entropy

MDPI

Article

Empirical Frequentist Coverage of Deep Learning Uncertainty Quantification Procedures

Benjamin Kompa [1], Jasper Snoek [2] and Andrew L. Beam [1,3,*]

[1] Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA; benjamin_kompa@hms.harvard.edu
[2] Google Research, Cambridge, MA 02142, USA; jsnoek@google.com
[3] Department of Epidemiology, Harvard School of Public Health, Boston, MA 02115, USA
* Correspondence: andrew_beam@hms.harvard.edu

Abstract: Uncertainty quantification for complex deep learning models is increasingly important as these techniques see growing use in high-stakes, real-world settings. Currently, the quality of a model's uncertainty is evaluated using point-prediction metrics, such as the negative log-likelihood (NLL), expected calibration error (ECE) or the Brier score on held-out data. Marginal coverage of prediction intervals or sets, a well-known concept in the statistical literature, is an intuitive alternative to these metrics but has yet to be systematically studied for many popular uncertainty quantification techniques for deep learning models. With marginal coverage and the complementary notion of the width of a prediction interval, downstream users of deployed machine learning models can better understand uncertainty quantification both on a global dataset level and on a per-sample basis. In this study, we provide the first large-scale evaluation of the empirical frequentist coverage properties of well-known uncertainty quantification techniques on a suite of regression and classification tasks. We find that, in general, some methods do achieve desirable coverage properties on *in distribution* samples, but that coverage is not maintained on out-of-distribution data. Our results demonstrate the failings of current uncertainty quantification techniques as dataset shift increases and reinforce coverage as an important metric in developing models for real-world applications.

Keywords: uncertainty quantification; coverage; Bayesian methods; dataset shift

Citation: Kompa, B.; Snoek, J.; Beam, A.L. Empirical Frequentist Coverage of Deep Learning Uncertainty Quantification Procedures. *Entropy* **2021**, *23*, 1608. https://doi.org/10.3390/e23121608

Academic Editors: Eric Nalisnick and Dustin Tran

Received: 1 October 2021
Accepted: 24 November 2021
Published: 30 November 2021

1. Introduction

Predictive models based on deep learning have seen a dramatic improvement in recent years [1], which has led to widespread adoption in many areas. For critical, high-stakes domains, such as medicine or self-driving cars, it is imperative that mechanisms are in place to ensure safe and reliable operation. Crucial to the notion of safe and reliable deep learning is the effective quantification and communication of predictive uncertainty to potential end-users of a system. In medicine, for instance, understanding predictive uncertainty could lead to better decision-making through improved allocation of hospital resources, detecting dataset shift in deployed algorithms, or helping machine learning models abstain from making a prediction [2]. For medical classification problems involving many possible labels (i.e., creating a differential diagnosis), methods that provide a set of possible diagnoses when uncertain are natural to consider and align more closely with the differential diagnosis procedure used by physicians. The prediction sets and intervals we propose in this work are an intuitive way to quantify uncertainty in machine learning models and provide interpretable metrics for downstream, nontechnical users.

Commonly used approaches to quantify uncertainty in deep learning generally fall into two broad categories: ensembles and approximate Bayesian methods. Deep ensembles [3] aggregate information from multiple individual models to provide a measure of uncertainty that reflects the ensembles' agreement about a given data point. Bayesian

methods offer direct access to predictive uncertainty through the posterior predictive distribution, which combines prior knowledge with the observed data. Although conceptually elegant, calculating exact posteriors of even simple neural models is computationally intractable [4,5], and many approximations have been developed [6–12]. Though approximate Bayesian methods scale to modern sized data and models, recent work has questioned the quality of the uncertainty provided by these approximations [4,13,14].

Previous work assessing the quality of uncertainty estimates has focused on calibration metrics and scoring rules, such as the negative log-likelihood (NLL), expected calibration error (ECE), and Brier score. Here we provide an alternative perspective based on the notion of empirical coverage, a well-established concept in the statistical literature [15] that evaluates the quality of a predictive set or interval instead of a point prediction. Informally, coverage asks the question: If a model produces a predictive uncertainty interval, how often does that interval actually contain the observed value? Ideally, predictions on examples for which a model is uncertain would produce larger intervals and thus be more likely to cover the observed value.

In this work, we focus on marginal coverage *over a dataset* $\mathcal{D}'$ for the canonical α value of 0.05, i.e., 95% prediction intervals. For a machine learning model that produces a 95% prediction interval $\hat{C}_n(x_n)$ based on the training dataset $\mathcal{D}$, we consider what fraction of the points in the dataset $\mathcal{D}'$ have their true label contained in $\hat{C}_n(x_{n+1})$ for $x_{n+1} \in \mathcal{D}'$. To measure the robustness of these intervals, we also consider cases when the generating distributions for $\mathcal{D}$ and $\mathcal{D}'$ are not the same (i.e., dataset shift).

Figure 1 provides a visual depiction of marginal coverage over a dataset for two hypothetical regression models. Throughout this work, we refer to "marginal coverage over a dataset" as "coverage".

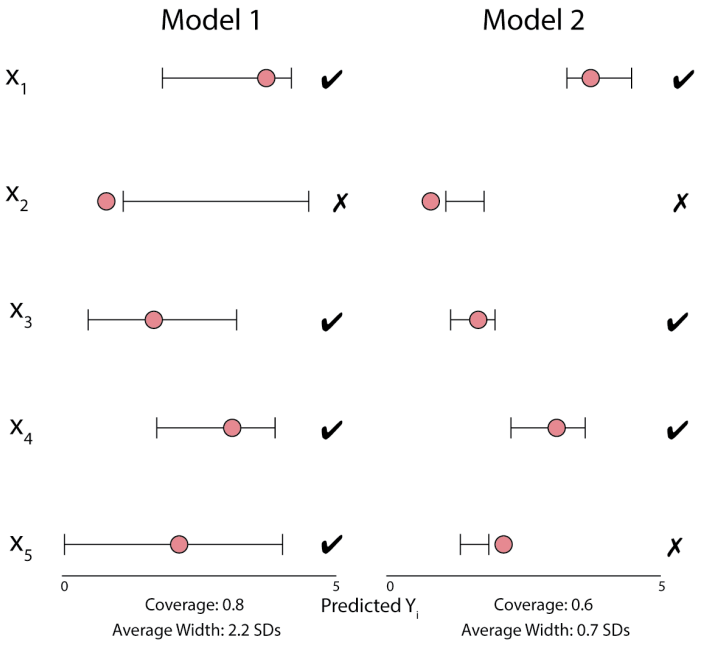


Figure 1. An example of the coverage properties for two methods of uncertainty quantification. In this scenario, each model produces an uncertainty interval for each x_i, which attempts to cover the true y_i, represented by the red points. Coverage is calculated as the fraction of true values contained in these regions, while the width of these regions is reported in terms of multiples of the standard deviation of the training set y_i values.

For a machine learning model that produces predictive uncertainty estimates (i.e., approximate Bayesian methods and ensembling), coverage encompasses both the aleatoric

and epistemic uncertainties [16] produced by these models. In a regression setting, the predictions from these models can be written as:

$$\hat{y} = f(x) + \epsilon \tag{1}$$

where epistemic uncertainty is captured in the $f(x)$ component, while aleatoric uncertainty is considered in the ϵ term. Since coverage captures how often the predicted interval of $\hat{y}$ contains the true value, it captures the contributions from both types of uncertainty.

A complementary metric to coverage is width, which is the size of the prediction interval or set. In regression problems, we typically measure width in terms of the standard deviation of the true label in the training set. As an example, an uncertainty quantification procedure could produce prediction intervals that have 90% marginal coverage with an average width of two standard deviations. For classification problems, width is simply the average size of a prediction set. Width can provide a relative ranking of different methods, i.e., given two methods with the same level of coverage, we should prefer the method that provides intervals with smaller widths.

Contributions: In this study, we investigate the empirical coverage properties of prediction intervals constructed from a catalog of popular uncertainty quantification techniques, such as ensembling, Monte Carlo dropout, Gaussian processes, and stochastic variational inference. We assess the coverage properties of these methods on nine regression tasks and two classification tasks with and without dataset shift. These tasks help us make the following contributions:

- We introduce coverage and width over a dataset as natural and interpretable metrics for evaluating predictive uncertainty for deep learning models.
- A comprehensive set of coverage evaluations on a suite of popular uncertainty quantification techniques.
- An examination of how dataset shift affects these coverage properties.

