Frequentist uncertainty estimates for deep learning

Natasa Tagasovska
HEC Lausanne
natasa.tagasovska@unil.ch

David Lopez-Paz
Facebook AI Research
dlp@fb.com

Abstract

We provide frequentist estimates of aleatoric and epistemic uncertainty for deep neural networks. To estimate aleatoric uncertainty we propose *simultaneous quantile regression*, a loss function to learn all the conditional quantiles of a given target variable. These quantiles lead to well-calibrated prediction intervals. To estimate epistemic uncertainty we propose *training certificates*, a collection of diverse non-trivial functions that map all training samples to zero. These certificates map out-of-distribution examples to non-zero values, signaling high epistemic uncertainty. We compare our proposals to prior art in various experiments.

1 Introduction

Deep learning permeates our lives, with prospects to drive our cars and decide our medical treatments. These ambitions will unlikely materialize if deep models remain unable to assess their confidence when performing in diverse situations. Being aware of uncertainty in prediction is crucial throughout multiple scenarios. First, when deciding to *abstain from prediction*. Abstaining is a reasonable strategy to deal with anomalies [4], outliers [13], out-of-distribution examples [25], defend against adversarial examples [27], or delegate high-risk predictions to humans [5]. Deep classifiers do not "know what they know", and may confidently assign one of the training categories to objects that they have never seen. Second, when *active learning* [24], the problem of deciding what examples should humans annotate to maximally improve the performance of a model. Third, when *comparing models* in cases where the structure of the noise must be captured with fidelity, including applications requiring well-calibrated prediction intervals, as well as causal discovery [21]. Finally, providing prediction uncertainties is one of the first steps towards model *interpretability* [1].

Most taxonomies consider three sources of uncertainty: approximation, aleatoric, and epistemic uncertainties [7]. First, approximation uncertainty describes the errors made by simplistic models unable to fit complex data (e.g., the error made by a linear model fitting a sinusoidal curve). Since the sequel focuses on deep neural networks, which are known to be universal approximators [6], we assume that the approximation uncertainty is negligible and omit its analysis. Second, aleatoric uncertainty (from the Greek word *alea*, meaning "rolling a dice") accounts for the *stochasticity of the data*. Aleatoric uncertainty describes the variance of the conditional distribution of our target variable given our features. This type of uncertainty arises due to unmeasured variables or measurement errors, and cannot be reduced by collecting more data from the same distribution. Third, epistemic uncertainty (from the Greek word *episteme*, meaning "knowledge") describes the errors made by the model's *lack of knowledge* about certain regions of the feature space. Therefore, epistemic uncertainty is inversely proportional to the density of our features as given by the training distribution, and can be reduced by collecting data in low density regions.

Prior literature on uncertainty estimation for deep neural networks is dominated by approximate Bayesian methods [12, 2, 14, 15, 29], many of them relying on Dropout at test time [10, 14]. On the other hand, frequentist approaches rely on expensive ensembling [3], and have been explored only recently [22, 19, 23]. Here we propose frequentist, scalable, single deep model methods to estimate aleatoric (Section 2) and epistemic (Section 3) uncertainty in deep models.

Third workshop on Bayesian Deep Learning (NeurIPS 2018), Montréal, Canada.

2 Simultaneous quantile regression for aleatoric uncertainty

Let $F(y) = P(Y \leq y)$ be the strictly monotone cumulative distribution function of a target variable Y taking real values y. Consequently, let $F^{-1}(\tau) = \inf\{y : F(y) \geq \tau\}$ denote the quantile distribution function of the same variable Y, for all quantile levels $0 \leq \tau \leq 1$. The goal of *quantile regression* is to estimate a given quantile level τ of the target variable Y, when conditioned to the values x taken by a feature variable X. That is, we are interested in building a model $\hat{y} = \hat{f}_\tau(x)$ approximating the conditional quantile distribution function $y = F^{-1}(\tau|X = x)$. One strategy to estimate such models is to minimize their *pinball loss* [9, 17, 16, 8]:

$$\ell_\tau(y, \hat{y}) = \tau(y - \hat{y}) \text{ if } y - \hat{y} \geq 0, \text{ else } (1 - \tau)(\hat{y} - y). \tag{1}$$

As expected, one recovers the absolute loss when building a quantile regressor for the level $\tau = \frac{1}{2}$. More generally, the pinball loss ℓ_τ estimates the τ-th quantile consistently [26].

