Evidential classification/regression

Contents

1 Parameter estimation **2**

1.1 Bernoulli/Categorical distribution . . . 2

1.2 Gaussian distribution . . . 2

1.3 Bayesian inference . . . 3

1.4 Prior/Posterior predictive . . . 4

1.5 Prior hyperparameters as fictitious evidence . . . 5

1.6 Why the introduction . . . 5

2 Classification/Regression **6**

2.1 Problem description . . . 6

2.2 Bayesian treatment . . . 7

2.3 Evidential treatment . . . 8

2.3.1 Introduction . . . 8

2.3.2 Making predictions . . . 8

2.3.3 Loss function for evidential learning . . . 10

2.4 Summary . . . 11

3 Key points through examples **12**

References **15**

1 Parameter estimation

1.1 Bernoulli/Categorical distribution

Consider an i.i.d dataset $\mathcal{D} = \{y_i\}_{i=1}^N$, where each y_i can take values from $\{1, 0\}$ and is sampled from a Bernoulli distribution with unknown probability μ. This can describe a series of coin tosses and the objective is to estimate the underlying probability of "Heads" occurrence. Let's denote Heads with the outcome 1. The data likelihood for this binary variable experiment is given as

$$p(\mathcal{D}|\mu) = \prod_{i=1}^{N} \mu^{y_i}(1-\mu)^{1-y_i} \tag{1}$$

and the maximum likelihood estimate (MLE) is given as

$$\mu_{MLE} = \frac{N_H}{N} \tag{2}$$

where N_H is the number of Heads in the N samples. The MLE is thus the sample mean and we can use it for making future predictions. However, if for example after 3 tosses we observe 3 Heads, then $\mu_{MLE} = 1$, i.e., it predicts that the next sample will be Heads with probability 1. This is an extreme case of overfitting.

More generally, for describing quantities that can take K possible values we can use a categorical distribution (also referred to as Multinoulli) with an unknown vector of discrete class probabilities $\mu = [\mu_1, \ldots, \mu_K]^T$ which sums to 1; i.e., it belongs to the $K-1$-simplex. The MLE for this set of parameters is given as $\mu_{MLE} = [N_1/N, \ldots, N_K/N]^T$.

In binary classification problems, for each x the class $y(x)$ is typically modeled with a Bernoulli distribution and in multi-class classification problems for each x the class $y(x)$ is typically modeled with a categorical distribution.

1.2 Gaussian distribution

Consider a i.i.d. dataset $\mathcal{D} = \{y_i\}_{i=1}^N$, where each y_i can take values in $\mathbb{R}$ and is sampled from a Gaussian distribution with unknown mean μ and variance σ^2. The data likelihood for this experiment is proportional to

$$p(\mathcal{D}|\mu, \sigma^2) \propto \prod_{i=1}^{N} exp\left(-\frac{(y_i-\mu)^2}{2\sigma^2}\right) \tag{3}$$

and the MLE as

$$\mu_{MLE} = \bar{y} = \frac{1}{N}\sum y_i, \;\; \sigma^2_{MLE} = \frac{1}{N}\sum (y_i - \bar{y})^2 \tag{4}$$

The MLE for σ^2 is biased and is typically corrected by multiplying by $\frac{N}{N-1}$. This is because both μ and σ^2 are estimated from the data.

In regression problems, for each x the variable $y(x)$ is typically modeled with a Gaussian distribution.

1.3 Bayesian inference

MLE gives only point estimates of the parameters of interest and can lead to overfitted results for small datasets. In this regard, the Bayesian framework treats the unknown parameters as random variables with associated probabilities corresponding to degrees of belief. Thus, for a parameter set w (this may refer to the μ vector in categorical or $\{\mu, \sigma^2\}$ in Gaussian distribution), instead of maximizing the likelihood $p(\mathcal{D}|w)$ and obtaining a point estimate w_{MLE}, we introduce a posterior probability distribution $p(w|\mathcal{D})$ which specifies the plausibility of each value for w based on the data. The posterior is given as

$$p(w|\mathcal{D}, \eta) = \frac{p(\mathcal{D}|w)p(w, \eta)}{p(\mathcal{D}, \eta)} \tag{5}$$

where $p(w, \eta)$ is a prior distribution for the parameters and η represents a set of hyperparameters (e.g., the prior variance). Note that the likelihood may have its own hyperparameters as well.

