The Distribution Coverage Loss: A customizable-distribution-based uncertainty quantification loss function for gradient-based models.

Jarne Verhaeghe [1] **Femke Ongenae** [1] **Sofie Van Hoecke** [1]

Abstract

Uncertainty quantification is important for model acceptance and risk assessment when using machine learning models in practice. Many deep-learning uncertainty quantifying regression solutions focus on developing prediction intervals, however, predictive distributions are also frequently required. Therefore, a new loss function is proposed in this work: the Distribution Coverage Loss (DC Loss), where a custom distribution can be specified beforehand. The DC loss is usable in gradient-based models and provides a regression output and a predictive distribution, by jointly optimizing sharpness, calibration, and regression performance. Benchmark results show a large potential for the DC Loss to create calibrated and narrow models.

1. Introduction

Uncertainty quantification (UQ) is important for model acceptance and risk assessment when using machine learning models in practice. In deep learning problems, the need for UQ is therefore even higher because of lower interpretability. In many deep-learning UQ regression solutions, the focus lies on developing prediction intervals (PIs) to accompany the prediction (S. Salem et al., 2020). However, some regression applications, such as drug attainment, require an entire predictive distribution or an UQ of how likely the actual value will be in a predefined interval, together with a regression output. These deep-learning problems are frequently solved using Bayesian NNs or distribution-based models such as mean-variance models, however, they bound the predictive distribution to a specific distribution type. Furthermore, these solutions tend to prefer homoscedastic UQ. Therefore, this work adapts a specific PI deep-learning solution by T. Pearce, the Quality-Driven Loss, and converts it into a distribution-based solution providing both regression output and UQ. The type of predictive distribution is specifiable and the model uses a framework and directly optimizes sharpness and calibration (Pearce et al., 2018). To assess performance, both models are evaluated on two benchmark datasets.

[1]IDLab, University of Ghent, Belgium. Correspondence to: Jarne Verhaeghe <jarne.verhaeghe@ugent.be>.

Presented at the ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning., Copyright 2021 by the author(s).

2. Background

In regression problems the observable targets y are theorized to consist of a ground truth function $f(x)$, given the input features x, and additive noise ϵ. When predicting the target variable, we try to find an estimator such that $\hat{y} = \hat{f}(x)$ closely resembles the targets y. On the contrary, in UQ the goal is to correctly approximate and describe the predictive distribution $\mathbb{P}_{\hat{\mathbf{Y}}}$ of the outputs $\hat{y}$ such that it correctly encompasses all sources of uncertainty.

Two different sources of uncertainty can be identified. The first source is defined as the aleatoric or data uncertainty and entails the noise or uncertainty ϵ that is intrinsically bound to the data and the target variables. Therefore this uncertainty cannot be reduced by adjusting the model or through the data, as a result, aleatoric uncertainty is also often named the irreducible uncertainty. This uncertainty can be categorized into homoscedastic aleatoric uncertainty, where the noise is equal across all datapoints $\forall x_1, x_2 \in X : \epsilon(x_1) = \epsilon(x_2)$, or heteroscedastic uncertainty describing varying noise (Pearce et al., 2018).

The other source of uncertainty originates from the model and is called epistemic uncertainty. In comparison to the aleatoric uncertainty, this one is reducible and therefore a perfect model contains no epistemic uncertainty. Various factors contribute to this uncertainty such as model parameter uncertainty of reaching the true optimum, model specification or bias of how closely the model approximates the ground truth model and the training data uncertainty or variance risen from the selecting a certain training set, how it represents the actual data and how sensitive the model is to other sets (Pearce et al., 2018)(Kiureghian & Ditlevsen, 2009).

A confidence interval (CI) essentially quantifies the epistemic uncertainty, while the prediction intervals provide the total uncertainty also encompassing the aleatoric uncertainty.