Now, let us collect a dataset iid feature-target pairs $(x_1, y_1), \ldots, (x_n, y_n)$. Then, we may estimate the conditional quantile distribution of Y given X at the quantile level τ as the empirical risk minimizer $\hat{f}_\tau = \arg\min_f \frac{1}{n} \sum_{i=1}^n \ell_\tau(f(x_i), y_i)$.

Instead, we propose to estimate *all the quantile levels simultaneously* by solving: $\arg\min_f \frac{1}{n} \sum_{i=1}^n \mathbb{E}_{\tau \sim U[0,1]} [\ell_\tau(f(x_i, \tau), y_i)]$. In practice, we minimize this expression using SGD, and sample fresh random quantile levels for all training points at each mini-batch during training. The resulting function $\hat{f}(x, \tau)$ can be used to compute any quantile of the conditional variable $Y|X = x$. This allows to compute the entire conditional distribution of the target variable and, in particular, our proposed estimate for **aleatoric uncertainty**, the $1 - \alpha$ prediction interval around the median:

$$u_a(x^\star) := f(x^\star, 1 - \alpha/2) - f(x^\star, \alpha/2). \tag{2}$$

In contrast to prior art [10, 19, 23], our solution provides a prediction model able to estimate the entire profile of non-Gaussian (e.g. skew, asymmetric, multimodal) heteroskedastic noises in data.

Finally, a known phenomenon in the statistics literature of quantile regression is the problem of *crossing quantiles* [16], that is, obtaining predictions $f(x, \tau) > f(x, \tau + \epsilon)$ for some $\epsilon > 0$. One way to alleviate this issue is adding a regularization term $\max(-\frac{\partial f(x_i, \tau)}{\partial \tau}, 0)$ for each training point x_i.

3 Training certificates for epistemic uncertainty

Consider the problem of binary classification between a positive distribution P and a negative distribution Q. Construct a binary classifier f, mapping samples from P to zero, and mapping samples from Q to one. The classifier f is determined by the relative position of P and Q. Thus, if we consider a second binary classification problem between the same P and a different Q', the new optimal classifier f' may differ significantly from f. However, both classifiers f and f' have one treat in common: they map samples from the positive distribution P to zero.

The previous thought experiment illustrates the difficulty of estimating epistemic uncertainty when learning from a positive distribution P without any reference to a negative, "out-of-domain" distribution Q. That is, we are interested not only in one binary classifier mapping samples from P to zero, but in the infinite collection of such classifiers. Considering the infinite collection of classifiers mapping samples from the positive distribution P to zero, their class-probabilities should depart significantly from zero only at samples *not* from P, signaling high epistemic uncertainty.

This intuition motivates our epistemic uncertainty estimate. Consider a deep model $y = f(\phi(x))$ trained on feature-target samples drawn from the joint distribution $P(X, Y)$. Construct the dataset of high-level representations of training examples, denoted by $\Phi = \{\phi(x_i)\}_{i=1}^n$. Second, train a collection of *training certificates* $c_1, \ldots, c_k$. Each training certificate c_j is a simple neural network trained to map the dataset Φ to zero. Since we want diverse, non-trivial certificates, we train each certificate from a different random initialization and require their Lipschitz constant to be equal to one. Finally, we define our estimate of **epistemic uncertainty** as:

$$u_e(x^\star) := \max_{j=1,\ldots,k} c_j(\phi(x^\star)). \tag{3}$$

Certificates extend the common epistemic uncertainty estimate $u(x^\star) = \min_{\phi(x_i) \in \Phi} \|\phi(x_i) - \phi(x^\star)\|_2^2$, which can be written as a collection of n linear certificates with coefficients fixed by the feature representation of each training example.