If we obtain the posterior $p(w|\mathcal{D}, \eta)$ we can use it to make future predictions. But how do we obtain the posterior? We observe that the posterior is proportional to the product of the likelihood and the prior, i.e.,

$$p(w|\mathcal{D}, \eta) \propto p(\mathcal{D}|w)p(w, \eta) \tag{6}$$

Therefore, if we can introduce a prior that interacts in an analytically convenient manner with the likelihood then we can obtain the posterior analytically. This is the case for conjugate priors. Such priors interact with the likelihood in such a way so that the posterior and the prior belong to the same probability distribution family.

Interestingly, for any likelihood function belonging to the exponential family (such as the Bernoulli, categorical and Gaussian distributions), there exists a conjugate prior

and we also have its functional form. For example, the conjugate prior for the Bernoulli likelihood is the Beta distribution, for the categorical the Dirichlet distribution and for the Gaussian with unknown mean and variance the normal-inverse-Gamma (NIG) distribution.

As a result, we can obtain analytically the posterior for such cases. For a Bernoulli likelihood the prior is $\text{Beta}(\mu|\alpha,\beta)$ and the posterior is $\text{Beta}(\mu|\alpha+N_H,\beta+N_T)$. Next, for a categorical likelihood the prior is $\text{Dirichlet}(\mu|\alpha)$, where $\alpha=[\alpha_1,\ldots,\alpha_K]^T$, and the posterior is $\text{Dirichlet}(\mu|\alpha+[N_1,\ldots,N_K]^T)$. Finally, for a Gaussian likelihood the prior is $\text{NIG}(\mu,\sigma^2|\gamma,v,\alpha,\beta)$ and the posterior is $\text{NIG}(\mu,\sigma^2|g(\gamma,v,\alpha,\beta,\bar{y},\sum(y_i-\bar{y})^2))$ where g is a known function (see Bishop (2006)). Overall, the posterior is fully specified by the prior and the sufficient statistics of the data.

1.4 Prior/Posterior predictive

After obtaining the posterior we would like to make predictions. Before any data had arrived the predictions could be based only on the prior; this is the prior predictive distribution given as

$$p(y|\eta)=\int p(y|w)p(w|\eta)dw \tag{7}$$

or, equivalently,

$$prior\ predictive(y|\eta)=\int likelihood(y|w)\times prior(w|\eta)dw \tag{8}$$

For example, for the categorical/Dirichlet pair the prior predictive is

$$p(y|\eta)=\int\prod_{j=1}^{K}\mu_j^{y_j}\times Dirichlet(\mu|\eta)d\mu \tag{9}$$

which degenerates to $p(y=\lambda|\eta)=\eta_\lambda/\sum_k\eta_j$. For the Gaussian/NIG pair the prior predictive is

$$p(y|\eta)=\int Gaussian(y|\mu,\sigma^2)\times NIG(\mu,\sigma^2|\eta)d\mu d\sigma^2 \tag{10}$$

which is expressed as a Student-t distribution that depends on η.

After the data has arrived we can incorporate it for making predictions; this is the posterior predictive given as

$$p(y|\eta,\mathcal{D})=\int p(y|w)p(w|\eta,\mathcal{D})dw \tag{11}$$

or, equivalently,

$$posterior\ predictive(y|\eta, \mathcal{D}) = \int likelihood(y|w) \times posterior(w|\eta, \mathcal{D})dw \tag{12}$$

Recall that the prior $p(w|\eta)$ and the posterior $p(w|\eta, \mathcal{D})$ have the same functional form due to conjugacy. Therefore, the analytical closed-form expressions we have for Eq. (7) (Eqs. (9)-(10)) apply to Eq. (11) as well with minimal changes.

