Uncertainty Toolbox: an Open-Source Library for Assessing, Visualizing, and Improving Uncertainty Quantification

Youngseog Chung [1 2] **Ian Char** [2] **Han Guo** [3] **Jeff Schneider** [1] **Willie Neiswanger** [4]

1. Introduction

As machine learning (ML) systems are increasingly deployed on an array of high-stakes tasks, there is a growing need to robustly quantify their predictive uncertainties. Uncertainty quantification (UQ) in machine learning generally refers to the task of quantifying the confidence of a given prediction, and this measure of confidence can be especially crucial in a variety of downstream applications, including Bayesian optimization (Jones et al., 1998; Shahriari et al., 2015), model-based reinforcement learning (Malik et al., 2019; Yu et al., 2020), and in high-stakes prediction settings where errors incur large costs (Wexler, 2017; Rudin, 2019).

UQ is often performed via *distributional predictions* (in contrast with *point predictions*). Hence, given inputs $x \in \mathcal{X}$ and targets $y \in \mathcal{Y}$, the typical goal in UQ is to approximate the true conditional distribution of y given x. In the supervised setting where we only have access to a limited data sample, we are then faced with the question, "how can one verify whether a distributional prediction is close to the true distribution using only a finite dataset?" Many works in UQ tend to be disjoint in the evaluation metric utilized, which sends divided signals about which metrics *should* or *should not* be used. For example, some works report likelihood on a test set (Lakshminarayanan et al., 2017; Detlefsen et al., 2019; Zhao et al., 2020), some works use other proper scoring rules (Maciejowska et al., 2016; Askanazi et al., 2018; Bowman et al., 2020; Bracher et al., 2021), while others focus on calibration metrics (Kuleshov et al., 2018; Cui et al., 2020). Further, with disparate implementations for each metric, it is often the case that reported numerical results are not directly comparable across different works, even if the same metric is used.

To address this, we present *Uncertainty Toolbox*: an open-source python library that helps to assess, visualize, and improve UQ. There are other libraries such as Uncertainty Baselines (Nado et al., 2021) and Robustness Metrics (Djolonga et al., 2020) that focus on aspects of UQ in the *classification* setting. Uncertainty Toolbox focuses on the *regression* setting and additionally aims to provide user-friendly utilities such as visualizations, a glossary, and an organized collection of key references in UQ.

We begin our discussion by first introducing the contents of Uncertainty Toolbox. We then provide an overview of evaluation metrics in UQ. Afterwards, we demonstrate the functionalities of the toolbox with a case study where we train probabilistic neural networks (PNNs) (Nix and Weigend, 1994; Lakshminarayanan et al., 2017) with a set of different loss functions, and evaluate the resulting trained models using metrics and visualizations in the Toolbox. This case study shows that certain evaluation metrics shed light on different aspects of UQ performance, and makes the case for using a suite of metrics for a comprehensive evaluation.

2. Toolbox Contents

Uncertainty Toolbox comprises four main functionalities, which we detail below.

Evaluation Metrics The Toolbox provides implementations for a suite of evaluation metrics. The main categories of metrics are: calibration, group calibration, sharpness, and proper scoring rules. We discuss each of these metric types in the following section (Section 3).

Recalibration We further implement recalibration methods that leverage isotonic regression (Kuleshov et al., 2018). Concretely, recalibration aims to improve the average calibration (defined in Eq. (1)) of distributional predictions.

Visualizations The Toolbox offers a range of easy-to-use visualization utilies to help in inspecting and evaluating UQ quality. These plotting utilities focus on visualizing the predicted distribution, calibration, and prediction accuracy.

Pedagogy For those unfamiliar with the area of predictive UQ, we provide a glossary that communicates the core concepts in this area, and a paper list which organizes and maintains key papers in the field.

We hope the Toolbox serves as an intuitive guide for those unfamiliar but interested in utilizing UQ, and as a practical tool and point of reference for those active in UQ research. Uncertainty Toolbox is available at

[1]Robotics Institute, [2]Machine Learning Department, [3]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA [4]Computer Science Department, Stanford University, California, USA.