Evaluating the predictive distribution is a frequently researched topic and is also one of the focuses of this work. In literature, two different taxonomies are used, however, both evaluate the same properties of the predictive distribution. The first stems from a more heuristic evaluation of the results using PICP and NMPIW (Khosravi et al., 2011)(Lakshminarayanan et al., 2017), while the second one builds upon probability theory discussing calibration and sharpness (Zhao et al., 2020)(Tran et al., 2020)(Kiureghian & Ditlevsen, 2009). The calibration quantifies how well the predictive distribution captures the ground truth uncertainty of the predictions by evaluating and comparing every quantile of the predictive distribution. The sharpness indicates the size of this predictive distribution and has a lower bound equal to the size of the aleatoric uncertainty. The sharpness and calibration are both required to effectively evaluate the predictive distribution and a trade-off exists between these two metrics. A distribution can be perfectly calibrated, however, be very large and therefore not provide qualitative estimations and on the contrary, a distribution can be small but very poorly calibrated and does not depict the total uncertainty. Zhao et al. (Zhao et al., 2020) thoroughly discussed different possible calibration metrics.

The strongest form of calibration, besides perfect calibration, is called individual calibration where for each estimate $\hat{y}$, the individual predictive distribution $\mathbb{P}_{\hat{\mathbf{Y}}|x}$ is equal to the ground truth probability distribution of that estimate $\mathbb{P}_{\mathbf{Y}|x}$. Achieving full individual calibration is the goal of UQ, however, evaluating it is practically impossible for finite datasets. It requires the conditional cumulative distribution function $\mathbb{F}_{\mathbf{Y}|x}$ of the target values for all inputs individually and as finite datasets only have at most one y per input x, this function cannot be extracted.

A much more relaxed form of calibration is average calibration where $\mathbb{P}_{\hat{\mathbf{Y}}|x}$ should be on average equal to $\mathbb{P}_{\mathbf{Y}|x}$, often simply referred to as calibration (Kuleshov et al., 2018). As a result, this relaxes the requirement of $\mathbb{F}_{\mathbf{Y}|x}$ to $\mathbb{F}_{\mathbf{Y}}$, which can be easily estimated from y and therefore enables quantifying average calibration. Given a dataset of size N, the average calibration $\hat{p}_{avg}$ for a certain quantile p and quantile function estimator $\hat{\mathbb{Q}}$ can be written as (Chung et al., 2020):

$$\hat{p}_{avg}(p) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\{y_i \leq \hat{\mathbb{Q}}(x_i, p)\} \tag{1}$$

Where $\mathbb{I}$ is the indicator function, resulting in 1 if the input is true and 0 otherwise. $\hat{p}_{avg}(p)$ effectively counts how many observed values lie below the estimated quantiles $\hat{\mathbb{Q}}(x, p)$. If $\hat{p}_{avg}(p) = p$ for all quantiles $p \in (0, 1)$ than the model is average calibrated. The sum of the absolute differences between both, for all quantile,s is then defined as the expected calibration error (ECE):

$$ECE = \int_0^1 |\hat{p}_{avg}(p) - p| dp \tag{2}$$

$\hat{p}_{avg}(p)$ can also be interpreted as the amount of samples that lie in between the PI]-$\infty, \hat{\mathbb{Q}}(x, p)$] with expected coverage probability of p. Generalizing this formula for any PI of arbitrary expected coverage probability p with lower and upper bound $[y_{L_i}, y_{U_i}]$ forms the Prediction Interval Coverage Probability (PICP) and is therefore also a metric for average calibration:

$$\begin{aligned} PICP(\mathbf{y}, \mathbf{y}_L, \mathbf{y}_U) &= \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\{y_{L_i} \leq y_i \leq y_{U_i}\} \\ &= \frac{1}{N}\sum_{i=1}^{N} k_i = \frac{c}{N} \end{aligned} \tag{3}$$

An important notice is that individually calibrated models are always average calibrated models, however, in the extreme case where the predictive distribution is equal to the general data distribution, i.e. $\mathbb{P}_{\hat{\mathbf{Y}}|x} = \mathbb{P}_{\mathbf{Y}}$ for each x, the model is perfectly average calibrated but not individually calibrated and provides no useful uncertainty estimation.