Finally, note that the hyperparameters η can be estimated from type-II MLE (also known as empirical Bayes), i.e., by maximizing the marginal likelihood $p(\mathcal{D}|\eta)$. The marginal likelihood is given as

$$p(\mathcal{D}|\eta) = \int p(\mathcal{D}|w)p(w|\eta)dw \tag{13}$$

1.5 Prior hyperparameters as fictitious evidence

Conjugacy has two more implications. First, the posterior predictive can be construed as a prior predictive with better-informed prior hyperparameters and can be used as the prior when subsequent data arrives. In this regard, the posterior hyperparameters are a function of the prior ones and the sufficient statistics of the data; say $\tilde{\eta} = g(\eta, s(\mathcal{D}))$. They incorporate the "evidence" provided by the data $\mathcal{D}$. Second, following the same rationale, the prior hyperparameters η can be construed as "pseudo-evidence", i.e., corresponding to some fictitious observations.

For example, α, β of the Beta prior can be interpreted as fictitious observations of α Heads and β Tails in a coin tossing experiment. Similarly for the Dirichlet prior. For the NIG prior the hyperparameters γ, v can be interpreted as v fictitious observations with sample mean γ and the hyperparameters α, β as 2α observations with sample mean γ and sum of square residuals 2β.

This is a general result for exponential distributions: hyperparameters can be interpreted as corresponding to fictitious observations.

1.6 Why the introduction

The term "evidential" in evidential classification/regression refers to this fictitious evidence corresponding to the prior hyperparameters. To elaborate further, consider the following extreme scenario: suppose we have no data (i.e., no access to a posterior

predictive), but somehow have very good prior hyperparameters η. Then we can use the prior predictive for making predictions with accuracy depending on the number of fictitious observations the hyperparameters η correspond.

In evidential classification/regression the parameters η for each x are produced by a neural network (NN). It is, thus, *as if the NN produced the corresponding fictitious evidence*. Finally, for predicting $y(x)$ the prior predictive is used with hyperparameters given by the NN output $\eta(x)$.

2 Classification/Regression

2.1 Problem description

In the standard binary classification setting the likelihood for the value $y(x)$ for every x can be modeled as a Bernoulli distribution, i.e.,

$$p(y|x, w) = Bernoulli(y|\mu_w(x)) \tag{14}$$

where the vector of probabilities $\mu_w(x)$ (e.g., $[0.8, 0.2]$) can be the output of a NN. The prediction $y^*(x)$ is typically taken as the class corresponding to the maximum probability in $\mu_w(x)$. Similarly for multi-class classification with a categorical instead of a Bernoulli distribution.

In the regression setting the likelihood for $y(x)$ for every x can be modeled as a Gaussian distribution, i.e.,

$$p(y|x, w) = Gaussian(y|\mu_w(x), \sigma_w^2(x) \text{ or } \sigma^2) \tag{15}$$

where the aleatoric uncertainty variance is either considered as a hyperparameter σ^2 or is also part of the NN output. Typically, if a single prediction value is required it can be given as $y^*(x) = \mu_w(x)$.

For fitting the aforementioned NNs and estimating the model parameters w, the MLE can be used. For classification, MLE corresponds to minimizing the cross-entropy loss, whereas for regression, MLE corresponds to minimizing the sum of squared errors loss. However, MLE leads to overfitting and thus, regularized MLE is used instead (which can be seen as a MAP estimate).

Nevertheless, these approaches provide only point estimates. Even if we estimate

aleatoric uncertainty through σ^2 or $\sigma_w^2(x)$ we still have no estimate for epistemic uncertainty, i.e., uncertainty related to estimating the model parameters w. For example, we would like to have an estimate for the variance of $\mu_w(x)$.

2.2 Bayesian treatment

The Bayesian treatment of classification/regression is briefly discussed here in order to draw a distinction between the Bayesian and the evidential frameworks.

In this regard, predictions are made using the posterior predictive of Eqs. (11)-(12). The likelihood $p(y|w)$ used in these equations is given by Eq. (14) for classification and by Eq. (15) for regression, and the posterior $p(w|\eta, \mathcal{D})$ is given by Eq. (5).