Presented at the ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning.,

https://github.com/uncertainty-toolbox/uncertainty-toolbox.

3. Evaluation Metrics in Predictive UQ

To summarize the notation and setting: **X, Y** denote random variables; x, y denote realized values; and $\mathcal{X}, \mathcal{Y}$ denote sets of possible values. Further, for any random variable, we denote the true CDF as $\mathbb{F}$, its inverse (i.e. the quantile function) as $\mathbb{Q}$, the corresponding density function as f, and the space of distributions as $\mathcal{F}$. Estimates of these true functions will be denoted with a hat, e.g. $\hat{\mathbb{F}}$ and $\hat{f}$. Lastly, we consider the regression setting where $\mathcal{Y} \subset \mathbb{R}$ and $\mathcal{X} \subset \mathbb{R}^n$.

One evaluation metric that many recent works have focused on are the notions of *calibration* and *sharpness* (Gneiting et al., 2007; Guo et al., 2017; Kuleshov et al., 2018; Song et al., 2019; Tran et al., 2020; Zhao et al., 2020; Fasiolo et al., 2020; Cui et al., 2020). Calibration in the regression setting is defined in terms of quantiles, and broadly speaking, calibration requires that the probability of observing the target random variable below a predicted p^{th} quantile is equal to the *expected probability* p, for all $p \in (0, 1)$. We refer to the former quantity as the *observed probability* (also referred to as empirical probability) and denote it $p^{\text{obs}}(p)$, for an expected probability p. Calibration requires $p^{\text{obs}}(p) = p$, $\forall p \in (0, 1)$. From this generic statement, we can describe different notions of calibration based on how p^{obs} is defined.

The most common form of calibration is **average calibration**, where $\hat{\mathbb{Q}}_p(x)$ is the estimated p^{th} quantile of $\mathbf{Y}|x$,

$$p^{\text{obs}}_{avg}(p) := \mathbb{E}_{x \sim \mathbb{F}_{\mathbf{X}}}[\mathbb{F}_{\mathbf{Y}|x}(\hat{\mathbb{Q}}_p(x))], \quad \forall p \in (0, 1), \tag{1}$$

i.e. the probability of observing the target below the quantile prediction, *averaged over* $\mathbb{F}_{\boldsymbol{X}}$, is equal to p. Average calibration is often referred to simply as "calibration" (Kuleshov et al., 2018; Cui et al., 2020), and it is amenable to estimation in finite datasets, as follows. Given a dataset $D = \{(x_i, y_i)\}_{i=1}^N$, we can estimate $p^{\text{obs}}_{avg}(p)$ with $\hat{p}^{\text{obs}}_{avg}(D, p) = \frac{1}{N}\sum_{i=1}^N \mathbb{I}\{y_i \leq \hat{\mathbb{Q}}_p(x_i)\}$. The degree of error in average calibration is commonly measured by *expected calibration error* (Guo et al., 2017; Tran et al., 2020; Cui et al., 2020), $\text{ECE}(D, \hat{\mathbb{Q}}) = \frac{1}{m}\sum_{j=1}^m \left|\hat{p}^{\text{obs}}_{avg}(D, p_j) - p_j\right|$, where p_j is a range of expected probabilities of interest. Note that if our quantile estimate achieves average calibration then $\hat{p}^{\text{obs}}_{avg} \to p$ (and thus ECE $\to 0$) as $N \to \infty$, $\forall p \in (0, 1)$.

It may be possible to have an uninformative, yet average calibrated model. For example, quantile predictions that match the true *marginal* quantiles of $\mathbb{F}_{\mathbf{Y}}$ will be average calibrated, but will hardly be useful since they do not depend on the input x. Therefore, the notion of **sharpness** is also considered, which quantifies the concentration of distributional predictions (Gneiting et al., 2007). For example, in predictions that parameterize a Gaussian, the variance of the predicted distribution is often taken as a measure of sharpness. There generally exists a tradeoff between average calibration and sharpness (Murphy, 1973; Gneiting et al., 2007).