2. Background and Related Work

2.1. Frequentist Coverage and Conformal Inference

Given features $x_i \in \mathbb{R}^d$ and a response $y_i \in \mathbb{R}$ for some dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, Barber et al. [17] define *distribution-free* marginal coverage in terms of a set $\hat{\mathcal{C}}_n(x)$ and a level $\alpha \in [0, 1]$. The set $\hat{\mathcal{C}}_n(x)$ is said to have coverage at the $1 - \alpha$ level if for all distributions P such that $(x, y) \in \mathbb{R}^d \times \mathbb{R}$ and $(x, y) \sim P$, the following inequality holds:

$$\mathbb{P}\{y_{n+1} \in \hat{\mathcal{C}}_n(x_{n+1})\} \geq 1 - \alpha \tag{2}$$

For new samples beyond the first n samples in the training data, there is a $1 - \alpha$ probability of the true label of the test point being contained in the set $\hat{\mathcal{C}}_n(x_{n+1})$. This set can be constructed using a variety of procedures. For example, in the case of simple linear regression, a prediction interval for a new point x_{n+1} can be constructed using a simple, closed-form solution [15].

Marginal coverage is typically considered in the limit of infinite samples. However, here we focus on marginal coverage over a dataset $\mathcal{D}$. We assess, for a given model and test set $\mathcal{D}$, the empirical coverage by assessing whether $y_{n+1} \in \hat{\mathcal{C}}_n(x_{n+1}) \; \forall x_{n+1} \in \mathcal{D}$. Additionally, we consider how marginal coverage changes as there is data distribution shift such that a new dataset $\mathcal{D}'$ has a different data generating distribution. Despite the lack of infinite samples, this work establishes the motivation of considering coverage in critical, high-risk situations, such as medicine.

An important and often overlooked distinction is that of marginal and conditional coverage. In conditional coverage, one considers:

$$\mathbb{P}\{y_{n+1} \in \hat{\mathcal{C}}_n(x_{n+1}) | x_{n+1} = x\} \geq 1 - \alpha \tag{3}$$

The probability has been conditioned on specific features. This is potentially a more useful version of coverage to consider because one could make claims for specific instances rather than over the broader distribution P. However, it is impossible in general to have conditional coverage guarantees [17].

Conformal inference [18,19] is one statistical framework that can provide marginal coverage under a certain set of assumptions (e.g., exchangeable data) that we do not assume here [20]. In this work, we specifically seek to measure the empirical coverage of the existing approximate Bayesian and alternative uncertainty quantification methods with and without dataset shift. These methods are extremely popular in practice, but nobody has yet considered the empirical coverage of their 95% posteriors. Conformal methods are not part of the approximate Bayesian methods that we set out to analyze in this work. There has been recent work on Bayes-optimal prediction with frequentist coverage control [21] and conformal inference under dataset shift [22,23]. However, adding the conformal framework to approximate Bayesian methods post hoc and measuring their coverage properties could be interesting future work. An additional distinction between our work and the broader conformal inference literature is that we do not aim to provide finite sample coverage guarantees.

Another important point to consider is that while the notion of a confidence interval may seem natural to consider in our analysis, confidence intervals estimate global statistics over repeated trials of data and generally come with guarantees about how often these statistics lie in said intervals. In our study, this is not the case. Although we estimate coverage across many datasets, we are not aiming to estimate an unknown statistic of the data. We would like to understand the empirical coverage properties of machine learning models.

2.2. Obtaining Predictive Uncertainty Estimates

Several lines of work focus on improving approximations of the posterior of a Bayesian neural network [6–12]. Yao et al. [4] provide a comparison of many of these methods and highlight issues with common metrics of comparison, such as test-set log-likelihood and RMSE. Good scores on these metrics often indicate that the model posterior happens to match the test data rather than the true posterior [4]. Maddox et al. [24] developed a technique to sample the approximate posterior from the first moment of stochastic gradient descent iterates. Wenzel et al. [13] demonstrated that despite advances in these approximations, in practice, approximate methods for the Bayesian modeling of deep networks do not perform as well as theory would suggest.

Alternative methods that do not rely on estimating a posterior over the weights of a model can also be used to provide uncertainty estimates. Gal and Ghahramani [16], for instance, demonstrated that Monte Carlo dropout is related to a variational approximation to the Bayesian posterior implied by the dropout procedure. Lakshminarayanan et al. [3] used an ensemble of several neural networks to obtain uncertainty estimates. Guo et al. [25] established that temperature scaling provides well-calibrated predictions on an i.i.d test set. More recently, van Amersfoort et al. [26] showed that the distance from the centroids in an RBF neural network yields high-quality uncertainty estimates. Liu et al. [27] also leveraged the notion of distance (in the form of an approximate Gaussian process covariance function) to obtain uncertainty estimates with their Spectral-normalized Neural Gaussian Processes.

2.3. Assessments of Uncertainty Properties under Dataset Shift

Ovadia et al. [14] analyzed the effect of dataset shift on the accuracy and calibration of a variety of deep learning methods. Their large-scale empirical study assessed these methods on standard datasets, such as MNIST, CIFAR-10, ImageNet, and other non-image-based datasets. Additionally, they used translations, rotations, and corruptions of these datasets [28] to quantify performance under dataset shift. They found stochastic variational inference (SVI) to be promising on simpler datasets, such as MNIST and CIFAR-10, but

more difficult to train on larger datasets. Deep ensembles had the most robust response to dataset shift.

3. Methods

For features $x_i \in \mathbb{R}^d$ and a response $y_i \in \mathbb{R}$ or $y_i \in \mathbb{Z}$ (for regression and classification, respectively) for some dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$, we consider the prediction intervals or sets $\hat{\mathcal{C}}_n(x)$ in regression and classification settings, respectively. Unlike in the definitions of marginal and conditional coverage, we do not assume that $(x, y) \sim P$ always holds true. Thus, we consider the marginal coverage on a dataset $\mathcal{D}'$ for some new test sets that may have undergone dataset shift from the generating distribution of the training set $\mathcal{D}$.

In both the regression and classification settings, we analyzed the coverage properties of prediction intervals and sets of five different approximate Bayesian and non-Bayesian approaches for uncertainty quantification. These include dropout [16,29], ensembles [3], Stochastic Variational Inference [7,8,11,12,30], and last layer approximations of SVI and dropout [31]. Additionally, we considered prediction intervals from linear regression and the 95% credible interval of a Gaussian process with the squared exponential kernel as baselines in regression tasks. For classification, we also considered temperature scaling [25] and the softmax output of vanilla deep networks [28]. For more detail on our modeling choices, see Appendix B.

3.1. Regression Methods and Metrics

We evaluated the coverage properties of these methods on nine large real-world regression datasets used as a benchmark in Hernández-Lobato and Adams [6] and later Gal and Ghahramani [16]. We used the training, validation, and testing splits publicly available from Gal and Ghahramani [16] and performed nested cross-validation to find hyperparameters. On the training sets, we did 100 trials of a random search over hyperparameter space of a multi-layer-perceptron architecture with an Adam optimizer [32] and selected hyperparameters based on RMSE on the validation set.

Each approach required slightly different ways to obtain a 95% prediction interval. For an ensemble of neural networks, we trained $N = 40$ vanilla networks and used the 2.5% and 97.5% quantiles as the boundaries of the prediction interval. For dropout and last layer dropout, we made 200 predictions per sample and similarly discarded the top and bottom 2.5% quantiles. For SVI, last layer SVI (LL SVI), and Gaussian processes we had approximate variances available for the posterior, which we used to calculate the prediction interval. We calculated 95% prediction intervals from linear regression using the closed-form solution.

Then we calculated two metrics:

- **Coverage**: A sample is considered covered if the true label is contained in this 95% prediction interval. We average over all samples in a test set to estimate a method's marginal coverage on this dataset.
- **Width**: The width is the average over the test set of the ranges of the 95% prediction intervals.

Coverage measures how often the true label is in the prediction region, while width measures how specific that prediction region is. Ideally, we would have high levels of coverage with low levels of width on in-distribution data. As data becomes increasingly out of distribution, we would like coverage to remain high while width increases to indicate model uncertainty.

3.2. Classification Methods and Metrics

Ovadia et al. [14] evaluated model uncertainty on a variety of datasets publicly available. These predictions were made with the five approximate Bayesian methods described above, plus vanilla neural networks, with and without temperature scaling. We focus on the predictions from MNIST, CIFAR-10, CIFAR-10-C, ImageNet, and ImageNet-C datasets. For MNIST, we calculated coverage and width of model prediction intervals on

rotated and translated versions of the test set. For CIFAR-10, Ovadia et al. [14] measured model predictions on translated and corrupted versions of the test set from CIFAR-10-C [28] (see Figure 2). For ImageNet, we only considered the coverage and width of prediction sets on the corrupted images of ImageNet-C [28]. Each of these transformations (rotation, translation, or any of the 16 corruptions) has multiple levels of shift. Rotations range from 15 to 180 degrees in 15 degrees increments. Translations shift images every 2 and 4 pixels for MNIST and CIFAR-10, respectively (see Figure 3). Corruptions have five increasing levels of intensity. Figure 2 shows the effects of the 16 corruptions in CIFAR-10-C at the first, third, and fifth levels of intensity.

