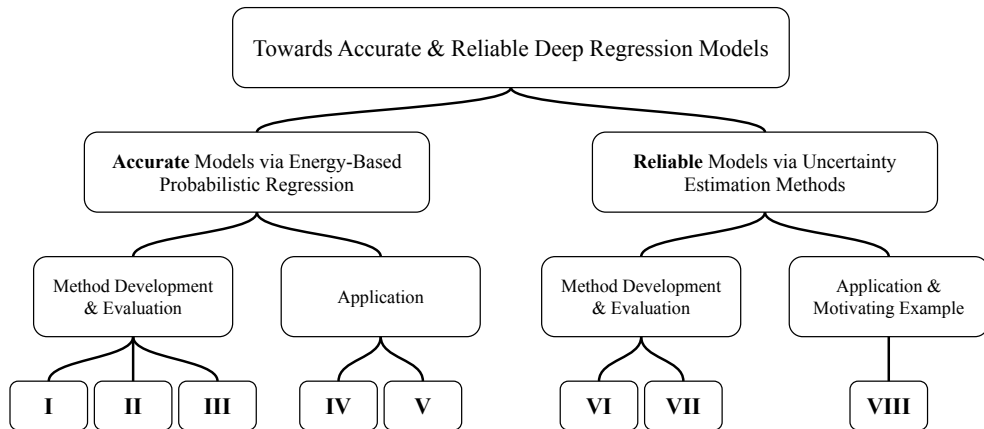
General Setting

Deep Regression Approaches

**Track 1: Energy-Based Probabilistic Regression**

Track 2: Uncertainty Estimation Methods

Conclusion



Track 1: Accurate Deep Regression Models via **Energy-Based Probabilistic Regression**.

The first main contribution of the thesis is the formulation and development of **energy-based probabilistic regression** in Paper I, II & III.

This is a general and conceptually simple regression framework with a clear probabilistic interpretation, using EBMs to model the true conditional target distribution  $p(y|x)$ .

The framework is formulated and initially evaluated in Paper I. A comprehensive study of how the EBMs should be trained for best possible regression performance is then conducted in Paper II, and some practical limitations of the approach are finally addressed in Paper III.

The framework has been applied to a number of regression problems, demonstrating particularly strong performance for 2D bounding box regression – improving the state-of-the-art when applied to the task of visual tracking.

Energy-based models have a rich history within machine learning.

An **energy-based model (EBM)** specifies a probability distribution  $p(x; \theta)$  over  $x \in \mathcal{X}$  directly via a parameterized scalar function  $f_\theta : \mathcal{X} \rightarrow \mathbb{R}$ :

$$p(x; \theta) = \frac{e^{f_\theta(x)}}{Z(\theta)}, \quad Z(\theta) = \int e^{f_\theta(\tilde{x})} d\tilde{x}$$

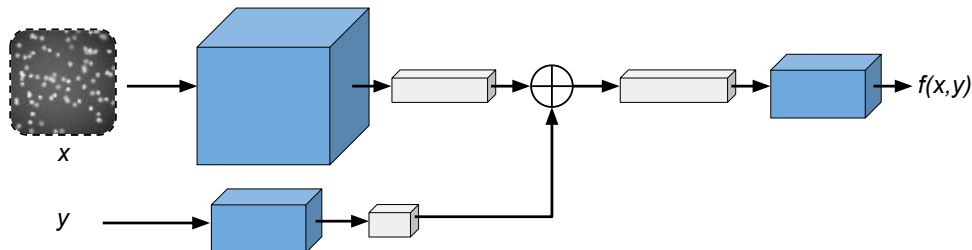
By defining  $f_\theta(x)$  using a DNN, the EBM  $p(x; \theta)$  becomes expressive enough to learn practically any distribution from observed data.

**Drawback:** The normalizing partition function  $Z(\theta) = \int e^{f_\theta(\tilde{x})} d\tilde{x}$  is intractable, which complicates evaluating or sampling from the EBM  $p(x; \theta)$ .