Recent works have suggested a notion of calibration stronger than average calibration, called adversarial group calibration (Zhao et al., 2020). This stems from the notion of **group calibration** (Kleinberg et al., 2016; Hébert-Johnson et al., 2017), which prescribes measurable subsets $\mathcal{S}_i \subset \mathcal{X}$ s.t. $P_{x \sim \mathbb{F}_{\mathbf{X}}}(x \in \mathcal{S}_i) > 0$, $i = 1, \ldots, k$, and requires the predictions to be average calibrated within each subset. Adversarial group calibration then requires average calibration for *any subset of* $\mathcal{X}$ *with non-zero measure*. Denote $\boldsymbol{X}_{\mathcal{S}}$ as a random variable that is conditioned on being in the set $\mathcal{S}$. For **adversarial group calibration**, the observed probability is

$$\begin{aligned} p^{\text{obs}}_{adv}(p) :=& \quad \mathbb{E}_{x \sim \mathbb{F}_{\mathbf{X}_{\mathcal{S}}}}[\mathbb{F}_{\mathbf{Y}|x}(\hat{\mathbb{Q}}_p(x))], \\ \forall p \in (0, 1), \ \forall \mathcal{S} \subset \mathcal{X} \text{ s.t. } & P_{x \sim \mathbb{F}_{\mathbf{X}}}(x \in \mathcal{S}) > 0. \end{aligned} \tag{2}$$

With a finite dataset, we can measure a proxy of adversarial group calibration by measuring average calibration within all subsets of the data with sufficiently many points.

An alternative but widely used family of evaluation metrics is **proper scoring rules** (Gneiting and Raftery, 2007). Proper scoring rules are summary statistics of overall performance of a distributional prediction, and are defined such that the true underlying distribution optimizes the expectation of the scoring rule. Given a scoring rule $S(\hat{\mathbb{F}}, (x, y))$, where $x \sim \mathbb{F}_{\boldsymbol{X}}$, $y \sim \mathbb{F}_{\boldsymbol{Y}|x}$, the expectation of the scoring rule is $S(\hat{\mathbb{F}}, \mathbb{F}) = \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y}}[S(\hat{\mathbb{F}}, (x, y))]$, and S is said to be a proper scoring rule if $S(\mathbb{F}, \mathbb{F}) \geq S(\hat{\mathbb{F}}, \mathbb{F}), \forall \hat{\mathbb{F}} \in \mathcal{F}$.

There are a variety of proper scoring rules, based on the representation of the distributional prediction. Since these rules consider both calibration and sharpness together in a single value (Gneiting et al., 2007), they also serve as optimization objectives for UQ. For example, the logarithmic score is a popular proper scoring rule for density predictions (Lakshminarayanan et al., 2017; Pearce et al., 2018; Detlefsen et al., 2019), and it is used as a loss function via *negative log-likelihood* (**NLL**). The **check score** is widely used for quantile predictions and also known as the *pinball loss*. The **interval score** is commonly used for prediction intervals (a pair of quantiles with a prescribed expected coverage), and the continuous ranked probability score (**CRPS**) is popular for CDF predictions [1]. We refer the reader to Gneiting and Raftery (2007) for the definition of each scoring rule.

Given the wide range of metrics available, one might naturally ask, "is there one metric to rule them all?" Previous work has investigated some aspects of this question. For

[1] Proper scoring rules are usually *positively oriented* (i.e. greater value is more desirable), and their negative is taken as a loss function to minimize. In our work, we always report proper scoring rules in their *negative orientation* (i.e. as a loss).

example, Chung et al. (2020) noted the mismatch between the check score and average calibration, and Gneiting and Raftery (2007) and Bracher et al. (2021) point out cases in which disagreements can occur between some scoring rules. Still, whether there exists a golden metric in UQ is an open research problem. We instead suggest that there is virtue in inspecting various metrics simultaneously, which is made easy by the Uncertainty Toolbox, as we show below.