4 Experiments

Prediction Intervals We evaluate our aleatoric uncertainty estimate (2) to construct 95% Prediction Intervals (PIs). These are intervals containing the true prediction about the target variable with at least 95% probability. The quality of prediction intervals is measured by their Prediction Interval Coverage Probability (PICP, the number of true observations falling inside the bounds of the estimated PI), and their Mean Prediction Interval Width (MPIW, the width of the interval averaged across all predictions). Figure 1 compares the PICP/MPIW metrics obtained by QualityDriven [23], ConditionalGaussian [19], Dropout [10], and our proposed ConditionalQuantile, across four common datasets [12]. The proposed simultaneous quantile regression (ConditionalQuantile) is able to provide the narrowest (smallest MPIW) and most calibrated (PICP $\approx 95\%$) prediction intervals.

In practice while using simultaneous estimation we did not observe the issue of quantile crossing, so we excluded the regularization term for this experiment.

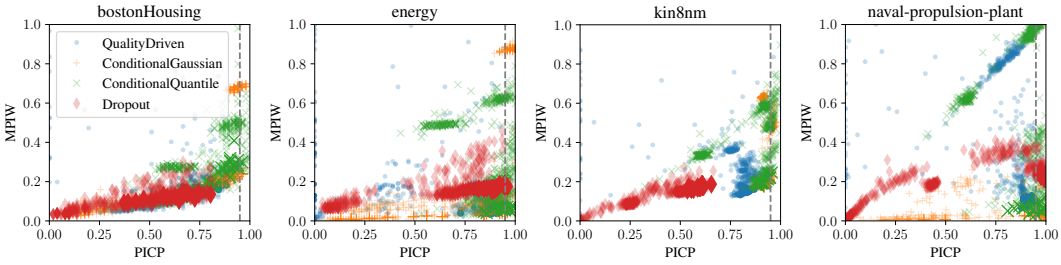


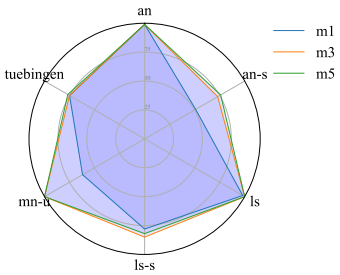
Figure 1: Experiments on prediction intervals. Multiple markers per baseline correspond to different hyperparameter sets, boldface markers are the best choices.

Causal Discovery Recently [28] it was suggested that the pinball loss can be used as a proxy to estimate Kolmogorov complexity [18]. In the context of detecting the causal direction between two variables X and Y, this means that we can interpret the regression model with lower pinball loss (from the one mapping X to Y and the one mapping Y to X) as the one following the true causal model. Figure 2 shows that using multiple quantile levels (m) to asses the quality of a model serves for causal discovery purposes, on par with the state-of-the-art on six different cause-effect datasets (See Sec. 4 from [28]).

Out-of-distribution examples We use our estimate of epistemic uncertainty (3) to detect out-of-distribution examples. First, consider a dataset with 10 classes, and split these 10 classes at random into 5 "in-distribution" classes and 5 "out-of-distribution" classes. Second, train a VGG-19 neural network on the train data from the in-distribution classes, and evaluate the epistemic uncertainty on the test data from the in-distribution and out-of-distribution classes. The training certificates were implemented as a single linear layer. The number of units in the layer corresponds to the number of certificates requested. We leverage such implementation to jointly train all of the certificates. The number of certificates used for each dataset was chosen from $[10, 100, 1000]$ according to the lowest accuracy obtained.

Figure 2 (right) shows the out-of-distribution detection accuracy obtained by different uncertainty estimation methods, where threshold to distinguish "in" or "out" of distribution examples is chosen optimally for each method. We compare against Gaussian process uncertainty (covariance), distance to nearest training sample in feature space, softmax entropy, softmax margin, geometrical margin, largest softmax score [11], the Odin detector [20], and an oracle trained on the test samples.