**Energy-Based Probabilistic Regression:** train a DNN  $f_\theta : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$  to predict a scalar  $f_\theta(x, y)$ , then model  $p(y|x)$  with the conditional EBM  $p(y|x; \theta)$ :

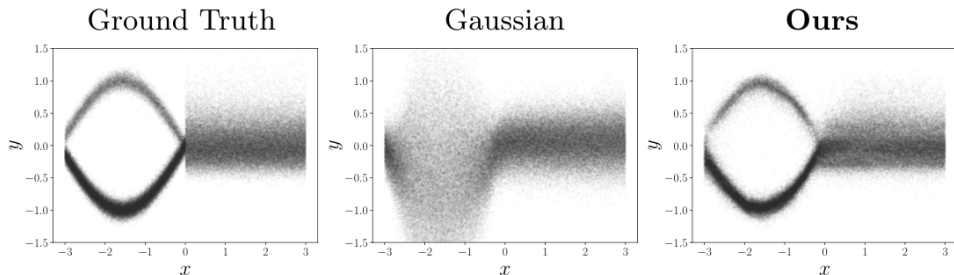
$$p(y|x; \theta) = \frac{e^{f_\theta(x, y)}}{Z(x, \theta)}, \quad Z(x, \theta) = \int e^{f_\theta(x, \tilde{y})} d\tilde{y}.$$



**Energy-Based Probabilistic Regression:** train a DNN  $f_\theta : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}$  to predict a scalar  $f_\theta(x, y)$ , then model  $p(y|x)$  with the conditional EBM  $p(y|x; \theta)$ :

$$p(y|x; \theta) = \frac{e^{f_\theta(x, y)}}{Z(x, \theta)}, \quad Z(x, \theta) = \int e^{f_\theta(x, \tilde{y})} d\tilde{y}.$$

The EBM  $p(y|x; \theta)$  can learn complex distributions  $p(y|x)$  directly from data:



## **I: Energy-Based Models for Deep Probabilistic Regression**

*Fredrik K. Gustafsson, Martin Danelljan, Goutam Bhat, Thomas B. Schön*

The European Conference on Computer Vision (ECCV), 2020

## **II: How to Train Your Energy-Based Model for Regression**

*Fredrik K. Gustafsson, Martin Danelljan, Radu Timofte, Thomas B. Schön*

The British Machine Vision Conference (BMVC), 2020

## **III: Learning Proposals for Practical Energy-Based Regression**

*Fredrik K. Gustafsson, Martin Danelljan, Thomas B. Schön*

The International Conference on Artificial Intelligence and Statistics (AISTATS), 2022

## **IV: Accurate 3D Object Detection using Energy-Based Models**

*Fredrik K. Gustafsson, Martin Danelljan, Thomas B. Schön*

Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops), 2021

## **V: Deep Energy-Based NARX Models**

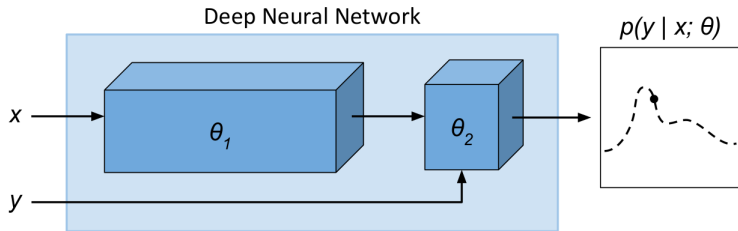
*Johannes Hendriks, Fredrik K. Gustafsson, Antônio H. Ribeiro, Adrian Wills, Thomas B. Schön*