4. Case Study on Training, Evaluating PNNs

To demonstrate the capabilities of Uncertainty Toolbox, we provide a case study on training PNNs with various loss objectives, and use the Toolbox to examine results.

A PNN is a neural network that assumes a conditional Gaussian for the predictive distribution, thus for any input point x, outputs an estimate of the mean and the covariance, $\hat{\mu}(x)$, $\hat{\Sigma}(x)$. This NN structure has been proposed as early as Nix and Weigend (1994), but it has been popularized as a UQ method in deep learning by Lakshminarayanan et al. (2017), and it remains one of the most popular UQ method to date.

The standard method of training PNNs is to optimize the logarithmic score, i.e. NLL loss. However, based on the training principle, "optimize a proper score to improve UQ quality" (Lakshminarayanan et al., 2017) (also referred to as "optimum score estimation" by Gneiting and Raftery (2007)), we can in fact optimize many more proper scoring rules. In this study, we train PNNs using several different methods by optimizing with respect to either NLL, CRPS, check score, or interval score. Afterwards, we assess the predictive UQ quality with Uncertainty Toolbox. We summarize main details of the experiment below (full details in Appendix A).

Dataset The data was generated with a mean function $y = \sin(x/2) + x\cos(0.8x)$ and heteroscedastic Gaussian noise was added to generate the observations, y, for each input $x \sim \text{unif}[-10, 10]$. The train, validation and test splits consisted of $200, 100, 100$ points, respectively.

Training A separate model was trained for each loss function with full batch gradient decent and learning rate $1e^{-3}$, for 2000 epochs while tracking the validation loss. To optimize the check and interval scores, a batch of 30 expected probabilities $p_i \sim \text{unif}(0, 1)$ was selected and the scores for each p_i were summed to compute the loss (Tagasovska and Lopez-Paz, 2019; Chung et al., 2020). All reported results are based on the model with best validation loss.

Analysis We first visually observe UQ performance on the test set. Figure 2 (1) shows all of the methods approximately recovering the true level of heteroscedastic noise. Notably, NLL converges to a solution s.t. for $x < -5$, there is high error in mean estimation, which is compensated for with high (and wrong) variance estimation. The widths of the prediction intervals (PIs) in Figure 2 (2) also show how NLL is erroneously too wide. Meanwhile, comparisons with the ground truth PIs (far right plot) show that CRPS, Check and Interval all tend to have too narrow (sharp) predictions. This is further confirmed via the *Sharpness* metric column in Table 1, and we can also observe the ramifications in *average calibration* in Figure 2 (3): NLL's observed proportions in an interval tend to be greater than the expected proportion, signaling under-confidence (i.e. PIs that are too wide). The opposite case occurs for the other methods, and over-confidence (due to PIs that are too sharp) is especially pronounced in CRPS and Check. While NLL may seem average calibrated (with second lowest ECE), adversarial group calibration in Figure 1 shows that CRPS and Interval are better calibrated for smaller subsets of the domain, and achieve better adversarial group calibration.

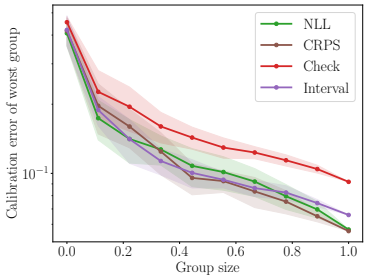

Figure 1. **Adversarial Group Calibration.** Group size refers to proportion of test dataset size, and the shades represent ± 1 standard error for the worst calibration error.

While the proper score metrics in Table 1 add another facet to the analysis, they also underscore the complex nature of assessing UQ. Each proper score has its own, separate ranking of the 4 methods, and they are also split on which one is best: Simply given this set of proper scoring rules, we believe it would be impossible to choose a single best method. Lastly, we note how a lower proper score may not necessarily indicate better calibration (Figure 2 (4)). Even while the proper scores improve on the test set (until around the validated epoch), calibration tends to get worse, while the predictions get sharper. Notably, CRPS and Check converge to a solution which is sharper than the true sharpness. This is problematic for calibration since a UQ sharper than the true sharpness will never be calibrated.