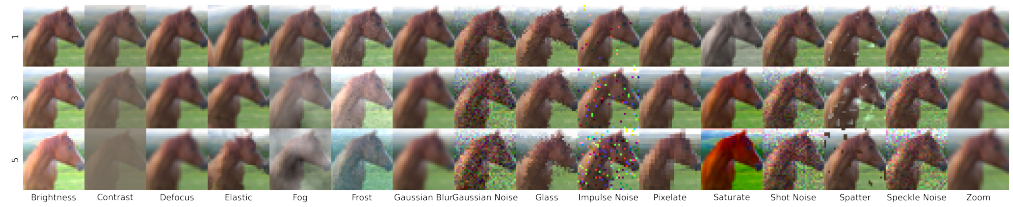


Figure 2. An example of the corruptions in CIFAR-10-C from [28]. The 16 different corruptions have 5 discrete levels of shift, of which 3 are shown here. The same corruptions were applied to ImageNet to form the ImageNet-C dataset.

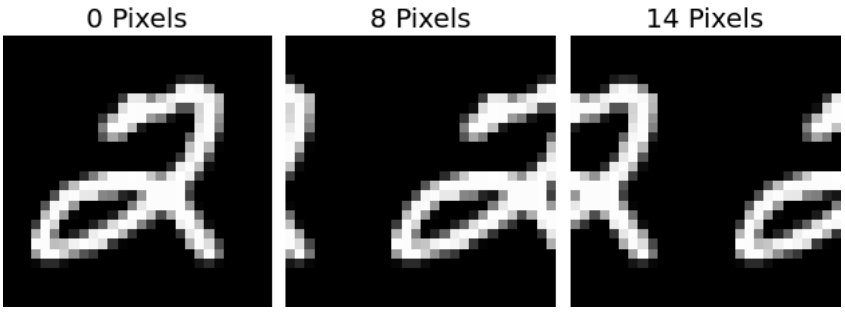


Figure 3. Several examples of the "rolling" translation shift that moves an image across an axis.

Given $\alpha \in (0, 1)$ and predicted probabilities $p(y_c|x_i)$ from a model for all K classes $c \in \{1, ..., K\}$, the $1 - \alpha$ prediction set $\mathcal{S}$ for a sample x_i is the minimum sized set of classes such that:

$$\sum_{c \in \mathcal{S}} p(y_c|x_i) \geq 1 - \alpha \tag{4}$$

This results in a set of size k_i, which consists of the largest probabilities in the full probability distribution over all classes $p(y_c|x_i)$ such that $1 - \alpha$ probability has been accumulated. This inherently assumes that the labels are unordered categorical classes such that including classes 1 and K does not imply that all classes between are also included in the set $\mathcal{S}$. Then we can define:

- **Coverage:** For each example in a dataset, we calculate the $1 - \alpha$ prediction set of the label probabilities, then coverage is what fraction of these prediction sets contain the true label.
- **Width:** The width of a prediction set is simply the number of labels in the set, $|\mathcal{S}|$. We report the average width of prediction sets over a dataset in our figures.

Although both calibration [25] and coverage can involve a probability over a model's output, calibration only considers the most likely label, and its corresponding probability, while coverage considers the top-k_i probabilities. In the classification setting, coverage is more robust to label errors as it does not penalize models for putting probability on similar classes.

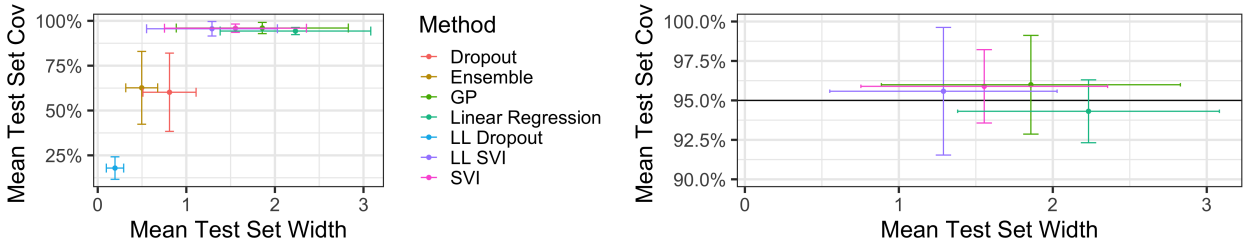
4. Results

4.1. Regression

Figure 4 plots the mean test set coverage and width for the regression methods we considered averaged over the nine regression datasets. Error bars demonstrate that for low-performing methods, such as ensembling, dropout, and LL dropout, there is high variability in coverage levels and widths across the datasets.

We observe that several methods perform well across the nine datasets. In particular, LL SVI, SVI, and GPs all exceed the 95% coverage threshold on average, and linear regression comes within the statistical sampling error of this threshold. Over the regression datasets, we considered, LL SVI had the lowest mean width while maintaining at least 95% coverage. For specific values of coverage and width for methods on a particular dataset, see Tables A1 and A2 in Appendix A.

Figure 4 also demonstrates an important point that will persist through our results. Coverage and width are directly related. Although high coverage can and ideally does occur when width is low, we typically observe that high levels of coverage occur in conjunction with high levels of width.



(**a**) Coverage and width across datasets (**b**) Detailed view

Figure 4. The mean coverage and widths of models' prediction intervals average over the nine regression datasets we considered (panel **a**). Error bars indicate the standard deviation for both coverage and width across all experiments. In general, one would desire a model with the highest coverage above some threshold (here 95%) with a minimum average test set width. Models in the upper left have the best empirical coverage. In (panel **b**), we observe that the four methods which maintained 95% coverage did so because they had appropriately wide prediction intervals. LL SVI had the lowest average width while maintaining at least 95% coverage.

4.2. MNIST

In the classification setting, we begin by calculating coverage and width for predictions from Ovadia et al. [14] on MNIST and shifted MNIST data. Ovadia et al. [14] used a LeNet architecture, and we refer to their manuscript for more details on their implementation.

Figure 5 shows how coverage and width co-vary as dataset shift increases. The elevated width for SVI on these dataset splits indicate that the posterior predictions of label probabilities were the most diffuse to begin with among all models. In Figure 5, all seven models have at least 95% coverage with a 15-degree rotation shift. Most models do not see an appreciable increase in the average width of the 95% prediction set, except for SVI. The average width for SVI jumps to over 2 at 15 degrees rotation. As the amount of shift increases, coverage decreases across all methods in a comparable way. In the rotation shifts, we observe that coverage increases and width decreases after about 120 degrees of shift. This is likely due to some of the natural symmetry of several digits (i.e., 0 and 8 look identical after 180 degrees of rotation).

SVI maintains higher levels of coverage but with a compensatory increase in width. In fact, there is a Pearson correlation of 0.9 between the width of the SVI prediction set and the distance from the maximum shift of 14 pixels. The maximum shift occurs when the original center of the image is broken across the edge as the image rolls to the right. Figure 3's right-most example is a case of the maximum shift of 14 pixels on a MNIST digit. This strong correlation between width and severity of shift for some methods makes the width of a prediction set at a fixed α level a natural proxy to detect dataset shift. For

this simple dataset, SVI outperforms other models with regards to coverage and width properties. It is the only model that has an average width that corresponds to the amount of shift observed and provides the highest level of average coverage.

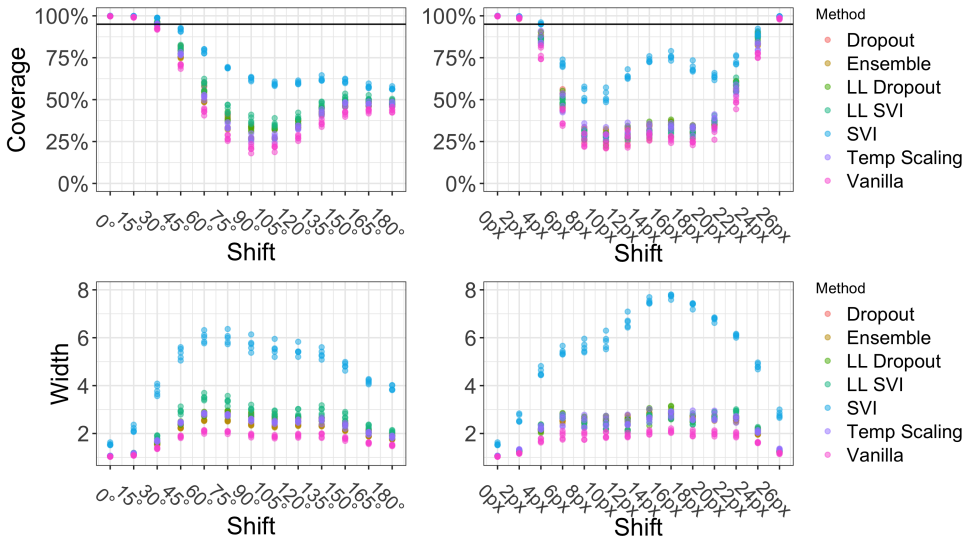


Figure 5. The effect of rotation and translation on coverage and width, respectively, for MNIST. 0 degrees or 0 pixels of shift indicate results on the test set of MNIST.