Finally, we also obtained interesting results for *detecting adversarial examples* and *active learning* scenarios, however, due to lack of space, we defer these experiments to future/extended versions of this manuscript.



	CIFAR	Fashion	MNIST	SVHN	mean
certificates	0.290	0.170	0.090	0.136	**0.171**
covariance	0.330	0.305	0.087	0.165	0.222
distance	0.345	0.362	0.094	0.218	0.255
entropy	0.338	0.176	0.120	0.130	0.191
functional	0.336	0.178	0.103	0.129	0.187
geometrical	0.386	0.386	0.142	0.391	0.326
largest	0.337	0.176	0.117	0.132	0.191
Odin	0.485	0.477	0.496	0.459	0.479
oracle	0.250	0.155	0.057	0.119	0.145

Figure 2: Experiment on causal discovery (left) and out-of-distribution detection (right).

References

[1] D. Alvarez-Melis and T. Jaakkola. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *EMNLP*, 2017.

[2] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural networks. In *ICML*, 2015.

[3] L. Breiman. Bagging predictors. *Machine learning*, 1996.

[4] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM computing surveys*, 2009.

[5] C. Cortes, G. DeSalvo, and M. Mohri. Learning with rejection. In *International Conference on Algorithmic Learning Theory*. Springer, 2016.

[6] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 1989.

[7] A. Der Kiureghian and O. Ditlevsen. Aleatory or epistemic? does it matter? *Structural Safety*, 2009.

[8] T. S. Ferguson. *Mathematical statistics: A decision theoretic approach*. Taylor & Francis Group, 1967.

[9] M. Fox and H. Rubin. Admissibility of quantile estimates of a single location parameter. *The Annals of Mathematical Statistics*, 1964.

[10] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016.

[11] D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *ICLR*, 2017.

[12] J. M. Hernández-Lobato and R. Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *ICML*, 2015.

[13] V. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial intelligence review*, 2004.

[14] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NeurIPS*, 2017.

[15] M. E. Khan, D. Nielsen, V. Tangkaratt, W. Lin, Y. Gal, and A. Srivastava. Fast and scalable bayesian deep learning by weight-perturbation in adam. In *ICML*, 2018.

[16] R. Koenker. *Quantile Regression*. Cambridge University Press, 2005.

[17] R. Koenker and G. Bassett Jr. Regression quantiles. *Econometrica: Journal of the Econometric Society*, 1978.

[18] A. N. Kolmogorov. On tables of random numbers. *Sankhyā: The Indian Journal of Statistics, Series A*, 1963.

[19] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017.

[20] S. Liang, Y. Li, and R. Srikant. Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. *ICLR*, 2018.

[21] D. Lopez-Paz. *From dependence to causation*. PhD thesis, University of Cambridge, 2016.

[22] I. Osband, C. Blundell, A. Pritzel, and B. Van Roy. Deep exploration via bootstrapped dqn. In *NeurIPS*, 2016.

[23] T. Pearce, M. Zaki, A. Brintrup, and A. Neely. High-Quality Prediction Intervals for Deep Learning: A Distribution-Free, Ensembled Approach. In *ICML*, 2018.

[24] B. Settles. Active learning literature survey. 2010. *Computer Sciences Technical Report*, 2014.

[25] A. Shafaei, M. Schmidt, and J. J. Little. Does your model know the digit 6 is not a cat? a less biased evaluation of" outlier" detectors. *arXiv preprint arXiv:1809.04729*, 2018.

[26] I. Steinwart, A. Christmann, et al. Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 2011.

[27] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks, 2014.

[28] N. Tagasovska, T. Vatter, and V. Chavez-Demoulin. Nonparametric quantile-based causal discovery. *arXiv preprints:1801.10579*, 2018.

[29] M. Teye, H. Azizpour, and K. Smith. Bayesian uncertainty estimation for batch normalized deep networks. *ICML*, 2018.