Conclusion This case study demonstrates that, even with numerous evaluation metrics at our disposal, the analysis of UQ for regression problems may not be straightforward. It also highlights limitations of the evaluation metrics, as relying on a single one (or small subset), may imply a conclusion counter to what other metrics signal. In the face of such limitations, we believe it is important to examine a suite of metrics simultaneously and perform a holistic evaluation of UQ quality. Not only does the Uncertainty Toolbox provide this functionality, but it also offers recalibration for pre-trained UQ models and resources giving key terms, explanations, and seminal works for those unfamiliar with the field. We hope that this toolbox is useful for accelerating and uniting efforts for uncertainty in machine learning.

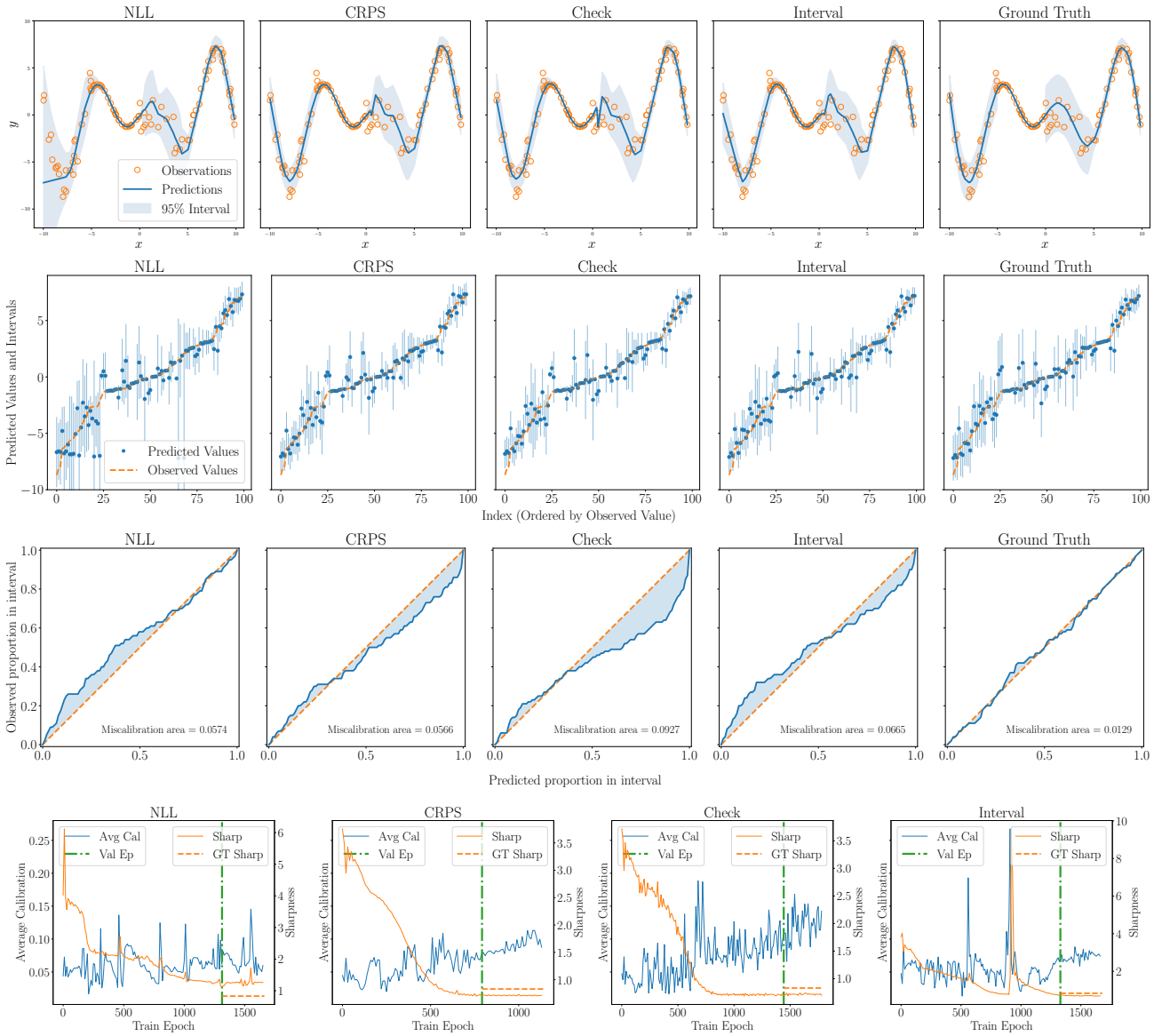

Figure 2. From top to bottom row, **(1)** Test observations, with the predicted mean and confidence bands **(2)** Test observations, the predicted mean and prediction interval, in order of test observations **(3)** Average calibration plot: predicted proportions (or expected probability) on x axis, observed proportions (or observed probability) on y axis **(4)** Training curves: average calibration (left y axis), sharpness (right y axis). GT Sharp denotes the true sharpness (noise level) of the data, and Val Ep denotes the epoch with lowest validation loss.

		Metrics							
		RMSE	MAE	ECE	Sharpness	NLL	CRPS	Check	Interval
Methods	NLL	1.689	0.852	0.057	1.451	2.214	0.604	0.305	2.990
	CRPS	**0.864**	0.568	**0.056**	0.729	1.266	**0.427**	**0.215**	2.323
	Check	0.880	**0.566**	0.092	**0.720**	4.264	0.434	0.219	2.434
	Interval	0.916	0.600	0.066	0.722	**0.780**	0.447	0.226	**2.309**
	Ground Truth	*0.824*	*0.530*	*0.013*	*0.831*	*−0.083*	*0.370*	*0.187*	*1.758*