In realistic settings the posterior predictive is approximated with a Monte Carlo (MC) estimate as

$$p(y|x, \eta, \mathcal{D}) = \int p(y|x, w)p(w|\eta, \mathcal{D})dw \approx \frac{1}{M}\sum_{j=1}^{M} p(y|x, \hat{w}_j) \tag{16}$$

with $\{\hat{w}_j\}_{j=1}^M$ sampled from the posterior $p(w|\eta, \mathcal{D})$. The prior hyperparameters η correspond to the prior of the w parameters which are the NN parameters. They do not correspond to the prior of the likelihood parameters as in Section 1. The likelihood parameters are given as the output of the NN (e.g., $\mu_w(x)$ for the Gaussian).

As an example, for regression we can obtain the following estimates:

Mean function	$\mathbb{E}[\mu(x)]$	$\frac{1}{M}\sum_{j=1}^{M}\mu_{\hat{w}_j}(x)$
Epistemic uncertainty	$Var[\mu(x)]$	$Var\left[\{\mu_{\hat{w}_j}(x)\}_{j=1}^{M}\right]$
Aleatoric uncertainty	$\mathbb{E}[\sigma^2]$	$\frac{1}{M}\sum_{j=1}^{M}\sigma_{\hat{w}_j}^2(x)$ or σ^2

Table 1. Interpretation of Bayesian regression estimates.

The main limitation of the Bayesian framework is that it is not straightforward to produce samples $\{\hat{w}_j\}_{j=1}^M$ from the posterior. Posterior approximation techniques, depending on their accuracy, are computationally costly.

Finally, note that that the hyperparameters η can be estimated by type-II MLE, i.e., by maximizing the marginal likelihood of Eq. (13).

2.3 Evidential treatment

2.3.1 Introduction

In the deterministic framework we consider a likelihood function for each x with parameters given by a NN which is itself parametrized. For example, our prediction for $y(x)$ in the regression setting is a Gaussian likelihood with mean given by a NN. And by optimizing the NN parameters we obtain the likelihood parameters.

In the Bayesian framework the same is considered except for the fact that many different values for the NN parameters are averaged in an MC fashion. The hyperparameters η in the Bayesian framework correspond to the prior of the NN parameters and do not depend on the location x.

The evidential treatment is completely orthogonal to the above. Instead of making predictions using the likelihood we make predictions using the prior predictive of Eqs. (11)-(12). The hyperparameters η of the prior predictive are given by the output of the NN and depend on the location x. Overall, there are three main differences with the Bayesian framework:

1. For each x instead of using the likelihood with parameters given by a NN, we use the prior predictive with hyperparameters η given by a NN
2. In the Bayesian framework η corresponds to the prior of the NN parameters and is shared among all x locations, whereas in the evidential framework it corresponds to the prior of the likelihood parameters and varies depending on x
3. We deterministically optimize the evidential NN (with regularized MLE). However, epistemic uncertainty is already embedded in the prior predictive distribution so we do have an estimate for it.

2.3.2 Making predictions

As explained in Section 2.3.1, for each x the prior predictive is used for predicting $y(x)$. The required hyperparameters $\eta(x)$ are given as the output of a NN.

Specifically, for the classification setting, the prior predictive is given by Eq. (9)

and the required hyperparameters $\eta = [\alpha_1, \ldots, \alpha_K]$ are given by the sum of $[1, \ldots, 1]$ and the output of a NN; i.e., $[1, \ldots, 1]$ represents the totally uncertain state, whereas the NN represents the additional (fictitious) evidence. This sum is also the vector of probabilities for x belonging in each class.

For the regression setting, the prior predictive is a Student-t distribution (derived from Eq. (10)) and the required hyperparameters $\eta = [\gamma, v, \alpha, \beta]$ are given by a NN; i.e., the NN outputs

$$\eta_w(x) = \begin{bmatrix} \gamma_w(x) \\ v_w(x) \\ \alpha_w(x) \\ \beta_w(x) \end{bmatrix} = \begin{bmatrix} \gamma \\ v \\ \alpha \\ \beta \end{bmatrix} \tag{17}$$

These hyperparameters have the following interpretations:

Mean function	$\mathbb{E}[\mu(x)]$	γ
Epistemic uncertainty	$Var[\mu(x)]$	$\frac{\beta}{v(\alpha-1)}$
Aleatoric uncertainty	$\mathbb{E}[\sigma^2]$	$\frac{\beta}{\alpha-1}$

Table 2. Interpretation of evidential regression output.