The 19th IFAC Symposium on System Identification (SYSID), 2021

## I: Energy-Based Models for Deep Probabilistic Regression

*Fredrik K. Gustafsson, Martin Danelljan, Goutam Bhat, Thomas B. Schön*

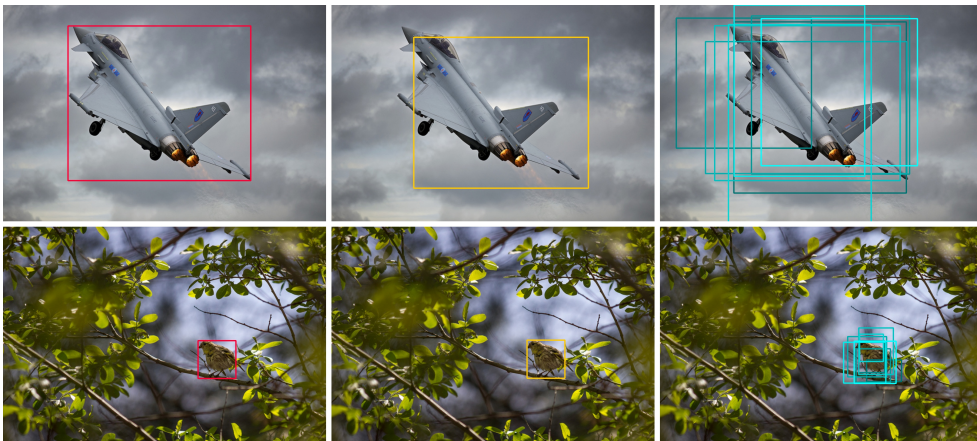
The European Conference on Computer Vision (ECCV), 2020



## II: How to Train Your Energy-Based Model for Regression

*Fredrik K. Gustafsson, Martin Danelljan, Radu Timofte, Thomas B. Schön*

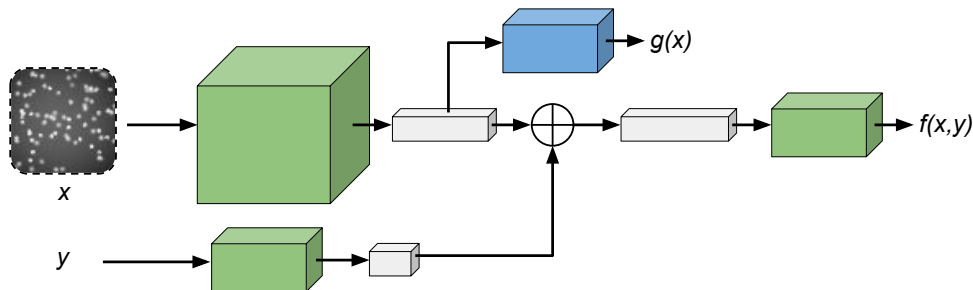
The British Machine Vision Conference (BMVC), 2020



## III: Learning Proposals for Practical Energy-Based Regression

*Fredrik K. Gustafsson, Martin Danelljan, Thomas B. Schön*

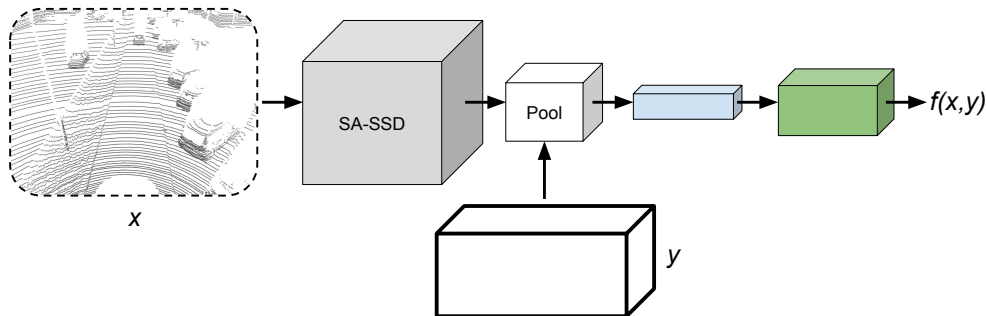
The International Conference on Artificial Intelligence and Statistics (AISTATS), 2022



## IV: Accurate 3D Object Detection using Energy-Based Models

*Fredrik K. Gustafsson, Martin Danelljan, Thomas B. Schön*

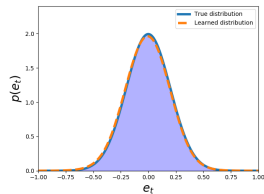
Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops), 2021



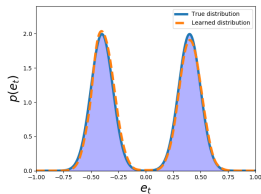
## V: Deep Energy-Based NARX Models

Johannes Hendriks, Fredrik K. Gustafsson, Antônio H. Ribeiro, Adrian Wills, Thomas B. Schön

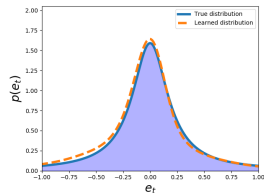
The 19th IFAC Symposium on System Identification (SYSID), 2021



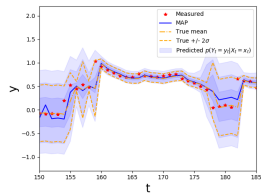
(a) Gaussian.