Table 1. **Scalar Evaluation Metrics.** Each row shows evaluations metrics for a single method (i.e. loss function). RMSE (root mean squared error) and MAE (mean absolute error) are accuracy metrics. The best method for each metric is in **bold**. While these values are based on one seed, we show results across 5 random seed with standard error in Appendix B.

References

Ross Askanazi, Francis X Diebold, Frank Schorfheide, and Minchul Shin. On the comparison of interval forecasts. *Journal of Time Series Analysis*, 39(6):953–965, 2018.

VE Bowman, DS Silk, U Dalrymple, and DC Woods. Uncertainty quantification for epidemiological forecasts of covid-19 through combinations of model predictions. *arXiv preprint arXiv:2006.10714*, 2020.

Johannes Bracher, Evan L Ray, Tilmann Gneiting, and Nicholas G Reich. Evaluating epidemic forecasts in an interval format. *PLoS computational biology*, 17(2): e1008618, 2021.

Youngseog Chung, Willie Neiswanger, Ian Char, and Jeff Schneider. Beyond pinball loss: Quantile methods for calibrated uncertainty quantification. *arXiv preprint arXiv:2011.09588*, 2020.

Peng Cui, Wenbo Hu, and Jun Zhu. Calibrated reliable regression using maximum mean discrepancy. *Advances in Neural Information Processing Systems*, 33, 2020.

Nicki S Detlefsen, Martin Jørgensen, and Søren Hauberg. Reliable training and estimation of variance networks. *arXiv preprint arXiv:1906.03260*, 2019.

Josip Djolonga, Frances Hubis, Matthias Minderer, Zachary Nado, Jeremy Nixon, Rob Romijnders, Dustin Tran, and Mario Lucic. Robustness Metrics, 2020. URL https://github.com/google-research/robustness_metrics.

Matteo Fasiolo, Simon N Wood, Margaux Zaffran, Raphaël Nedellec, and Yannig Goude. Fast calibrated additive quantile regression. *Journal of the American Statistical Association*, pages 1–11, 2020.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1321–1330. JMLR. org, 2017.

Ursula Hébert-Johnson, Michael P Kim, Omer Reingold, and Guy N Rothblum. Calibration for the (computationally-identifiable) masses. *arXiv preprint arXiv:1711.08513*, 2017.

Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. *arXiv preprint arXiv:1807.00263*, 2018.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, pages 6402–6413, 2017.

Katarzyna Maciejowska, Jakub Nowotarski, and Rafał Weron. Probabilistic forecasting of electricity spot prices using factor quantile regression averaging. *International Journal of Forecasting*, 32(3):957–965, 2016.

Ali Malik, Volodymyr Kuleshov, Jiaming Song, Danny Nemer, Harlan Seymour, and Stefano Ermon. Calibrated model-based deep reinforcement learning. *arXiv preprint arXiv:1906.08312*, 2019.

Allan H Murphy. A new vector partition of the probability score. *Journal of applied Meteorology*, 12(4):595–600, 1973.

Zachary Nado, Neil Band, Mark Collier, Josip Djolonga, Michael Dusenberry, Sebastian Farquhar, Angelos Filos, Marton Havasi, Rodolphe Jenatton, Ghassen Jerfel, Jeremiah Liu, Zelda Mariet, Jeremy Nixon, Shreyas Padhy, Jie Ren, Tim Rudner, Yeming Wen, Florian Wenzel, Kevin Murphy, D. Sculley, Balaji Lakshminarayanan, Jasper Snoek, Yarin Gal, and Dustin Tran. Uncertainty Baselines: Benchmarks for uncertainty & robustness in deep learning. *arXiv preprint arXiv:2106.04015*, 2021.

David A Nix and Andreas S Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 ieee international conference on neural networks (ICNN'94)*, volume 1, pages 55–60. IEEE, 1994.

Tim Pearce, Felix Leibfried, Alexandra Brintrup, Mohamed Zaki, and Andy Neely. Uncertainty in neural networks: Approximately bayesian ensembling. *arXiv preprint arXiv:1810.05546*, 2018.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2015.

Hao Song, Tom Diethe, Meelis Kull, and Peter Flach. Distribution calibration for regression. In *International Conference on Machine Learning*, pages 5897–5906. PMLR, 2019.

Natasa Tagasovska and David Lopez-Paz. Single-model uncertainties for deep learning. In *Advances in Neural Information Processing Systems*, pages 6414–6425, 2019.

Kevin Tran, Willie Neiswanger, Junwoong Yoon, Qingyang Zhang, Eric Xing, and Zachary W Ulissi. Methods for comparing uncertainty quantifications for material property predictions. *Machine Learning: Science and Technology*, 1(2):025006, 2020.

Rebecca Wexler. When a computer program keeps you in jail: How computers are harming criminal justice. *New York Times*, 13, 2017.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *arXiv preprint arXiv:2005.13239*, 2020.

Shengjia Zhao, Tengyu Ma, and Stefano Ermon. Individual calibration with randomized forecasting. *arXiv preprint arXiv:2006.10288*, 2020.

Appendix

A. Details of the Case Study Experiment

A.1. Details on Dataset

The synthetic dataset in Section 4 was created with a mean function $y = \sin(x/2) + x\cos(0.8x)$ with $x \sim \text{unif}[-10, 10]$. The support $[-10, 10]$ was partitioned into 4 quadrants, and different levels of 0 mean, Gaussian noise was added to the mean function to create the y observations.

$$
\begin{aligned}
-10 \leq x < -5\text{: noise} &\sim \mathcal{N}(0, 1^2) \\
-5 \leq x < 0\text{: noise} &\sim \mathcal{N}(0, 0.01^2) \\
0 \leq x < 5\text{: noise} &\sim \mathcal{N}(0, 1.5^2) \\
5 \leq x \leq 10\text{: noise} &\sim \mathcal{N}(0, 0.5^2)
\end{aligned}
$$

A.2. Model Details

We used the same neural network architecture across all methods (i.e. loss functions): 3 layers of 64 hidden units with ReLU non-linearities, and 2 output units: one for the conditional mean $\hat{\mu}(x)$ and one for the conditional log-variance $\log \hat{\sigma}(x)$. We used the same learning rate $1e^{-3}$ and full batch size (200) for all methods. During training, we track the corresponding loss function on the validation set, and at the end of 2000 epochs, the final model was backtracked to the model with lowest validation loss. All reported test results are based on this backtracked model.