4.3. CIFAR-10

Next, we consider a more complex image dataset, CIFAR-10. Ovadia et al. [14] trained 20-layer and 50-layer ResNets. Figure 6 shows how the width of the prediction sets increases as the translation shift increases. This shift "rolls" the image pixel by pixel such that the right-most column in the image becomes the left-most image. Temperature scaling and ensemble, in particular, have at least 95% coverage for every translation, although all methods have high levels of coverage on average (though not exceeding 95%). We find that this high coverage comes with increases in width as shift increases. Figure 6 shows that temperature scaling has the highest average width across all models and shifts. Ensembling has the lowest width for the methods that maintain coverage of at least 95% across all shifts.

All models have the same encouraging pattern of width increasing as shift increases up to 16 pixels, then decreasing. As CIFAR-10 images are 28 pixels in width and height, this maximum width occurs when the original center of the image is rolled over to and broken by the edge of the image. This likely breaks common features that the methods have learned for classification onto both sides of the image, resulting in decreased classification accuracy and higher levels of uncertainty.

Between the models which satisfy 95% coverage levels on all shifts, ensemble models have lower width than temperature scaling models. Under translation shifts on CIFAR-10, ensemble methods perform the best given their high coverage and lower width.

Additionally, we consider the coverage properties of models on 16 different corruptions of CIFAR-10 from Hendrycks and Gimpel [28]. Figure 7 shows coverage vs. width over varying levels of shift intensity. Models that have more dispersed points to the right have higher widths for the same level of coverage. An ideal model would have a cluster of points above the 95% coverage line and be far to the left portion of each facet. For models that have similar levels of coverage, the superior method will have points further to the left.

Figure 7 demonstrates that at the lowest shift intensity, ensemble models, dropout, temperature scaling, and SVI were able to generally provide high levels of coverage on most corruption types. However, as the intensity of the shift increases, coverage decreases. Ensembles and dropout models have, for at least half of their 80 model-corruption evaluations, at least 95% coverage up to the third intensity level. At higher levels of shift

intensity, ensembles, dropout, and temperature scaling consistently have the highest levels of coverage. Although these higher-performing methods have similar levels of coverage, they have different widths.

We also present a way to quantify the relative strength of each method over a specific level of corruption. In Figure 8, for instance, we plot only the coverage and widths of methods at the third level of corruption and use the fraction of the points of a particular method that lie above the regression line. Methods that are more effective are providing higher coverage levels at lower widths and will have more points above this regression line.

For each of the five corruption levels, we calculated a regression line that modeled coverage as a function of width. Figure 9 presents the fraction of marginal coverages on various CIFAR-10-C datasets for each method that exceeded the linear regression prediction. The larger the fraction, the better the marginal coverage of a method given a prediction interval/set of a particular width. We observe that dropout and ensembles have a strong relative performance to the other methods across all five levels of shift.

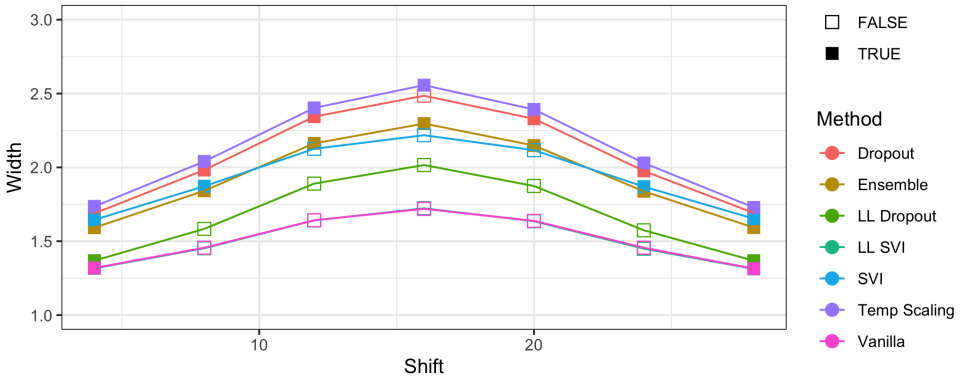


Figure 6. The effect of translation shifts on coverage and width in CIFAR-10 images. Coverage remains robust across all pixel shifts while width increases. The shading of points indicates whether 95% coverage was maintained when translated. In general, models with every point shaded maintain high levels of coverage. Therefore, the models with the best empirical coverage properties are the lowest width models such that coverage is maintained.

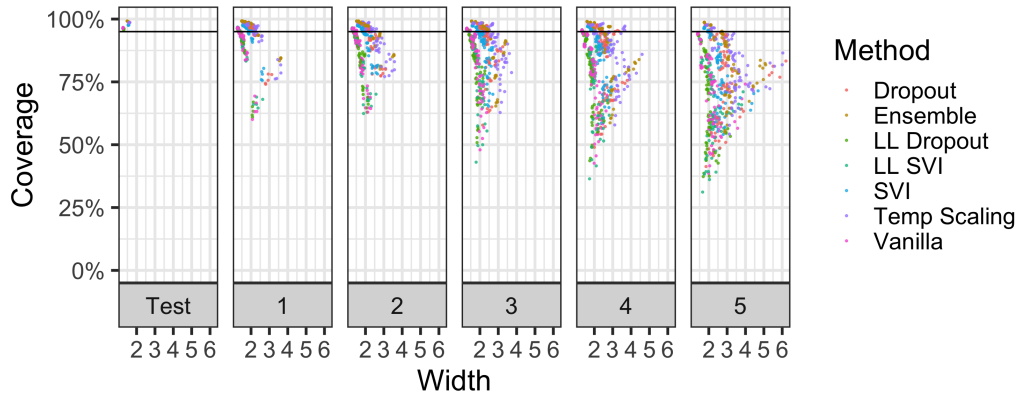


Figure 7. The effect of corruption intensity on coverage levels vs. width in CIFAR-10-C. Each facet panel represents a different corruption level, while points are the coverage of a model on one of 16 corruptions. Each facet has 80 points per method since 5 iterations were trained per method. For methods with points at the same coverage level, the superior method is to the left as it has a lower width. Please see Figures 8 and 9 for the additional synthesis of these results.

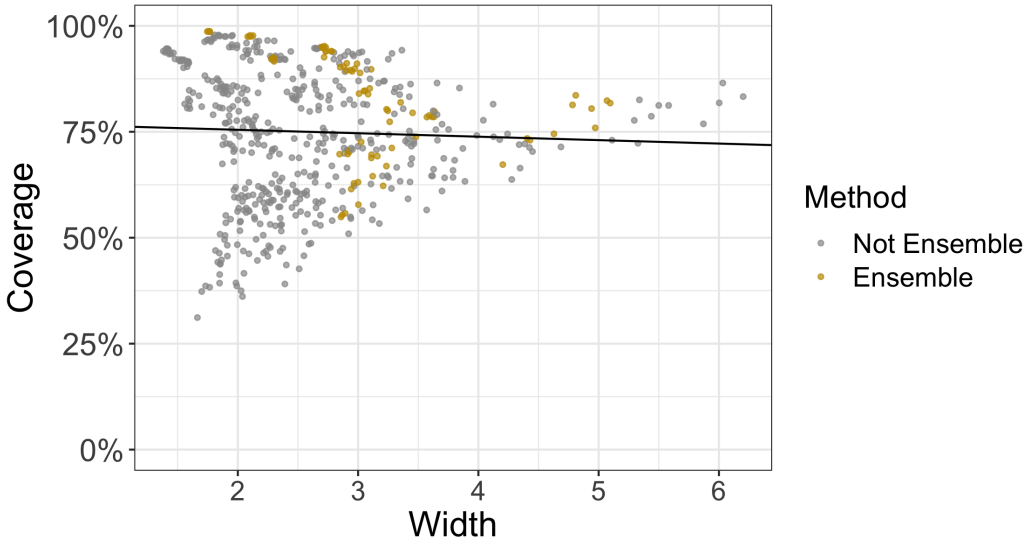


Figure 8. The coverage and width of ensemble and non-ensemble methods at the fifth level out of five levels of corruption in CIFAR-10-C. The black line is a simple linear regression of coverage against width. We then can consider the fraction of points for a particular method (in this case, ensembling) that are above the regression line (see Figures 9 and 10). The higher the fraction of these points above the regression line, the better the method is at providing higher coverage at a relatively smaller width than other methods.