To combat this, it is advised to accompany the average calibration with sharpness. In the mentioned case, the size of $\mathbb{P}_{\hat{\mathbf{Y}}|x}$ is approximately equal to the range of the observed values. The sharpness of a predictive distribution is often quantified by the variance of the predictive distribution but can also be quantified by the (Normalized) Mean Prediction Interval Width ((N)MPIW).

$$MPIW = \frac{1}{N}\sum_{i=1}^{N} y_{U_i} - y_{L_i} \tag{4}$$

$$NMPIW = \frac{MPIW}{y_{max} - y_{min}} \tag{5}$$

Comparing the (N)MPIW for a specific coverage probability p with the width of the $(1-p, p)$ quantile interval of the complete data distribution $\mathbf{Y}$, can detect whether the model is perfectly average calibrated but not individually:

Theorem 1 *A calibration estimation model is perfectly average calibrated and not individually calibration if and only if:*

$$ECE = 0 \quad \textit{and} \quad \mathbb{P}_{\hat{\mathbf{Y}}|x} = \mathbb{P}_{\mathbf{Y}}$$

Then the (N)MPIW of $\mathbb{P}_{\hat{\mathbf{Y}}|x}$ for any coverage p:

$$(N)MPIW = \mathbb{Q}_{\mathbf{Y}}(p) - \mathbb{Q}_{\mathbf{Y}}(1-p)$$

Where $\mathbb{Q}$ is the quantile function of $\mathbb{P}_{\mathbf{Y}}$

As a result, if the $(N)MPIW < \mathbb{Q}_{\mathbf{Y}}(p) - \mathbb{Q}_{\mathbf{Y}}(1-p)$, then $\mathbb{P}_{\hat{\mathbf{Y}}|x} = \mathbb{P}_{\mathbf{Y}}$ is disproven. The lower NMPIW, the more individual calibration is much more plausible.

3. Methods

3.1. The Quality-Driven loss

This work can be seen as a distribution-extension of the UQ solution by T. Pierce, the Quality-Driven (QD) loss (Pearce et al., 2018). The QD loss directly optimizes calibration and sharpness by using a variant of the MPIW definition, $MPIW_{capt}$, where only the MPIW of samples within the PI are taken into consideration. This loss function can be used in multi-output neural networks, outputting a lower y_{Li} and upper bound y_{Ui}:

$$MPIW_{capt} = \frac{1}{c}\sum_{i=1}^{N}(y_{Ui} - y_{Li}) \cdot k_i \tag{6}$$

The QD loss is then defined as:

$$L_{QD} = MPIW_{capt} + \lambda \frac{N}{\alpha(1-\alpha)} max(0, (1-\alpha) - PICP)^2 \tag{7}$$

Where α is the specified fraction that is covered by the PIs and λ controls the calibration-sharpness trade-off. The PICP is not derivable, therefore all k values in the above formulas are replaced by its derivable variant k_{soft}, where s is the soften of the sigmoid, σ, function, and a hyperparameter. As a result, this loss can be used in gradient-descent-based models. (Pearce et al., 2018).

$$k_{soft}(\mathbf{y}, \mathbf{y_L}, \mathbf{y_U}) = \sigma(s(\mathbf{y} - \mathbf{y_L})) \odot \sigma(s(\mathbf{y_U} - \mathbf{y})) \tag{8}$$

3.2. The Distribution Coverage Loss

The QD-loss was developed for PIs, and therefore not suited for predictive distribution estimation and providing a regression output. Therefore, this work tries to add a regression output and direct predictive distribution optimization to the QD-loss solution, while optimizing sharpness, calibration, and regression performance for multi-output regression models outputting a lower y_{Li} and upper bound y_{Ui}. The proposed loss, the Distribution Coverage Loss (DC Loss), works with any continuous distribution having at most two parameters to specify and a closed-form quantile function (such as the Gaussian Distribution, Logistic Distribution, Shifted Rayleigh, see A in the appendix). Other distributions with more parameters P are possible, however, then P-2 parameters are considered hyperparameters.
The first step consists of defining a coverage function that calculates the PICP, given the quantile function Q of the distribution, the parameters of the distribution θ, and the required coverage percentage p.