A.3. Calculation of Evaluation Metrics

This section describes how each of the reported metrics are computed within Uncertainty Toolbox, given a finite dataset $D = \{(x_i, y_i)_{i=1}^N\}$.

Accuracy Metrics
The root mean squared error (RMSE) and mean absolute error (MAE) are computed with the mean prediction $\hat{\mu}(x)$, following the standard definitions.

$$
\text{RMSE}(D, \hat{\mu}) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\mu}(x_i))^2}
$$

$$
\text{MAE}(D, \hat{\mu}) = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{\mu}(x_i)|
$$

Calibration Metrics
To measure the calibration metrics (average calibration, adversarial group calibration), expected probabilities are discretized from 0.01 to 0.99 in 0.01 increments (i.e. 0.01, 0.02, $\ldots$, 0.97, 0.98, 0.99). and the observed probabilities are calculated for each of these 99 expected probabilities.

ECE (measure of average calibration) is computed following the definition given in Section 3.

The procedure in which we measure adversarial group calibration is the following. For a given test set, we scale group size between 1% and 100% of the full test set size, in 10 equi-spaced intervals. With each group size, we draw 20 random groups from the test set and record the worst calibration incurred across these 20 random groups. The adversarial group calibration figure (Figure 1) plots the mean worst calibration incurred with ± 1 standard error in shades, for each group size. This is also the method used by Zhao et al. (2020) to measure adversarial group calibration.

Sharpness
Sharpness is measured as the mean of the standard deviation predictions on the test set. Note that sharpness is a property of the prediction *only*, and does not take into consideration the true distribution.

Proper Scoring Rules
The proper scoring rules (NLL, CRPS, check score, interval score) are measured as the mean of the score on the test set.

B. Numerical Results Across Multiple Trials

The results presented in Section 4 are based on one random seed. Below, we present the numerical results across 5 random seeds: $[0, 1, 2, 3, 4]$.

		Metrics			
		RMSE	MAE	ECE	Sharpness
Methods	NLL	2.048 ± 0.125	1.073 ± 0.080	$\mathbf{0.029} \pm 0.007$	1.746 ± 0.155
	CRPS	$\mathbf{1.023} \pm 0.090$	$\mathbf{0.661} \pm 0.054$	0.044 ± 0.005	0.897 ± 0.114
	Check	1.045 ± 0.105	0.672 ± 0.065	0.050 ± 0.011	$\mathbf{0.874} \pm 0.117$
	Interval	1.169 ± 0.187	0.745 ± 0.101	0.039 ± 0.009	0.915 ± 0.130
	Ground Truth	$\mathit{0.962 \pm 0.064}$	$\mathit{0.618 \pm 0.042}$	$\mathit{0.019 \pm 0.002}$	$\mathit{0.925 \pm 0.052}$

		Metrics			
		NLL	CRPS	Check	Interval
Methods	NLL	1.677 ± 0.343	0.766 ± 0.060	0.386 ± 0.030	3.885 ± 0.330
	CRPS	1.112 ± 0.111	$\mathbf{0.492} \pm 0.040$	$\mathbf{0.248} \pm 0.020$	$\mathbf{2.687} \pm 0.186$
	Check	1.635 ± 0.661	0.501 ± 0.048	0.253 ± 0.024	2.741 ± 0.224
	Interval	$\mathbf{0.961} \pm 0.062$	0.546 ± 0.073	0.276 ± 0.037	2.875 ± 0.352
	Ground Truth	$\mathit{0.187 \pm 0.115}$	$\mathit{0.435 \pm 0.033}$	$\mathit{0.219 \pm 0.017}$	$\mathit{2.122 \pm 0.177}$

Table 2. **Scalar Evaluation Metrics.** Each row shows evaluation metrics for a single method (i.e. loss function), and the mean with ± 1 standard error is shown. The best mean for each metric has been **bolded**.