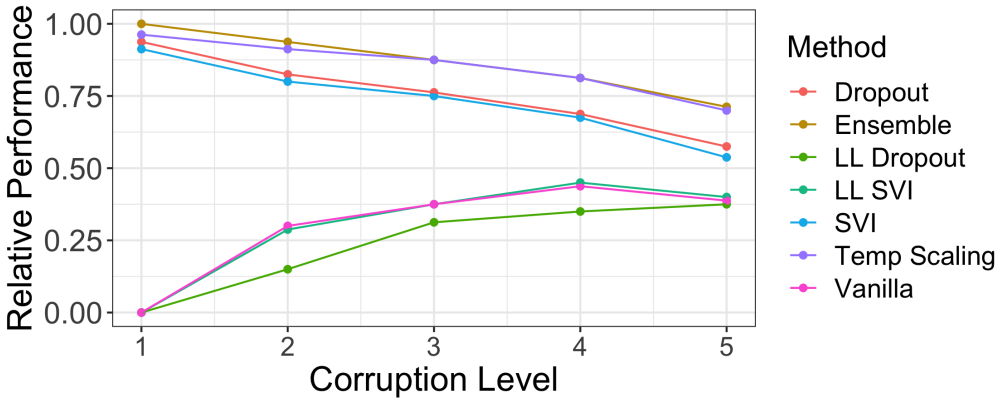


Figure 9. The fraction of marginal coverage levels achieved on CIFAR-10-C corruptions by our assessed methods that are above a regression line of coverage vs. width at a specific corruption level. Methods that have better coverage levels at the same width will have a higher fraction of points above the regression line (see Figure 8 for an example). At low levels of shift, dropout, ensemble, SVI, and temperature scaling have strictly better relative performance. As shift increases, poor coverage levels in general cause models to have more parity.

Finally, we compared the relative rank order of these methods across coverage, as well as two common metrics in uncertainty quantification literature: Brier score and ECE. Figure 11 shows that the rankings are similar across methods. In particular, coverage has a nearly identical pattern to ECE, with changes only in the lower ranking methods.

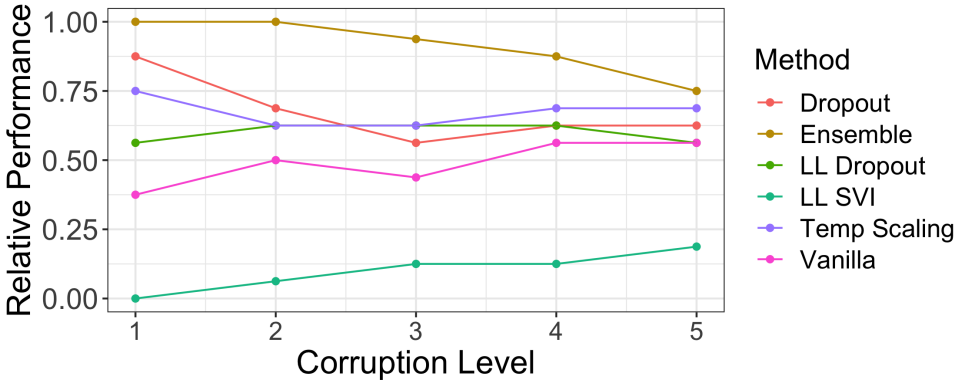


Figure 10. The fraction of marginal coverage levels achieved on ImageNet-C corruptions by our assessed methods that are above a regression line of coverage vs. width at a specific corruption level. Methods that have better coverage levels at the same width will have a higher fraction of points above the regression line (see Figure 8 for an example). Ensembling produces the best coverage levels given specific widths across all levels of corruption. However, at a higher level of dataset shift, there is more parity between methods.

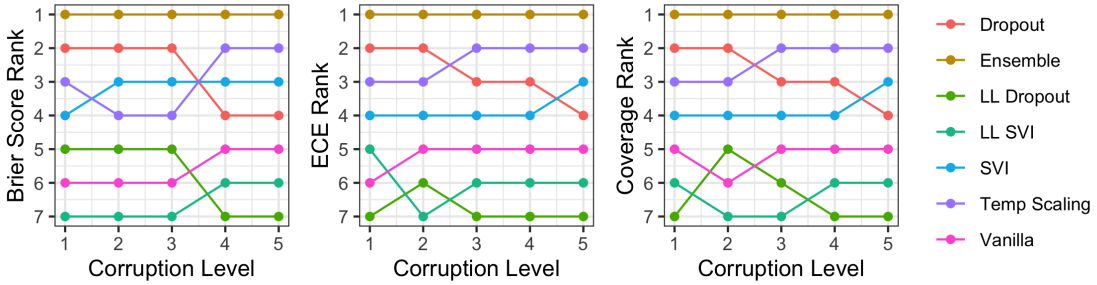


Figure 11. The ranks of each method's performance with respect to each metric we consider on CIFAR-10-C. For Brier Score and ECE, lower is better, while for coverage, higher is better. We observe that all three metrics have a generally consistent ordering, with coverage closely corresponding to the rankings of ECE.

4.4. ImageNet

Finally, we analyze coverage and width on ImageNet and ImageNet-C from Hendrycks and Gimpel [28]. Figure A1 shows similar coverage vs. width plots to Figure 7. We find that over the 16 different corruptions at 5 levels, ensembles, temperature scaling, and dropout models had consistently higher levels of coverage. Unsurprisingly, Figure A1 shows that these methods have correspondingly higher widths. Figure 10 reports the relative performance of each method across corruption levels. Ensembles had the highest fraction of marginal coverage on ImageNet-C datasets above the regression lines at each corruption level. Dropout, LL dropout, and temperature scaling all had similar performances, while LL SVI had a much lower fraction of marginal coverage above the regression lines. None of the methods have a commensurate increase in width to maintain the 95% coverage levels seen on in-distribution test data as dataset shift increases.

5. Discussion

We have provided the first comprehensive empirical study of the frequentist-style coverage properties of popular uncertainty quantification techniques for deep learning models. In regression tasks, LL SVI, SVI, and Gaussian processes all had high levels of coverage across nearly all benchmarks. LL SVI, in particular, had the lowest widths amongst methods with high coverage. SVI also had excellent coverage properties across most tasks with tighter intervals than GPs and linear regression. In contrast, the methods based on ensembles and Monte Carlo dropout had significantly worse coverage due to their overly confident and tight prediction intervals.

In the classification setting, all methods showed very high coverage in the i.i.d setting (i.e., no dataset shift), as coverage is reflective of top-1 accuracy in this scenario. On MNIST data, SVI had the best performance, maintaining high levels of coverage under slight dataset shift and scaling the width of its prediction intervals more appropriately as shift increased relative to other methods. On CIFAR-10 data and ImageNet, ensemble models were superior. They had the highest coverage relative to other methods, as demonstrated in Figures 9 and 10.

An important consideration throughout this work is the choice of hyperparameters in most all of the analyzed methods makes a significant impact on the uncertainty estimates. We set hyperparameters and optimized model parameters according to community best practices in an attempt to reflect what a "real-world" machine learning practitioner might do: selecting hyperparameters based on minimizing validation loss over nested cross-validation. Our work is a measurement of the empirical coverage properties of these methods as one would typically utilize them, rather than an exploration of how pathological hyperparameters can skew uncertainty estimates to 0 or to infinity, while this is an inherent limitation in the applicability of our work to every context, our sensible choices will provide a relevant benchmark for models in practice.

Of particular note is that the width of a prediction interval or set typically correlated with the degree of dataset shift. For instance, when the translation shift is applied to MNIST, both prediction set width and dataset shift is maximized at around 14 pixels. There is a 0.9 Pearson correlation between width and shift. Width can serve as a soft proxy of dataset shift and potentially detect shift in real-world scenarios.

Simultaneously, the ranks of coverage, Brier score, and ECE are all generally consistent. However, coverage is arguably the most interpretable to downstream users of machine learning models. Clinicians, for instance, may not have the technical training to have an intuition about what specific values of Brier score or ECE mean in practice, while coverage and width are readily understandable. Manrai et al. [33] already demonstrated clinicians' general lack of intuition about the positive predictive value, and these uncertainty quantification metrics are more difficult to internalize than PPV.

Moreover, proper scoring rules (e.g., Brier score and negative log-likelihood) can be misleading under model misspecification [34]. Negative log-likelihood, specifically, suffers from the potential impact of a few points with low probability. These points can contribute near-infinite terms to NLL that distort interpretation. In contrast, marginal coverage over a dataset is less sensitive to the impacts of outlying data.

In summary, we find that popular uncertainty quantification methods for deep learning models do not provide good coverage properties under moderate levels of dataset shift. Although the width of prediction regions do increase under increasing amounts of shift, these changes are not enough to maintain the levels of coverage seen on i.i.d data. We conclude that the methods we evaluated for uncertainty quantification are likely insufficient for use in high-stakes, real-world applications, where dataset shift is likely to occur. However, marginal coverage of a prediction interval or set is a natural and intuitive metric to quantify uncertainty. The width of a prediction interval/set is an additional tool that captures dataset shift and provides additional interpretable information to downstream users of machine learning models.