$$C(\mathbf{y}, p, \theta) = PICP\left(\mathbf{y}, Q\left(\frac{1-p}{2}, \tilde{\theta}\right), Q\left(\frac{1+p}{2}, \tilde{\theta}\right)\right) \tag{9}$$

Given C, the Root Mean Square Error (RMSE), and $MPIW_{capt}$, the Distribution Coverage Loss can now be defined as:

$$L_{DC}(\mathbf{y}, \tilde{\mathbf{y}}, \tilde{\theta}) = \lambda_r(RMSE(\mathbf{y}, \tilde{\mathbf{y}}) + \frac{1}{N}\sum_{i}^{N}(m_{u_i} - m_{l_i})) + \lambda_s\, MPIW_{capt} + \frac{\lambda_c}{N}\sum_{i=1}^{S}\left(min\left(0, C(y, i/S, \tilde{\theta}) - \frac{i}{S}\right)\right)^2 \tag{10}$$

With $\tilde{\mathbf{y}}$ the inferred regression mean, the distribution parameters $\tilde{\theta}$, and m_{u_i} and m_{l_i} the modes of the distribution with parameters $\tilde{\theta}$ inferred by the predicted upper and lower bound respectively (see A). m_{u_i} and m_{l_i} are defined to reduce the distribution parameter and regression mean inference error. $\tilde{\mathbf{y}}$ and $\tilde{\theta}$ are calculated using the outputted upper and lower bound, depending on the used distribution. λ_r, λ_s, and λ_c are hyperparameters to balance the regression performance, sharpness and calibration respectively.
The min makes sure the coverage is only quadratically penalised if it is below the required coverage, as having a higher coverage is often more preferred and therefore not penalised in the loss function, also resulting in more stable optimization.
As was done in the QD-loss, the PICP-base of the coverage C formula (equation 9) is converted into a continuous variant using the sigmoid definition in equation 8, for use in gradient descent-based models.

3.3. Calibration Errors

As mentioned in section 2, measuring the calibration and sharpness together is important to verify individual calibration and avoid only having average calibration. Mostly only the PICP is used to verify the PI coverage, however, this does not work for distributions. Therefore, the (Absolute) Distribution Coverage Error is proposed, based upon the ECE in equation 2, for heuristic calibration calculation in distributions. A sampling rate S is defined for the heuristic calculation of the ECE, corresponding to the step size of the percentages. To bound the absolute values of DCE and ADCE to $[0, 1]$, both are multiplied by 2:

$$DCE(y, \tilde{\theta}) = \frac{2}{S}\sum_{i=0}^{S}\left(C(y, i/S, \tilde{\theta}) - \frac{i}{S}\right) \tag{11}$$

$$ADCE(y, \tilde{\theta}) = \frac{2}{S}\sum_{i=0}^{S}\left|C(y, i/S, \tilde{\theta}) - \frac{i}{S}\right| \tag{12}$$

Table 1. Test results on benchmark datasets, using a coverage percentage of 95% for PI extraction for the PICP and NMPIW.