(b) Bimodal Gaussian.



(c) Cauchy.



(d) Dependent variance  
Gaussian.



Thesis Overview

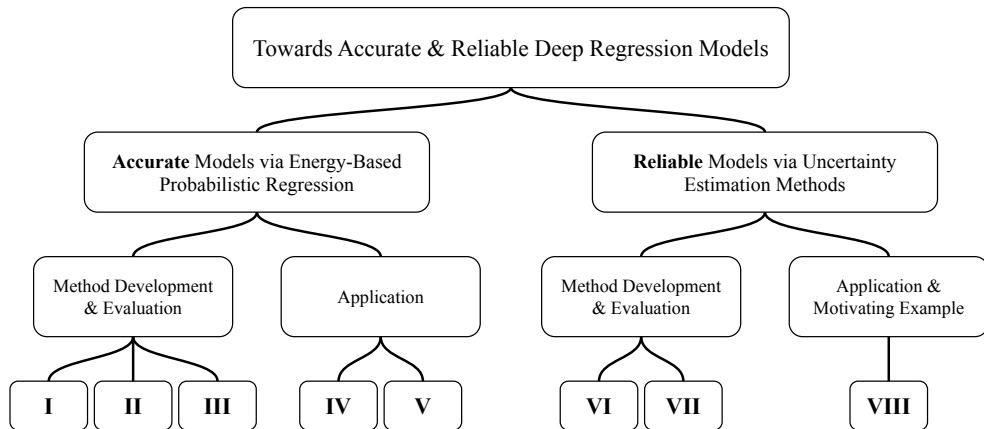
General Setting

Deep Regression Approaches

Track 1: Energy-Based Probabilistic Regression

Track 2: Uncertainty Estimation Methods

Conclusion



Track 2: Reliable Deep Regression Models via **Uncertainty Estimation Methods**.

The second main contribution of the thesis is the **critical evaluation of various uncertainty estimation methods** conducted in Paper VI & VII.

A general introduction to the problem of estimating the predictive uncertainty of deep models is provided in Paper VI, together with an extensive comparison of the two popular methods ensembling and MC-dropout.

In Paper VII, ensembling and other uncertainty estimation methods are then further evaluated, specifically examining their reliability under *real-world distribution shifts*.

This evaluation *uncovers important limitations of current methods* and serves as a challenge to the research community. It demonstrates that more work is required in order to develop truly reliable uncertainty estimation methods for regression.

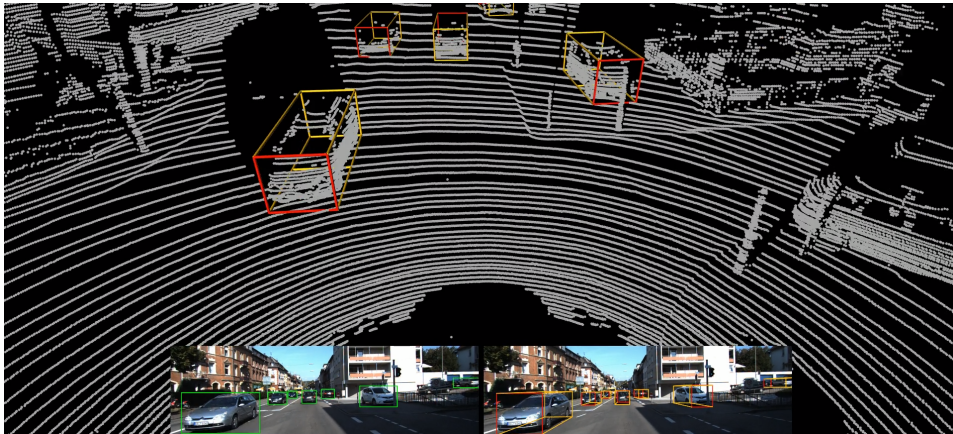
DNNs  $f_\theta : \mathcal{X} \rightarrow \mathcal{O}$  have become the go-to approach within computer vision and many other domains due to their impressive predictive power. However, they generally fail to properly capture the uncertainty inherent in their predictions.