Author Contributions: Conceptualization, B.K., J.S. and A.L.B.; methodology, B.K. and A.L.B.; software, B.K. and J.S.; validation, B.K.; formal analysis, B.K., J.S. and A.L.B.; investigation, B.K.; resources, A.L.B.; data curation, B.K.; writing—original draft preparation, B.K.; writing—review and editing, B.K, J.S. and A.L.B.; visualization, B.K.; supervision, A.L.B.; project administration, A.L.B.; funding acquisition, A.L.B. All authors have read and agreed to the published version of the manuscript.

Funding: Beam was supported by award 5K01HL141771 from the NHLBI.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Code is available at https://github.com/beamlab-hsph/coverage-quantification (accessed on 28 November 2021) with additional links to the public datasets, model parameters, and model predictions used in this work.

Acknowledgments: We would like to thank Alex D'Amour and Balaji Lakshminarayanan for their insightful comments on a draft of this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Additional Results

Appendix A.1. Regression Tables

Table A1. The average coverage of six methods across nine datasets with the standard error over 20 cross-validation folds in parentheses.

Dataset | Method	Linear Regression	GP	Ensemble	Dropout	LL Dropout	SVI	LL SVI
Boston Housing	9.461×10^{-1} (5.61×10^{-3})	9.765×10^{-1} (5.05×10^{-3})	5.912×10^{-1} (1.43×10^{-2})	6.02×10^{-1} (1.64×10^{-2})	1.902×10^{-1} (2.01×10^{-2})	9.434×10^{-1} (6.04×10^{-3})	9.3399×10^{-1} (8.48×10^{-3})
Concrete	9.437×10^{-1} (2.68×10^{-3})	9.67×10^{-1} (3.02×10^{-3})	5.854×10^{-1} (1.04×10^{-2})	7.2882×10^{-1} (1.17×10^{-2})	9.32×10^{-2} (1.75×10^{-2})	9.581×10^{-1} (3.61×10^{-3})	9.443×10^{-1} (6.72×10^{-3})
Energy	8.957×10^{-1} (4.66×10^{-3})	8.857×10^{-1} (6.96×10^{-3})	8.669×10^{-1} (5.26×10^{-3})	8.013×10^{-1} (2.00×10^{-2})	2.597×10^{-1} (2.75×10^{-2})	9.773×10^{-1} (3.02×10^{-3})	9.938×10^{-1} (2.99×10^{-3})
Kin8nm	9.514×10^{-1} (1.20×10^{-3})	9.705×10^{-1} (1.53×10^{-3})	6.706×10^{-1} (4.43×10^{-3})	8.037×10^{-1} (8.15×10^{-3})	1.984×10^{-1} (1.36×10^{-2})	9.618×10^{-1} (2.63×10^{-3})	9.633×10^{-1} (1.36×10^{-3})
Naval Propulsion Plant	9.373×10^{-1} (1.59×10^{-3})	9.994×10^{-1} (2.12×10^{-4})	8.036×10^{-1} (5.99×10^{-3})	9.212×10^{-1} (6.76×10^{-3})	2.683×10^{-1} (2.51×10^{-2})	9.797×10^{-1} (1.88×10^{-3})	9.941×10^{-1} (1.25×10^{-3})
Power Plant	9.646×10^{-1} (1.14×10^{-3})	9.614×10^{-1} (1.26×10^{-3})	4.008×10^{-1} (1.12×10^{-2})	4.32×10^{-1} (1.47×10^{-2})	1.138×10^{-1} (1.41×10^{-2})	9.626×10^{-1} (1.13×10^{-3})	9.623×10^{-1} (1.60×10^{-3})
Protein Tertiary Structure	9.619×10^{-1} (4.71×10^{-4})	9.59×10^{-1} (4.72×10^{-4})	4.125×10^{-1} (2.98×10^{-3})	3.846×10^{-1} (1.36×10^{-2})	1.182×10^{-1} (1.35×10^{-2})	9.609×10^{-1} (2.27×10^{-3})	9.559×10^{-1} (1.72×10^{-3})
Wine Quality Red	9.425×10^{-1} (2.32×10^{-3})	9.472×10^{-1} (3.28×10^{-3})	3.919×10^{-1} (1.18×10^{-2})	3.556×10^{-1} (1.83×10^{-2})	1.616×10^{-1} (7.45×10^{-3})	9.059×10^{-1} (8.19×10^{-3})	8.647×10^{-1} (8.77×10^{-3})
Yacht Hydrodynamics	9.449×10^{-1} (7.86×10^{-3})	9.726×10^{-1} (6.73×10^{-3})	9.161×10^{-1} (7.38×10^{-3})	3.871×10^{-1} (2.82×10^{-2})	2.081×10^{-1} (2.54×10^{-2})	9.807×10^{-1} (6.97×10^{-3})	9.899×10^{-1} (6.03×10^{-3})

Table A2. The average width of the posterior prediction interval of six methods across nine datasets with the standard error over 20 cross-validation folds in parentheses. Width is reported in terms of standard deviations of the response variable in the training set.

Dataset | Method	Linear Regression	GP	Ensemble	Dropout	LL Dropout	SVI	LL SVI
Boston Housing	2.0424×10^{0} (6.87×10^{-3})	1.8716×10^{0} (1.17×10^{-2})	4.432×10^{-1} (7.82×10^{-3})	6.882×10^{-1} (2.19×10^{-2})	1.855×10^{-1} (2.05×10^{-2})	1.301×10^{0} (2.56×10^{-2})	1.148×10^{0} (2.36×10^{-2})
Concrete	2.4562×10^{0} (2.22×10^{-3})	2×10^{0} (3.32×10^{-3})	4.776×10^{-1} (9.03×10^{-3})	1.0342×10^{0} (1.79×10^{-2})	1.028×10^{-1} (2.04×10^{-2})	1.5116×10^{0} (1.72×10^{-2})	1.2293×10^{0} (1.41×10^{-2})
Energy	1.144×10^{0} (2.29×10^{-3})	1.0773×10^{0} (2.64×10^{-3})	2.394×10^{-1} (2.56×10^{-3})	5.928×10^{-1} (1.22×10^{-2})	1.1417×10^{-1} (1.61×10^{-2})	8.426×10^{-1} (1.73×10^{-2})	7.974×10^{-1} (1.95×10^{-2})
Kin8nm	3.0039×10^{0} (9.76×10^{-4})	2.3795×10^{0} (7.02×10^{-3})	5.493×10^{-1} (2.37×10^{-3})	1.2355×10^{0} (1.37×10^{-2})	2.024×10^{-1} (1.22×10^{-2})	1.6697×10^{0} (7.75×10^{-3})	1.2624×10^{0} (2.99×10^{-3})
Naval Propulsion Plant	1.5551×10^{0} (7.12×10^{-4})	3.403×10^{-1} (1.00×10^{-2})	6.048×10^{-1} (4.86×10^{-3})	1.1593×10^{0} (6.45×10^{-3})	2.281×10^{-1} (1.83×10^{-2})	1.3064×10^{0} (1.38×10^{-1})	4.88×10^{-1} (5.44×10^{-3})
Power Plant	1.0475×10^{0} (7.09×10^{-4})	9.768×10^{-1} (9.63×10^{-4})	2.494×10^{-1} (6.72×10^{-3})	3.385×10^{-1} (1.69×10^{-2})	9.18×10^{-2} (9.06×10^{-3})	1.0035×10^{0} (1.88×10^{-3})	9.818×10^{-1} (3.64×10^{-3})
Protein Tertiary Structure	3.3182×10^{0} (3.21×10^{-4})	3.2123×10^{0} (3.47×10^{-3})	6.804×10^{-1} (3.77×10^{-3})	9.144×10^{-1} (1.41×10^{-2})	3.454×10^{-1} (1.99×10^{-2})	2.9535×10^{0} (3.82×10^{-2})	2.6506×10^{0} (2.20×10^{-2})
Wine Quality Red	3.1573×10^{0} (1.82×10^{-3})	3.1629×10^{0} (4.07×10^{-3})	7.763×10^{-1} (1.31×10^{-2})	7.841×10^{-1} (2.91×10^{-2})	3.481×10^{-1} (1.61×10^{-2})	2.7469×10^{0} (2.72×10^{-2})	2.3597×10^{0} (2.70×10^{-2})
Yacht Hydrodynamics	2.3636×10^{0} (2.89×10^{-3})	1.6974×10^{0} (7.57×10^{-3})	4.475×10^{-1} (9.76×10^{-3})	5.443×10^{-1} (2.22×10^{-2})	1.081×10^{-1} (9.83×10^{-3})	6.57×10^{-1} (3.54×10^{-2})	6.9×10^{-1} (3.81×10^{-2})

Appendix A.2. Classification Results

Table A3. MNIST average coverage and width for the test set, rotation shift, and translation shift.