As also mentioned earlier, epistemic uncertainty is already embedded in the output of the prior predictive; thus it is not affected by the fact that we deterministically optimize our NN). See Fig. 1 for a sketch of the evidential regression NN outputs as compared to the one of standard regression NN.

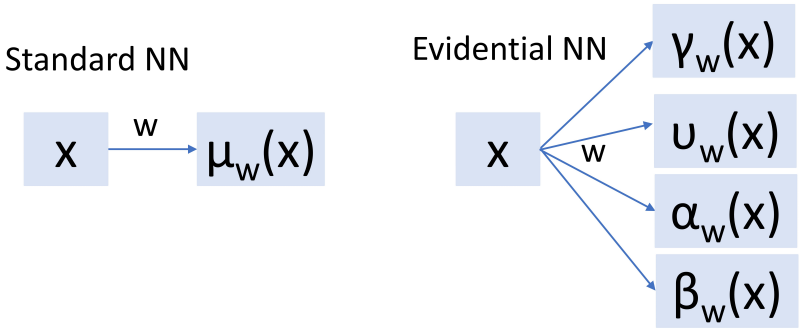

Fig. 1. Standard vs evidential NN outputs.

According to the original paper Amini et al. (2020), the evidential distribution (the prior predictive) can be construed as a higher-order distribution on top of the unknown lower-order likelihood distribution from which observations are drawn. As depicted in

Fig. 2-A, different values of fictitious evidence output by the NN correspond to different higher-order distributions. And based on NN-predicted higher-order distribution (Fig. 2-B), different likelihood functions can be drawn as shown in Fig. 2-C.

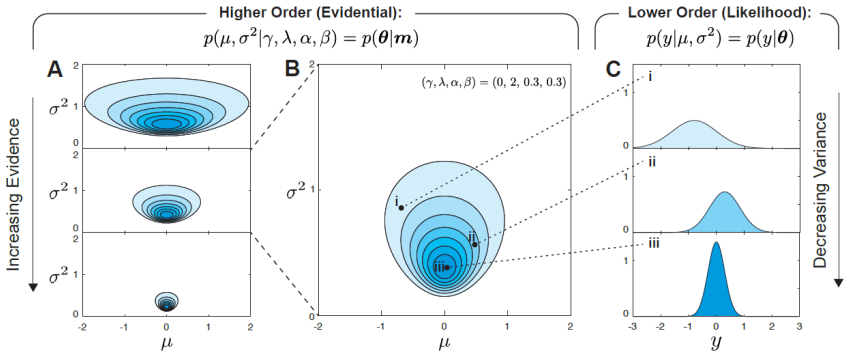

Fig. 2. According to the original paper Amini et al. (2020), the evidential distribution (the prior predictive) can be construed as a higher-order distribution on top of the unknown lower-order likelihood distribution from which observations are drawn. As depicted in part A of the figure, different values of fictitious evidence output by the NN correspond to different higher-order distributions. And based on NN-predicted higher-order distribution (part B), different likelihood functions can be drawn as shown in part C.

2.3.3 Loss function for evidential learning

As explained in previous sections, in the Bayesian framework the hyperparameters η, which parametrize the prior of the NN parameters and are shared among x locations, are tuned with type-II MLE (marginal likelihood maximization). In this regard, recall the typical interpretation of marginal likelihood: it is the probability for the dataset to occur under the prior predictive. For this reason, the marginal likelihood can also be used for defining a loss function for evidential training, in which predictions are based on the prior predictive. As a result, the loss function can be expressed as

$$Loss(w) = \sum_{i=1}^{N} -log[p(y_i|\eta_w(x_i))] \tag{18}$$

Because we have an analytical expression for $p(y_i|\eta_w(x_i))$ (Student-t for regression), it is straightforward to compute the loss corresponding to w. Training with the loss function of Eq. (18) can also be interpreted as standard MLE training with a Student-t likelihood instead of a Gaussian likelihood.