*Bayesian deep learning* is one approach that aims to address this issue in a principled manner. It deals with predictive uncertainty by decomposing it into the distinct types of *aleatoric* and *epistemic* uncertainty.

**Aleatoric** uncertainty captures *inherent* and *irreducible* ambiguity in the inputs  $x$ .

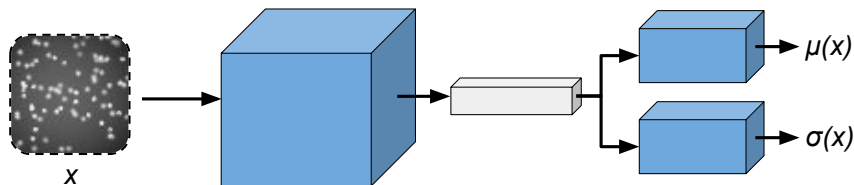
**Epistemic** uncertainty accounts for uncertainty in the DNN model parameters  $\theta$ . More broadly, it is *reducible* uncertainty related to a lack of knowledge (*“uncertainty due to things one could in principle know but does not in practice”*).

Input-dependent aleatoric uncertainty arises whenever the target  $y$  is expected to be inherently more uncertain for some inputs  $x$  than others. This is true e.g. in automotive 3D object detection, where it is inherently more difficult to estimate the 3D position and size of distant or partially occluded vehicles.



To estimate input-dependent aleatoric uncertainty, the DNN  $f_\theta : \mathcal{X} \rightarrow \mathcal{O}$  can be used to specify a model  $p(y|x; \theta)$  of the conditional target distribution.

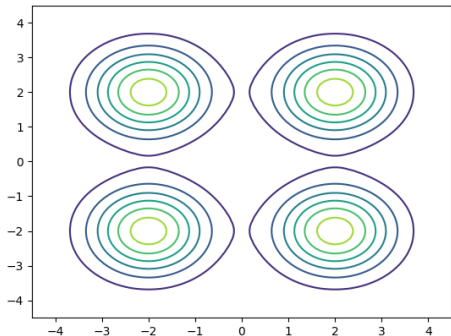
For example, a Gaussian model can be used,  $p(y|x; \theta) = \mathcal{N}(y; \mu_\theta(x), \sigma_\theta^2(x))$ .



The mean can then be taken as a prediction,  $\hat{y}(x) = \mu_\theta(x)$ , whereas the variance  $\sigma_\theta^2(x)$  naturally can be interpreted as a measure of aleatoric uncertainty for this prediction.

Using DNNs to specify models  $p(y|x; \theta)$  of the conditional target distribution does however not capture **epistemic** uncertainty, as information about the uncertainty in the model parameters  $\theta$  is disregarded.

Large epistemic uncertainty is present whenever a large set of model parameters explains the given training data (approximately) equally well.



This is often the case for DNNs, since the corresponding optimization landscapes are highly multi-modal.

Disregarding the epistemic model uncertainty can lead to highly confident yet incorrect predictions, especially for inputs  $x$  which are not well-represented by the training data.

Epistemic uncertainty can be estimated in a principled manner by performing *approximate Bayesian inference*.

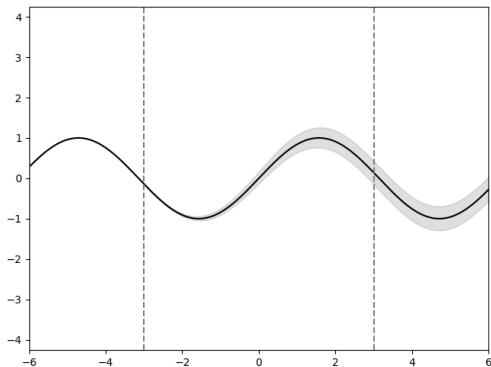
Instead of just finding a single point estimate  $\hat{\theta}$  of the model parameters  $\theta$ , by minimizing the negative log-likelihood  $\mathcal{L}(\theta) = \sum_{i=1}^N -\log p(y_i|x_i; \theta)$  over the training set  $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ , Bayesian inference entails estimating the full *posterior distribution*  $p(\theta|\mathcal{D})$ .