Method	Mean Test Set Coverage (SE)	Mean Test Set Width (SE)	Mean Rotation Shift Coverage (SE)	Mean Rotation Shift Width (SE)	Mean Translation Shift Coverage (SE)	Mean Translation Shift Width (SE)
Dropout	9.987×10^{-1} (6.32×10^{-5})	1.06×10^{0} (1.38×10^{-4})	5.519×10^{-1} (2.91×10^{-2})	2.3279×10^{0} (6.64×10^{-2})	5.333×10^{-1} (3.54×10^{-2})	2.3527×10^{0} (6.34×10^{-2})
Ensemble	9.9984×10^{-1} (7.07×10^{-5})	1.0424×10^{0} (2.07×10^{-4})	5.157×10^{-1} (3.11×10^{-2})	2.0892×10^{0} (5.44×10^{-2})	5.424×10^{-1} (3.33×10^{-2})	2.3276×10^{0} (6.66×10^{-2})
LL Dropout	9.985×10^{-1} (1.05×10^{-4})	1.0561×10^{0} (1.89×10^{-3})	5.52×10^{-1} (2.93×10^{-2})	2.3162×10^{0} (6.73×10^{-2})	5.388×10^{-1} (3.52×10^{-2})	2.3658×10^{0} (6.66×10^{-2})
LL SVI	9.984×10^{-1} (1.14×10^{-4})	1.0637×10^{0} (1.65×10^{-3})	5.746×10^{-1} (2.77×10^{-2})	2.6324×10^{0} (8.41×10^{-2})	5.35×10^{-1} (3.51×10^{-2})	2.3294×10^{0} (6.46×10^{-2})
SVI	9.9997×10^{-1} (7.35×10^{-5})	1.5492×10^{0} (2.19×10^{-2})	7.148×10^{-1} (2.06×10^{-2})	4.8549×10^{0} (1.44×10^{-1})	7.54×10^{-1} (1.96×10^{-2})	5.6803×10^{0} (1.99×10^{-1})
Temp scaling	9.986×10^{-1} (1.36×10^{-4})	1.0642×10^{0} (1.98×10^{-3})	5.243×10^{-1} (3.10×10^{-2})	2.2683×10^{0} (6.17×10^{-2})	5.375×10^{-1} (3.33×10^{-2})	2.347×10^{0} (6.21×10^{-2})
Vanilla	9.972×10^{-1} (1.16×10^{-4})	1.032×10^{0} (9.06×10^{-4})	4.715×10^{-1} (3.28×10^{-2})	1.7492×10^{0} (3.78×10^{-2})	4.798×10^{-1} (3.50×10^{-2})	1.801×10^{0} (3.84×10^{-2})

Table A4. CIFAR-10 average coverage and width for the test set and translation shift.

Method	Mean Test Set Coverage (SE)	Mean Test Set Width (SE)	Mean Translation Shift Coverage (SE)	Mean Translation Shift Width (SE)
Dropout	9.8883×10^{-1} (3.79×10^{-4})	1.5778×10^{0} (2.68×10^{-3})	9.696×10^{-1} (2.48×10^{-3})	2.0709×10^{0} (5.11×10^{-2})
Ensemble	9.922×10^{-1} (3.08×10^{-4})	1.4925×10^{0} (1.52×10^{-3})	9.806×10^{-1} (1.65×10^{-3})	1.9246×10^{0} (4.49×10^{-2})
LL Dropout	9.628×10^{-1} (1.40×10^{-3})	1.3007×10^{0} (3.99×10^{-3})	9.184×10^{-1} (5.59×10^{-3})	1.6678×10^{0} (4.16×10^{-2})
LL SVI	9.677×10^{-1} (1.10×10^{-3})	1.2585×10^{0} (2.60×10^{-3})	9.29×10^{-1} (4.55×10^{-3})	1.5044×10^{0} (2.61×10^{-2})
SVI	9.789×10^{-1} (6.41×10^{-4})	1.5579×10^{0} (6.31×10^{-3})	9.543×10^{-1} (2.89×10^{-3})	1.9286×10^{0} (3.69×10^{-2})
Temp scaling	9.871×10^{-1} (3.51×10^{-4})	1.5987×10^{0} (1.19×10^{-2})	9.707×10^{-1} (1.97×10^{-3})	2.1266×10^{0} (5.30×10^{-2})
Vanilla	9.686×10^{-1} (6.06×10^{-4})	1.2611×10^{0} (3.90×10^{-3})	9.296×10^{-1} (4.36×10^{-3})	1.5064×10^{0} (2.58×10^{-2})

Table A5. The mean coverage and widths on the test set of CIFAR-10, as well as on the mean coverage and width averaged over 16 corruptions and 5 intensities.

Method	Mean Test Set Coverage (SE)	Mean Test Set Width (SE)	Mean Corruption Coverage (SE)	Mean Corruption Width (SE)
Dropout	9.87×10^{-1} (3.72×10^{-4})	1.578×10^{0} (2.68×10^{-3})	8.86×10^{-1} (6.34×10^{-3})	2.313×10^{0} (3.03×10^{-2})
Ensemble	9.92×10^{-1} (9.70×10^{-5})	1.492×10^{0} (1.52×10^{-3})	9.11×10^{-1} (5.16×10^{-3})	2.425×10^{0} (3.69×10^{-2})
LL Dropout	9.60×10^{-1} (8.77×10^{-4})	1.301×10^{0} (3.99×10^{-3})	8.15×10^{-1} (7.48×10^{-3})	1.699×10^{0} (1.53×10^{-2})
LL SVI	9.64×10^{-1} (6.64×10^{-4})	1.258×10^{0} (2.60×10^{-3})	8.17×10^{-1} (7.52×10^{-3})	1.781×10^{0} (2.15×10^{-2})
SVI	9.76×10^{-1} (5.10×10^{-4})	1.558×10^{0} (6.31×10^{-3})	8.81×10^{-1} (5.45×10^{-3})	2.161×10^{0} (2.32×10^{-2})
Temp Scaling	9.85×10^{-1} (4.54×10^{-4})	1.599×10^{0} (1.19×10^{-2})	8.99×10^{-1} (4.85×10^{-3})	2.636×10^{0} (3.86×10^{-2})
Vanilla	9.64×10^{-1} (6.36×10^{-4})	1.261×10^{0} (3.90×10^{-3})	8.23×10^{-1} (7.10×10^{-3})	1.790×10^{0} (2.16×10^{-2})

Table A6. The mean coverage and widths on the test set of ImageNet, as well as on the mean coverage and width averaged over 16 corruptions and 5 intensities.

Method	Mean Test Set Coverage	Mean Test Set Width	Mean Corruption Coverage (SE)	Mean Corruption Width (SE)
Dropout	9.613×10^{-1}	1.32699×10^{1}	8.579×10^{-1} (1.61×10^{-2})	8.75784×10^{1} (7.80×10^{0})
Ensemble	9.701×10^{-1}	1.30613×10^{1}	9.231×10^{-1} (7.13×10^{-3})	1.053608×10^{2} (8.57×10^{0})
LL Dropout	9.552×10^{-1}	1.07707×10^{1}	8.688×10^{-1} (1.18×10^{-2})	8.80326×10^{1} (8.04×10^{0})
LL SVI	9.327×10^{-1}	1.05624×10^{1}	7.77×10^{-1} (1.76×10^{-2})	6.59982×10^{1} (5.01×10^{0})
Temp Scaling	9.613×10^{-1}	1.54811×10^{1}	8.829×10^{-1} (1.10×10^{-2})	1.051409×10^{2} (8.43×10^{0})
Vanilla	9.525×10^{-1}	1.10255×10^{1}	8.529×10^{-1} (1.27×10^{-2})	8.0687×10^{1} (7.16×10^{0})

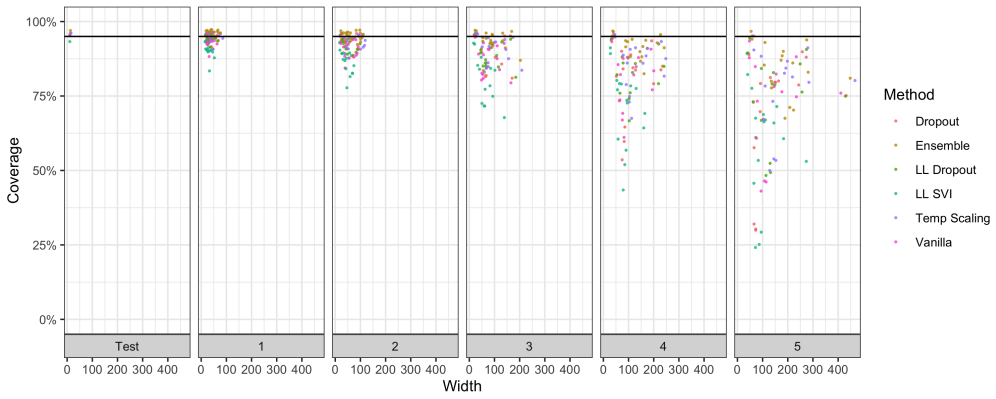


Figure A1. The effect of corruption intensity on coverage levels vs. width in ImageNet-C. Each facet panel represents a different corruption level, while points are the coverage of a model on one of 16 corruptions. Each facet has 16 points per method, as only 1 iteration was trained per method. For methods equal coverage, the superior method is to the left as it has a lower width.