For regularization, the approach proposed by Sensoy et al. (2018) is to penalize the network for leaving the "I do not know" state; i.e., a change in w should make a significant data misfit reduction, otherwise the NN output should correspond to large uncertainty. In evidential terms, large uncertainty corresponds to small fictitious evidence (see Section 1.5).

Specifically, for the Dirichlet prior the "I do not know" state corresponds to $\eta = [1, \ldots, 1]$ which is a uniform prior that does not favor any one direction. The penalty term for this case can be given as the KL divergence of the proposed prior Dirichlet$(\eta_w(x))$ from the totally uncertain prior Dirichlet$([1, \ldots, 1])$. See Sensoy et al. (2018) for more information regarding the classification case.

For the regression case, the KL divergence between the NN-produced NIG prior and the zero-evidence NIG prior is undefined (see Amini et al., 2020). For proposing an alternative approach the following properties are desirable:

- A larger penalty value should be applied for larger data misfit
- A larger penalty value should be applied for larger fictitious evidence. Recall that prior hyperparameters can be interpreted as fictitious evidence/observations (see Section 1.5). More evidence produced by the NN corresponds to less uncertainty and it is desirable this less uncertainty to be backed by significant reduction in data misfit.

Based on the above properties, Amini et al. (2020) proposes the loss function to be regularized by

$$R(w) = \sum_{i=1}^{N} |y_i - \gamma_w(x_i)|(2\alpha_w(x_i) + v_w(x_i)) \tag{19}$$

where $2\alpha_w(x_i) + v_w(x_i)$ is the total number of fictitious observations that are produced by the NN (see Section 1.5 for more information). Note that in the first version of Amini et al. (2020) the total evidence is given as in Eq. (19), whereas in the second version it is given as $\alpha_w(x_i) + 2v_w(x_i)$, which should be a typo.

2.4 Summary

In Table 3 the evidential regression framework is compared with standard regression.

Regression / Training	Predictive function	Optimization
Standard / MLE	$\mathcal{N}(y\|\mu_{\hat{w}}(x), \sigma^2)$	$\min_w \sum_{i=1}^{N} \|\|y_i - \mu_w(x_i)\|\|_2^2$
Standard / Reg. MLE	$\mathcal{N}(y\|\mu_{\hat{w}}(x), \sigma^2)$	$\min_w \sum_{i=1}^{N} \|\|y_i - \mu_w(x_i)\|\|_2^2$ $+\lambda\|\|w\|\|_2^2$
Standard / Bayesian	$\frac{1}{M}\sum_{j=1}^{M} \mathcal{N}(y\|\mu_{\hat{w}_j}(x), \sigma^2)$	Posterior sampling
Evidential / Reg. MLE	*Student-t*$(y\|\eta_{\hat{w}}(x))$	Eq. (18) + Eq. (19)

Table 3. Evidential vs standard regression.

3 Key points through examples

A few key points based on the numerical examples of Amini et al. (2020) are summarized below:

- Evidential regression does not rely on sampling during prediction (also seen as "inference") and thus is faster than sampling-based alternatives (e.g., dropout, Bayes by backprop, HMC). However, this cost may not be very significant in general.

	N	# Parameters		Inference Speed		RMSE	NLL
		Absolute	Relative	Seconds	Relative		
Evidential (Ours)	-	7,846,776	1.00	0.003	1.00	0.024 ± 0.032	-1.128 ± 0.290
Spatial Dropout	2	7,846,657	1.00	0.028	10.20	0.033 ± 0.037	-0.564 ± 0.231
Spatial Dropout	5	7,846,657	1.00	0.031	11.48	0.031 ± 0.033	-1.227 ± 0.374
Spatial Dropout	10	7,846,657	1.00	0.037	13.69	0.035 ± 0.042	-1.139 ± 0.379
Spatial Dropout	25	7,846,657	1.00	0.065	23.99	0.032 ± 0.035	-1.137 ± 0.327
Spatial Dropout	50	7,846,657	1.00	0.107	39.36	0.032 ± 0.036	-1.110 ± 0.381
Ensembles	2	15,693,314	2.00	0.005	1.94	0.026 ± 0.032	-1.080 ± 3.334
Ensembles	5	39,233,285	5.00	0.010	3.72	0.023 ± 0.027	-1.077 ± 0.298
Ensembles	10	78,466,570	10.00	0.019	6.82	0.025 ± 0.038	-0.980 ± 0.298
Ensembles	25	196,166,425	25.00	0.045	16.45	0.022 ± 0.029	-1.000 ± 0.259
Ensembles	50	392,332,850	50.00	0.112	41.26	0.022 ± 0.031	-0.996 ± 0.275

Fig. 3. Inference speed comparison for the depth estimation regression problem. See more info in Amini et al. (2020).