The posterior  $p(\theta|\mathcal{D})$  is obtained from the data likelihood  $\prod_{i=1}^N p(y_i|x_i; \theta)$  and a chosen prior  $p(\theta)$  by applying Bayes' theorem,  $p(\theta|\mathcal{D}) \propto \prod_{i=1}^N p(y_i|x_i; \theta)p(\theta)$ .

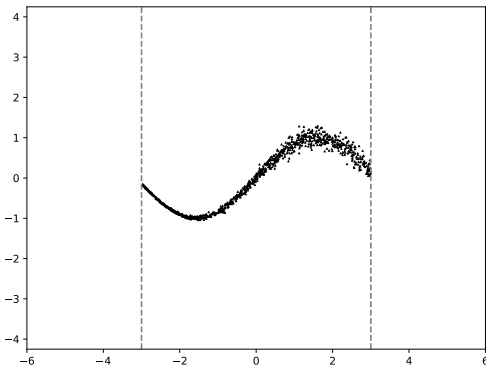


Let us consider the following simple 1D regression problem:

$$p(y|x) = \mathcal{N}(y; \mu(x), \sigma^2(x)), \quad \mu(x) = \sin(x), \quad \sigma(x) = \frac{0.15}{1 + e^{-x}}.$$

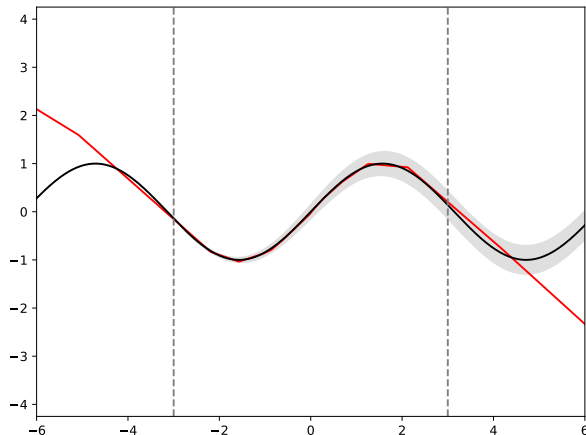


(a) True data generator  $p(y|x)$ .

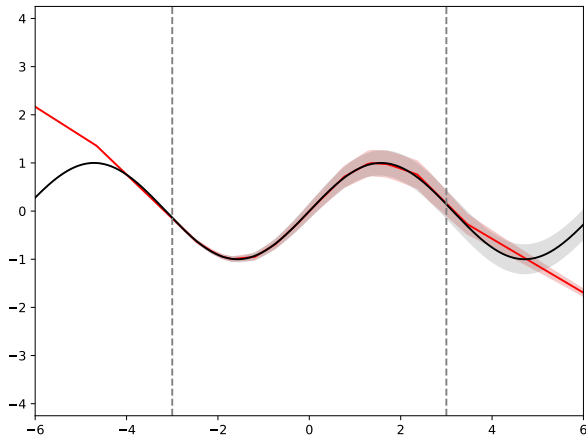


(b) Training dataset  $\{(x_i, y_i)\}_{i=1}^{1000}$ .

A DNN  $f_\theta$  trained to directly output predicted targets,  $\hat{y}(x) = f_\theta(x)$ , is able to accurately regress the mean  $\mu(x) = \sin(x)$  for  $x \in [-3, 3]$ . However, this model fails to capture any notion of uncertainty.