MODEL	RMSE	R^2	ADCE	DCE	PICP	NMPIW	$NMPIW_{0.95}$	λ	λ_c
BOSTON HOUSING									
DC GAUSSIAN	**3.071**	**0.863**	**0.040**	-0.017	0.925	**0.187**	0.219	-	0.1
DC LOGISTIC	3.165	0.850	0.075	**-0.002**	**0.957**	0.223	**0.209**	-	0.15
DC SHIFTED RAYLEIGH	3.084	0.856	0.099	0.085	0.916	0.215	0.234	-	0.045
QD	4.018	0.721	0.061	0.009	0.933	0.202	0.231	0.4	-
CONCRETE COMPRESSION STRENGTH									
DC GAUSSIAN	**5.284**	**0.875**	**0.033**	-0.012	0.930	0.249	**0.273**	-	0.125
DC LOGISTIC	5.346	0.871	0.047	**-0.001**	**0.931**	0.256	0.287	-	0.05
DC SHIFTED RAYLEIGH	5.305	0.874	**0.033**	0.008	0.920	**0.246**	0.287	-	0.1
QD	5.901	0.833	0.046	-0.029	0.974	0.312	0.279	0.4	-

The ADCE quantifies the average calibration of the complete predictive distribution. The DCE shows any calibration biases, either consistently underestimating (negative) or overestimating its coverage (positive), however, DCE can be 0 while ADCE can be 1, but not vice versa.

The calibration results can be plotted in calibration plots for further visual inspection of the calibration performance if required. Furthermore, using the calibration plot we can define the $NMPIW_p$, to compare the sharpness-calibration trade-off of the models. As multiple different distributions are used, a standard formula for sharpness for distributions, cannot be used reliably. Additionally, the NMPIW can only be compared between models if the coverage of the PIs is the same, therefore, for each model, the width of the PI corresponding with the measured coverage of p is written as $NMPIW_p$.

4. Results

For model evaluation, the same model structure as in the work of T. Pierce is used, using an ensemble of 5 multi-output neural networks (NN) with two outputs; upper and lower bound of a PI (Pearce et al., 2018). A coverage percentage p (or $1-\alpha$) of 90% is specified that corresponds with the outputted lower and upper bound of all models. Afterward, the output of the ensemble is calculated by the mean the outputs of the 5 NNs added or subtracted with $\sqrt{(2)} * erf^{-1}(p)$ for the final upper and lower bound respectively. Five-fold cross-validation was used for hyperparameter tuning. The QD model is converted into a predictive distribution model using the DC framework for comparison. For each dataset, the reported values are the averages of five runs with different NN initializations.

Two benchmark datasets were used for this work: the Concrete Compression Strength and Boston Housing datasets, with a test-set size of 20% and sklearn random state of 42. Both models used all features of both datasets, using a network structure of $(input, 50, 2)$ with learning rates of 0.005 using the Adam optimizer with 0.98 decay in Tensorflow. 400 epochs were used for the Boston dataset, 800 epochs for the concrete dataset. All DC models used a λ_r and λ_s of 1. The results are shown in table 1.

The regression performance of the DC models is higher compared to the QD models. The general calibration is slightly better and the sharpness is slightly more narrow, looking at the $NMPIW_{0.95}$, this difference is somewhat larger for the Boston dataset. Interestingly, the $NMPIW$ of the QD Concrete model is the worst, but the $NMPIW_{0.95}$ is the second best, showing the need for fair comparison of sharpness in PIs and distributions.

5. Conclusion

In this paper, a variant on the QD-loss was proposed that can incorporate and directly optimize different distributions in a framework based upon the theory of calibration and sharpness, with the requirement that the distributions have a closed-form quantile function. These models can achieve high calibration and sharpness, whilst still providing an accurate regression output. Additionally, the DC-loss is not limited to deep-learning models and can be used in any gradient-based model with multi-output regression, which can be explored in future work. Furthermore, this paper proposed a way of reliably comparing the sharpness and calibration of distribution-based models using the $NMPIW_p$, DCE, and $ADCE$ metrics, resulting in an informed sharpness-calibration trade-off for predictive distributions and extracted PIs.

Software and Data

Source files can found at https://github.com/juthsty/The-Distribution-Coverage-Loss. Data is freely available.

Acknowledgements

The authors thank FWO for funding this research in the FWO Junior Research project HEROI2C.

References

Chung, Y., Neiswanger, W., Char, I., and Schneider, J. Beyond pinball loss: Quantile methods for calibrated uncertainty quantification, 2020.