- Predicted epistemic uncertainty depends highly on the regularization parameter λ of Eq.(19), as opposed to aleatoric uncertainty

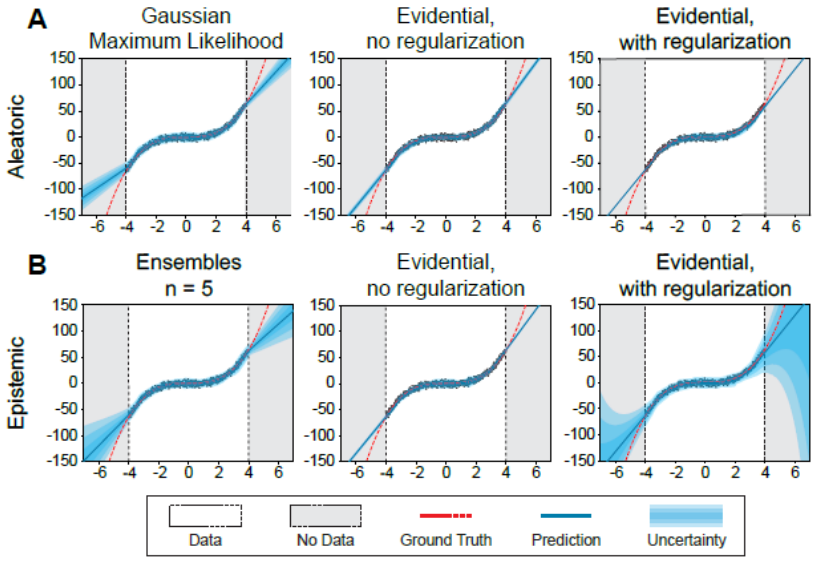

Fig. 4. Effect of regularization for the cubic function regression problem.

- With the right λ value evidential regression can give better predictions for out-of-distribution epistemic uncertainty as compared to Dropout and Deep Ensembles

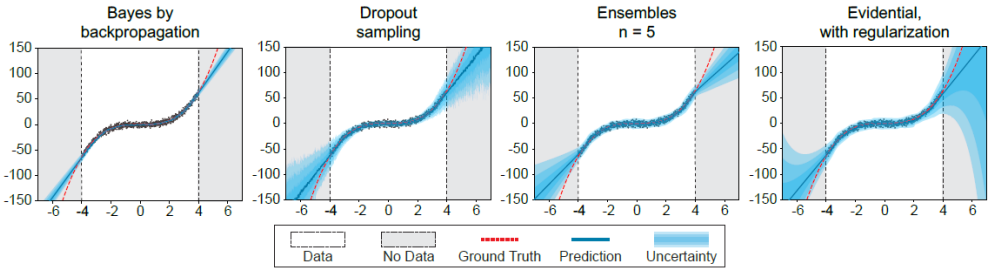

Fig. 5. Total uncertainty comparison for the cubic function regression problem.

- Evidential regression is competitive both in terms of accuracy (RMSE) and uncertainty estimation (negative log-likelihood; NLL). Note that NLL although has a term that corresponds to sum of squared errors, it also has a term that relates to the predicted uncertainty and how the data agrees with this uncertainty