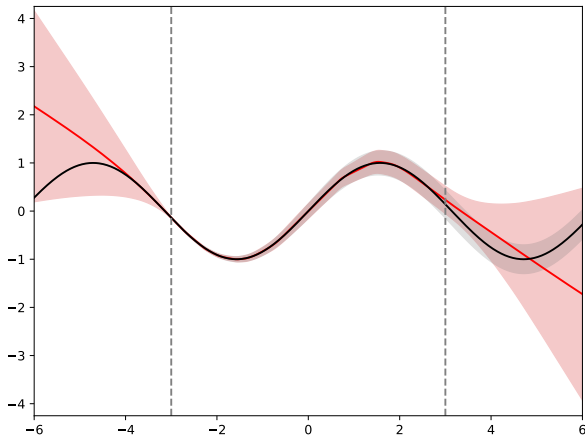


Instead, the DNN  $f_\theta$  can be used to specify a Gaussian model  $p(y|x; \theta) = \mathcal{N}(y; \mu_\theta(x), \sigma_\theta^2(x))$ , trained by minimizing the NLL  $\mathcal{L}(\theta)$ . The model closely matches the true  $p(y|x)$  for  $x \in [-3, 3]$ , accounting for *aleatoric* uncertainty.



For inputs  $|x| > 3$  not seen during training, however, the estimated mean  $\mu_\theta(x)$  deviates significantly from the true  $\mu(x) = \sin(x)$ , while the estimated uncertainty  $\sigma_\theta^2(x)$  remains very small. That is, the model becomes **overconfident** for inputs  $|x| > 3$ .

The Gaussian DNN model  $p(y|x; \theta) = \mathcal{N}(y; \mu_{\theta}(x), \sigma_{\theta}^2(x))$  can instead be estimated via approximate Bayesian inference, in order to account for both *aleatoric* and *epistemic* uncertainty.



The model now predicts a more reasonable uncertainty  $\sigma_{\theta}^2(x)$  in the region with no available training data.

While the estimated mean  $\mu_{\theta}(x)$  still deviates from the true  $\mu(x) = \sin(x)$  for  $|x| > 3$ , the uncertainty  $\sigma_{\theta}^2(x)$  also increases accordingly – the model does *not* become overconfident.

### **VI: Evaluating Scalable Bayesian Deep Learning Methods for Robust Computer Vision**

*Fredrik K. Gustafsson, Martin Danelljan, Thomas B. Schön*

Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops), 2020

### **VII: How Reliable is Your Regression Model's Uncertainty Under Real-World Distribution Shifts?**

*Fredrik K. Gustafsson, Martin Danelljan, Thomas B. Schön*

Transactions on Machine Learning Research (TMLR), 2023

### **VIII: ECG-Based Electrolyte Prediction: Evaluating Regression and Probabilistic Methods**

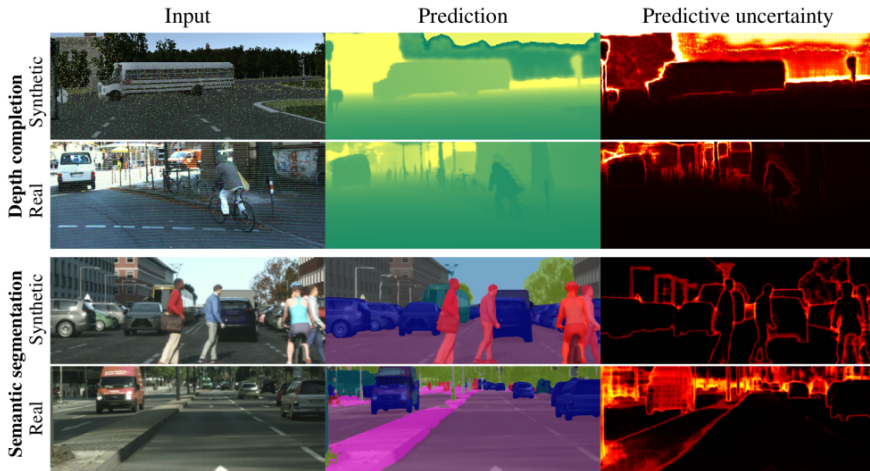
*Philipp Von Bachmann, Daniel Gedon, Fredrik K. Gustafsson, Antônio H. Ribeiro, Erik Lampa, Stefan Gustafsson, Johan Sundström, Thomas B. Schön*