Khosravi, A., Nahavandi, S., Creighton, D., and Atiya, A. F. Lower upper bound estimation method for construction of neural network-based prediction intervals. *IEEE Transactions on Neural Networks*, 22(3):337–346, 2011. doi: 10.1109/TNN.2010.2096824.

Kiureghian, A. D. and Ditlevsen, O. Aleatory or epistemic? does it matter? *Structural Safety*, 31(2):105 – 112, 2009. ISSN 0167-4730. doi: https://doi.org/10.1016/j.strusafe.2008.06.020. URL http://www.sciencedirect.com/science/article/pii/S0167473008000556. Risk Acceptance and Risk Communication.

Kuleshov, V., Fenner, N., and Ermon, S. Accurate uncertainties for deep learning using calibrated regression. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2796–2804, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/kuleshov18a.html.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6405–6416, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Pearce, T., Brintrup, A., Zaki, M., and Neely, A. High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4075–4084, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/pearce18a.html.

S. Salem, T., Langseth, H., and Ramampiaro, H. Prediction intervals: Split normal mixture from quality-driven deep ensembles. In Peters, J. and Sontag, D. (eds.), *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pp. 1179–1187. PMLR, 2020. URL http://proceedings.mlr.press/v124/saleh-salem20a.html.

Tran, K., Neiswanger, W., Yoon, J., Zhang, Q., Xing, E., and Ulissi, Z. W. Methods for comparing uncertainty quantifications for material property predictions. *Machine Learning: Science and Technology*, 1(2):025006, may 2020. doi: 10.1088/2632-2153/ab7e1a. URL https://doi.org/10.1088/2632-2153/ab7e1a.

Zhao, S., Ma, T., and Ermon, S. Individual calibration with randomized forecasting, 2020.

A. Distribution parameter calculation from upper and lower bound output

The following calculations start from predicted upper $\tilde{y_U}$ and lower $\tilde{y_L}$ bound and predefined coverage estimation percentage p.

A.1. Gaussian Distribution

The quantile function of the Gaussian distribution is:

$$Q(p;\mu,\sigma) = \mu + \sigma\sqrt{2}erf^{-1}(2p-1) \tag{13}$$

With μ and σ parameters of the Gaussian distribution.

Given the quantile function, an upper $\tilde{y_U}$ and lower $\tilde{y_L}$ bound of a centered prediction interval, and a specifed coverage p, it is possible to infer the distribution parameters:

$$\tilde{\sigma} = \left| \frac{\tilde{y_U} - \tilde{y_L}}{\sqrt{2}(erf^{-1}(p) - erf^{-1}(-p))} \right| \tag{14}$$

To avoid a negative σ, the absolute value is taken, in the event that $\tilde{y_L} > \tilde{y_U}$. And:

$$\tilde{\mu_u} = \tilde{y_U} - \sigma\sqrt{2}erf^{-1}(p) \tag{15}$$

$$\tilde{\mu_l} = \tilde{y_L} - \sigma\sqrt{2}erf^{-1}(-p) \tag{16}$$

$$\tilde{\mu} = \frac{\mu_u + \mu_l}{2} \tag{17}$$

Where the difference between μ_l and μ_u is minimalized in the loss DC function. An alternative would be to directly take the mean value of $\tilde{y_U}$ and $\tilde{y_L}$. $\tilde{\mu}$ can also be interpreted as the regression output, where the regression output is assumed to be the mode of the distribution.

Now, given the estimated distribution parameters $\tilde{\mu}$ and $\tilde{\sigma}$ and the distribution quantile function, any prediction interval can now be estimated for a given coverage p:

$$[y_{L_p}, y_{U_p}] = \left[Q\left(\frac{1-p}{2}, \tilde{\theta}\right), Q\left(\frac{1+p}{2}, \tilde{\theta}\right) \right] \tag{18}$$

A.2. Logistic Distribution

The quantile function of the Logistic distribution is:

$$Q(p;\mu,\sigma) = \mu + \sigma ln\left(\frac{p}{1-p}\right) \tag{19}$$

With μ and σ parameters of the Logistic distribution.