Appendix B. Hyperparameter Search and Model Details

A brief summary of the models utilized in this work:

- **Vanilla networks** in the style of [1], which are feedforward networks that were simply fully connected dense layers. Since there is no element of variability in the model's prediction for the same sample, we could not consider the coverage of vanilla networks in regression tasks. They simply produce a single-point estimate given the same sample.
- **Temperature Scaling** was considered in classification tasks. This is a post-training calibration measure using a validation set as in [25].
- **Dropout** as in [16]. Feedforward networks had Monte Carlo dropout in between their dense layers. At test time, dropout still applied. In our work, we sampled networks 200 times to obtain a distribution of predictions.
- **Ensembles** as found in [3]. We took the outputs from 40 independently-trained vanilla networks, and these formed a predictive distribution.
- **Stochastic Variational Inference** (SVI) models, such as those of [7,8,11,30]. SVI models had difficult convergence properties in our experience and required the use of empirical Bayes for prior standard deviations.
- **Last layer methods** We considered LL dropout and LL SVI where dropout and mean-field stochastic variational inference were applied to only the last layer of an otherwise vanilla network, respectively.
- **Gaussian Processes** We implement sparse Gaussian processes [35] for regression with a RBF kernel and 10 inducing points.
- **Linear regression** We use the standard `lm` function in `R` to obtain prediction intervals for linear regression.

All models were implemented in `Keras`, with the exception of GPs (`GPy`) and linear regression (`R`). Hyperparameters were found for each model over 50 trials with a random sampling of the values described below.

Table A7. The hyperparameters considered in our search for vanilla, dropout, ensemble, SVI, and LL models.

Hyperparameter	Range	Sampling Strategy
Dropout rate	[0, 0.5]	Uniform
Number of Hidden Layers	{1,2,3}	Uniform
Layer Width	{16, 32, 48, 64}	Uniform
Learning Rate	$[1 \times 10^{-4}, 1 \times 10^{-1}]$	Log uniform
Batch Size	32	Fixed
Max Epochs	50	Fixed

We performed our hyperparameter search as part of K-fold cross validation. In regression tasks, we had 20 folds. On the larger split of each fold, we split the data 80/20 to form train/val splits for hyperparameter evaluation. In classification tasks, we were able to reuse the published predictions from these models from [14] for each sample in the held-out test set.

References

1. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
2. Kompa, B.; Snoek, J.; Beam, A.L. Second opinion needed: Communicating uncertainty in medical machine learning. *NPJ Digit. Med.* **2021**, *4*, 4. [CrossRef]
3. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4 December 2017; pp. 6405–6416.
4. Yao, J.; Pan, W.; Ghosh, S.; Doshi-Velez, F. Quality of Uncertainty Quantification for Bayesian Neural Network Inference. *arXiv* **2019**, arXiv:1906.09686.
5. Neal, R.M. *Bayesian Learning for Neural Networks*; Springer: New York, NY, USA, 1996.
6. Hernández-Lobato, J.M.; Adams, R.P. Probabilistic Backpropagation for Scalable Learning of Bayesian Neural Networks. In Proceedings of the 32nd International Conference on International Conference on Machine Learning (ICML'15), Lille, France, 6–11 July 2015; pp. 1861–1869
7. Blundell, C.; Cornebise, J.; Kavukcuoglu, K.; Wierstra, D. Weight uncertainty in neural networks. In Proceedings of the 32nd International Conference on International Conference on Machine Learning (ICML'15), Lille, France, 6–11 July 2015; pp. 1613–1622.
8. Graves, A. Practical Variational Inference for Neural Networks. In Proceedings of the 25th Conference on Neural Information Processing Systems (NeurIPS 2011), Grenada, Spain, 12–17 December 2011.
9. Pawlowski, N.; Brock, A.; Lee, M.C.H.; Rajchl, M.; Glocker, B. Implicit Weight Uncertainty in Neural Networks. *arXiv* **2017**, arXiv:1711.01297.
10. Hernández-Lobato, J.M.; Li, Y.; Rowland, M.; Hernández-Lobato, D.; Bui, T.; Turner, R.E. Black-box α-divergence Minimization. In Proceedings of The 33rd International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016.
11. Louizos, C.; Welling, M. Structured and Efficient Variational Deep Learning with Matrix Gaussian Posteriors. In Proceedings of the 33rd International Conference on International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016.
12. Louizos, C.; Welling, M. Multiplicative normalizing flows for variational Bayesian neural networks. In Proceedings of the International Conference of Machine Learning, Sydney, Australia, 6–11 August 2017.
13. Wenzel, F.; Roth, K.; Veeling, B.S.; Świątkowski, J.; Tran, L.; Mandt, S.; Snoek, J.; Salimans, T.; Jenatton, R.; Nowozin, S. How Good is the Bayes Posterior in Deep Neural Networks Really? In Proceedings of the International Conference on Machine Learning, Vienna, Australia, 12–18 July 2020.
14. Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J.V.; Lakshminarayanan, B.; Snoek, J. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. In Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, BC, Canada, 8–13 December 2019.
15. Wasserman, L. *All of Statistics: A Concise Course in Statistical Inference*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2013.
16. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In Proceedings of the 33rd International Conference on International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016.
17. Barber, R.F.; Candès, E.J.; Ramdas, A.; Tibshirani, R.J. The limits of distribution-free conditional predictive inference. *Inf. Inference J. IMA* **2021**, *10*, 455–482. [CrossRef]
18. Vovk, V.; Gammerman, A.; Shafer, G. *Algorithmic Learning in a Random World*; Springer: Boston, MA, USA, 2005.
19. Angelopoulos, A.N.; Bates, S. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. *arXiv* **2021**, arXiv:2107.07511.

20. Shafer, G.; Vovk, V. A Tutorial on Conformal Prediction. *J. Mach. Learn. Res.* **2008**, *9*, 371–421.
21. Hoff, P. Bayes-optimal prediction with frequentist coverage control. *arXiv* **2021**, arXiv:2105.14045.
22. Cauchois, M.; Gupta, S.; Ali, A.; Duchi, J.C. Robust Validation: Confident Predictions Even When Distributions Shift. *arXiv* **2020**, arXiv:2008.04267.
23. Barber, R.F.; Candes, E.J.; Ramdas, A.; Tibshirani, R.J. Conformal Prediction Under Covariate Shift. *arXiv* **2019**, arXiv:1904.06019.
24. Maddox, W.; Garipov, T.; Izmailov, P.; Vetrov, D.; Wilson, A.G. A Simple Baseline for Bayesian Uncertainty in Deep Learning. In Proceedings of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, Vancouver, BC, Canada, 8–14 December 2019.
25. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017.
26. van Amersfoort, J.; Smith, L.; Teh, Y.W.; Gal, Y. Uncertainty Estimation Using a Single Deep Deterministic Neural Network. In Proceedings of the International Conference on Machine Learning, Vienna, Austria, 12–18 July 2020.
27. Liu, J.Z.; Lin, Z.; Padhy, S.; Tran, D.; Bedrax-Weiss, T.; Lakshminarayanan, B. Simple and Principled Uncertainty Estimation with Deterministic Deep Learning via Distance Awareness. In Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, BC, Canada, 6 December 2020.
28. Hendrycks, D.; Gimpel, K. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In Proceedings of the International Conference on Learning Representations, Palais des Congrès Neptune, Toulon, France, 24–26 April 2017.
29. Srivastava, R.K.; Greff, K.; Schmidhuber, J. Training Very Deep Networks. In Proceedings of the Advances in Neural Information Processing Systems, Annual Conference on Neural Information Processing Systems 2015, Montreal, QC, Canada, 7–12 December 2015.
30. Wen, Y.; Vicol, P.; Ba, J.; Tran, D.; Grosse, R. Flipout: Efficient Pseudo-Independent Weight Perturbations on Mini-Batches. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
31. Riquelme, C.; Tucker, G.; Snoek, J. Deep Bayesian Bandits Showdown: An Empirical Comparison of Bayesian Deep Networks for Thompson Sampling. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
32. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
33. Manrai, A.K.; Bhatia, G.; Strymish, J.; Kohane, I.S.; Jain, S.H. Medicine's uncomfortable relationship with math: Calculating positive predictive value. *JAMA Intern. Med.* **2014**, *174*, 991–993. [CrossRef]
34. Martin, G.M.; Loaiza-Maya, R.; Frazier, D.T.; Maneesoonthorn, W.; Hassan, A.R. Optimal probabilistic forecasts: When do they work? *Int. J. Forecast.* **2021**, in press. [CrossRef]
35. Snelson, E.; Ghahramani, Z. Local and global sparse Gaussian process approximations. In Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, PMLR, San Juan, Puerto Rico, 21–24 March 2007; Volume 2, pp. 524–531.