	RMSE			NLL			Inference Speed (ms)		
Dataset	Dropout	Ensembles	Evidential	Dropout	Ensembles	Evidential	Dropout	Ensemble	Evidential
Boston	**2.97 ± 0.19**	**3.28 ± 1.00**	**3.06 ± 0.16**	**2.46 ± 0.06**	**2.41 ± 0.25**	**2.35 ± 0.06**	3.24	3.35	**0.85**
Concrete	**5.23 ± 0.12**	6.03 ± 0.58	5.85 ± 0.15	3.04 ± 0.02	3.06 ± 0.18	**3.01 ± 0.02**	2.99	3.43	**0.94**
Energy	**1.66 ± 0.04**	2.09 ± 0.29	2.06 ± 0.10	1.99 ± 0.02	**1.38 ± 0.22**	**1.39 ± 0.06**	3.08	3.80	**0.87**
Kin8nm	0.10 ± 0.00	**0.09 ± 0.00**	**0.09 ± 0.00**	-0.95 ± 0.01	-1.20 ± 0.02	**-1.24 ± 0.01**	3.24	3.79	**0.97**
Naval	0.01 ± 0.00	**0.00 ± 0.00**	**0.00 ± 0.00**	-3.80 ± 0.01	**-5.63 ± 0.05**	**-5.73 ± 0.07**	3.31	3.37	**0.84**
Power	**4.02 ± 0.04**	**4.11 ± 0.17**	4.23 ± 0.09	**2.80 ± 0.01**	**2.79 ± 0.04**	**2.81 ± 0.07**	2.93	3.36	**0.85**
Protein	**4.36 ± 0.01**	4.71 ± 0.06	4.64 ± 0.03	2.89 ± 0.00	2.83 ± 0.02	**2.63 ± 0.00**	3.45	3.68	**1.18**
Wine	**0.62 ± 0.01**	**0.64 ± 0.04**	**0.61 ± 0.02**	**0.93 ± 0.01**	**0.94 ± 0.12**	**0.89 ± 0.05**	3.00	3.32	**0.86**
Yacht	**1.11 ± 0.09**	**1.58 ± 0.48**	**1.57 ± 0.56**	1.55 ± 0.03	**1.18 ± 0.21**	**1.03 ± 0.19**	2.99	3.36	**0.87**

Fig. 6. Accuracy and uncertainty comparison for benchmark regression problems.

- Evidential regression effectively disentangles epistemic and aleatoric uncertainties.

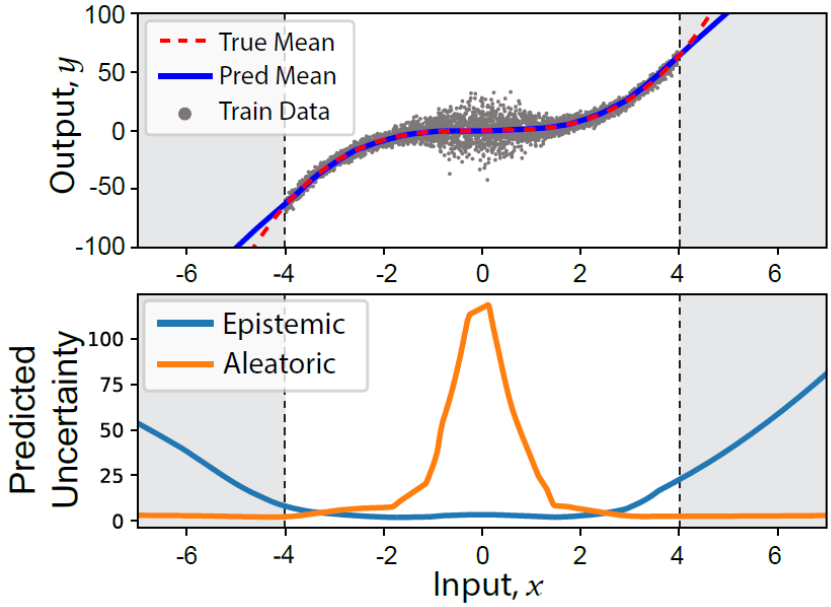

Fig. 7. Disentangled uncertainty for the cubic function regression problem with more added noise towards the center of the in-distribution region. Predicted aleatoric uncertainty increases in the middle (as it should) and predicted epistemic uncertainty increases where there is no data (as it should).

References

Amini, A., Schwarting, W., Soleimany, A., and Rus, D. (2020). "Deep Evidential Regression", p. 11.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer.

Sensoy, M., Kaplan, L., and Kandemir, M. (2018). "Evidential Deep Learning to Quantify Classification Uncertainty", p. 11.