Given the quantile function, an upper $\tilde{y_U}$ and lower $\tilde{y_L}$ bound of a centered prediction interval, and a specifed coverage p, it is possible to infer the distribution parameters:

$$\tilde{\sigma} = \left| \frac{\tilde{y_U} - \tilde{y_L}}{ln\left(\frac{p^2}{(1-p)^2}\right)} \right| \tag{20}$$

To avoid a negative σ, the absolute value is taken, in the event that $\tilde{y_L} > \tilde{y_U}$. And:

$$\tilde{\mu_u} = \tilde{y_U} - \sigma ln\left(\frac{p}{1-p}\right) \tag{21}$$

$$\tilde{\mu_l} = \tilde{y_L} - \sigma ln\left(\frac{1-p}{p}\right) \tag{22}$$

$$\tilde{\mu} = \frac{\mu_u + \mu_l}{2} \tag{23}$$

Where the difference between μ_l and μ_u is minimalized in the loss DC function. An alternative would be to directly take the mean value of $\tilde{y_U}$ and $\tilde{y_L}$. $\tilde{\mu}$ can also be interpreted as the regression output, where the regression output is assumed to be the mode of the distribution.

Now, given the estimated distribution parameters $\tilde{\mu}$ and $\tilde{\sigma}$ and the distribution quantile function, any prediction interval can now be estimated for a given coverage p:

$$[y_{L_p}, y_{U_p}] = \left[Q\left(\frac{1-p}{2}, \tilde{\theta}\right), Q\left(\frac{1+p}{2}, \tilde{\theta}\right) \right] \tag{24}$$

A.3. Shifted Rayleigh Distribution

A Rayleigh distribution always starts in 0, therefore a shift β is defined to make the origin start in any possible location.

The quantile function of the shifted Rayleigh distribution is:

$$Q(p;\beta,\sigma) = \beta + \sigma\sqrt{-2ln(1-p)} \tag{25}$$

With β and σ parameters of the shifted Rayleigh distribution.

Given the quantile function, an upper $\tilde{y_U}$ and lower $\tilde{y_L}$ bound of a centered prediction interval, and a specifed coverage p, it is possible to infer the distribution parameters:

$$\tilde{\sigma} = \left| \frac{\tilde{y_U} - \tilde{y_L}}{\sqrt{-2ln(1-p)} - \sqrt{-2ln(p)}} \right| \tag{26}$$

To avoid a negative σ, the absolute value is taken, in the event that $\tilde{y_L} > \tilde{y_U}$. And:

$$\tilde{\beta_u} = \tilde{y_U} - \sigma\sqrt{-2ln(1-p)} \tag{27}$$

$$\tilde{\mu_u} = \sigma + \beta_u \tag{28}$$

$$\tilde{\beta_l} = \tilde{y_L} - \sigma\sqrt{-2ln(p)} \tag{29}$$

$$\tilde{\mu_l} = \sigma + \beta_l \tag{30}$$

$$\tilde{\mu} = \frac{\mu_u + \mu_l}{2} \tag{31}$$

Where the difference between μ_l and μ_u is minimalized in the loss DC function. An alternative would be to directly take the mean value of $\tilde{y_U}$ and $\tilde{y_L}$. $\tilde{\mu}$ can also be interpreted

as the regression output, where the regression output is assumed to be the mode of the distribution.

Now, given the estimated distribution parameters $\tilde{\mu}$ and $\tilde{\sigma}$ and the distribution quantile function, any prediction interval can now be estimated for a given coverage p:

$$[y_{L_p}, y_{U_p}] = \left[Q\left(\frac{1-p}{2}, \tilde{\theta}\right), Q\left(\frac{1+p}{2}, \tilde{\theta}\right) \right] \quad (32)$$