In Preparation, 2023

## VI: Evaluating Scalable Bayesian Deep Learning Methods for Robust Computer Vision

*Fredrik K. Gustafsson, Martin Danelljan, Thomas B. Schön*

















Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops), 2020



## VII: How Reliable is Your Regression Model's Uncertainty Under Real-World Distribution Shifts?

*Fredrik K. Gustafsson, Martin Danelljan, Thomas B. Schön*

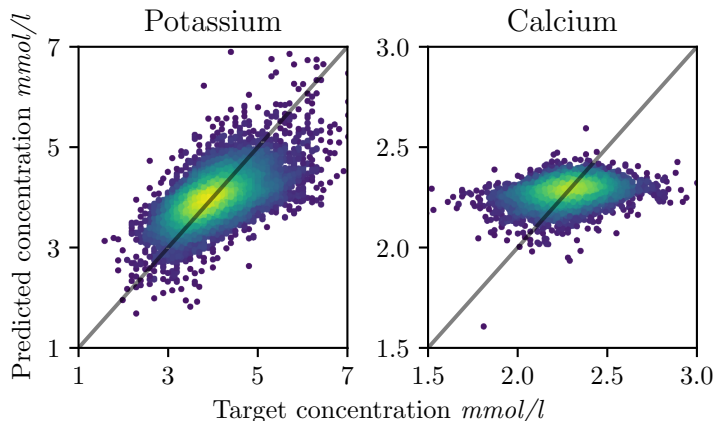
Transactions on Machine Learning Research (TMLR), 2023

	Cells-Tails	ChairAngle-Gap	AssetWealth	Ventricular Volume	BrainTumour Pixels	SkinLesion Pixels	Histology NucleiPixels	AerialBuilding Pixels
Train	 y = 100	 y = 80.5	 y = 1.594	 y = 40.38	 y = 252	 y = 500	 y = 1257	 y = 1097
Test	 y = 198	 y = 43.8	 y = 1.314	 y = 85.93	 y = 273	 y = 516	 y = 1156	 y = 433

### VIII: ECG-Based Electrolyte Prediction: Evaluating Regression and Probabilistic Methods

*Philipp Von Bachmann, Daniel Gedon, Fredrik K. Gustafsson, Antônio H. Ribeiro, Erik Lampa, Stefan Gustafsson, Johan Sundström, Thomas B. Schön*

In Preparation, 2023





Thesis Overview

General Setting

Deep Regression Approaches

Track 1: Energy-Based Probabilistic Regression

Track 2: Uncertainty Estimation Methods

Conclusion

Overarching ultimate goal: Develop deep regression models which are *accurate* and *reliable* enough for real-world deployment within safety-critical domains.

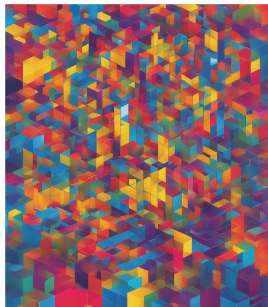
The first main contribution of the thesis is the formulation and development of **energy-based probabilistic regression** in Paper I, II & III.

- The framework is applied to a number of regression problems and demonstrates particularly strong performance for 2D bounding box regression, improving the state-of-the-art when applied to visual tracking.

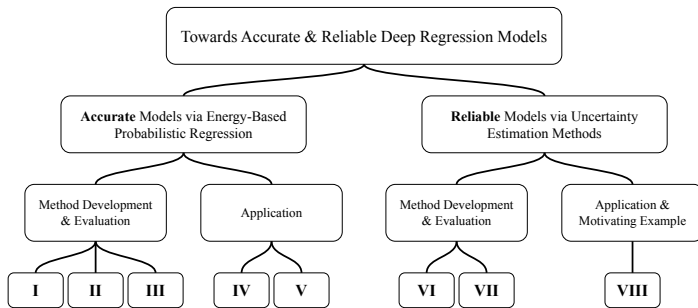
The second main contribution of the thesis is the **critical evaluation of various uncertainty estimation methods** conducted in Paper VI & VII.

- This evaluation *uncovers important limitations of current methods* and serves as a challenge to the research community. It demonstrates that more work is required in order to develop truly reliable uncertainty estimation methods for regression.

## Towards Accurate and Reliable Deep Regression Models



Fredrik K. Gustafsson



Fredrik K. Gustafsson

*fredrik.gustafsson@it.uu.se*

[www.fregu856.com](http://www.fregu856.com)

Please feel free to leave any type of **anonymous** feedback on this presentation:

[www.fregu856.com/post/feedback](http://www.fregu856.com/post/feedback)

