Adaptive, Distribution-Free Prediction Intervals for Deep Neural Networks

Danijel Kivaranovica Kory D. Johnsona Hannes Leeb.a,b
aDepartment of Statistics and Operations Research bDataScience@UniVienna
University of Vienna
{danijel.kivaranovic,kory.johnson,hannes.leeb}@univie.ac.at

Abstract

This paper addresses the problem of assessing the variability of predictions from deep neural networks. There is a growing literature on using and improving the predictive accuracy of deep networks, but a concomitant improvement in the quantification of their uncertainty is lacking. We provide a prediction interval network (PI-Network) which is a transparent, tractable modification of the standard predictive loss used to train deep networks. The PI-Network outputs three values instead of a single point estimate and optimizes a loss function inspired by quantile regression. We go beyond merely motivating the construction of these networks and provide two prediction interval methods with provable, finite sample coverage guarantees without any assumptions on the underlying distribution from which our data is drawn. We only require that the observations are independent and identically distributed. Furthermore, our intervals adapt to heteroskedasticity and asymmetry in the conditional distribution of the response given the covariates. The first method leverages the conformal inference framework and provides average coverage. The second method provides a new, stronger guarantee by conditioning on the observed data. Lastly, our loss function does not compromise the predictive accuracy of the network like other prediction interval methods. We demonstrate the ease of use of the PI-Network as well as its improvements over other methods on both simulated and real data. As the PI-Network can be used with a host of deep learning methods with only minor modifications, its use should become standard practice, much like reporting standard errors along with mean estimates.

1 Introduction

Deep neural networks have gained tremendous popularity in the last decade due to their superior predictive performance over other machine learning algorithms. They have become the state-of-the-art algorithm in many challenging tasks such as computer vision [15, 11], speech recognition [9], natural language processing [4], and bioinformatics [1].

Despite these successes, there is a paucity of research on the uncertainty of neural network predictions on new samples. While there is research on understanding the uncertainty of networks through Bayesian approaches [18, 6], the development of accurate prediction intervals (PIs) for neural networks is a challenging task that is only beginning to gain research interest. Several authors have provided motivation for modified loss function intended to encourage desirable properties [19, 8, 14, 16, 22, 12, 25]. That being said, some provide a PI without a point estimate [22, 12] or use loss functions which cannot be optimized with stochastic gradient descent [13]. Others have distributional assumptions [19, 16] or provide intervals in which the lower bound is not guaranteed to be smaller than the upper bound [25]. These methods also do not provide rigorous guarantees that the intervals posses the desired statistical properties.

Preprint. Under review.
One notable exception to these critiques is the work of Papadopoulos and Haralambous [20]. They use, among others, a simple method from conformal inference to create provably valid prediction intervals. Unfortunately, the intervals are fixed-width and thus cannot adapt to any heteroskedasticity in the distribution of the response given covariates. We discuss conformal inference in more detail in Section 2.1 as it motivates one of our prediction intervals.

These problems are all remedied in this paper: we show how to train a neural network (PI-Network) that outputs both accurate predictions and valid PIs that adapt to heteroskedasticity. We advocate using one of two prediction interval methods that provably provide average coverage over new samples drawn independently from the same data generating distribution. These intervals require minimal additional effort to compute and only minor changes to the predictive architecture of the neural network. Their transparent construction invites their use in all relevant contexts.

In Section 1.1 we formally set up the prediction problem and explain in detail what we mean by accurate PIs. This includes standard concepts such as average coverage but also a new, related idea we term Probably-Approximately-Valid (PAV). Section 2 introduces our PI-Network and the algorithms to produce valid confidence intervals. Extensive simulations and real-data examples are given in Section 3 that highlight the improvements our methods make over competing algorithms. In particular, we show that other methods either fail to provide adequate coverage, compromise predictive accuracy, or fail to account for heteroskedasticity. Furthermore, we show via simulations that our proposed methods have asymptotically nearly optimal performance.

### 1.1 Prediction Intervals

We consider a non-parametric regression setting, where $X$ denotes the $\mathbb{R}^d$-valued covariate vector and $Y$ the $\mathbb{R}$-valued response. The data set $D = (X_i, Y_i)_{1 \leq i \leq n}$ contains $n$ i.i.d. copies of the random variable $(X, Y)$. Throughout the paper, we assume that $(X, Y)$ is independent of the data set $D$.

There are various formal criteria that prediction intervals could satisfy. In general, a prediction interval $\Gamma_\alpha(X) = \Gamma_{D,\alpha}(X)$ is an interval-valued function of $X$, the data $D$, and the confidence level $\alpha$, such that, loosely speaking, a new observation falls within the interval with probability at least $1 - \alpha$. This loose definition can be made precise in various ways. In Section 2.1, we propose a conformal prediction interval, $\Gamma_c^\alpha(X)$, and Theorem 1 demonstrates that it provides average coverage:

$$\mathbb{P}(Y \in \Gamma_c^\alpha(X)) \geq 1 - \alpha. \quad (1)$$

Note that the probability in equation (1) is over all of the random variables included: the new observation $(X, Y)$ and the data $D$. As such, $\Gamma_c^\alpha(X)$ provides $(1 - \alpha)\%$-confidence in the statistical sense: it is the result of a method such that in repeated applications on independent data sets, $(1 - \alpha)\%$ of the new realizations satisfy $Y \in \Gamma_c^\alpha(X)$. Shafer and Vovk [25] provide a cogent discussion of the difference between this and related notions as it pertains to conformal inference.

One may object to average coverage because one would like a coverage guarantee given a particular data set instead of averaging over all potential data sets. Our second prediction interval method takes steps toward alleviating this concern. We define a notion of prediction interval validity called Probably Approximately Valid (PAV), which is closely related to the theory of probably approximately correct (PAC) learning [26]. A task is PAC-learnable if, loosely speaking, regardless of the data generating distribution, one can approximate the task arbitrarily well with high probability given sufficient data. Similarly, a PAV interval provides a coverage guarantee with high probability regardless of the data generating distribution:

**Definition 1.** Let $p(\Gamma_\alpha|D)$ be the conditional probability that $\Gamma_\alpha(X)$ contains $Y$ conditional on $D$: $p(\Gamma_\alpha|D) = \mathbb{P}(Y \in \Gamma_\alpha(X)|D)$. For all $\epsilon > 0$, all $\delta > 0$, all $n \geq n_0(\epsilon, \delta)$, and all distributions over $(X, Y)$, $\Gamma_\alpha(X)$ is Probably Approximately Valid (PAV) if

$$\mathbb{P}(p(\Gamma_\alpha|D) \leq 1 - \alpha - \epsilon) \geq \delta. \quad (2)$$

$\Gamma_\alpha(X)$ is PAV if the coverage probability is at least $1 - \alpha - \epsilon$ for all but $\delta100\%$ of data sets $D$. Ideally, the required sample size, $n_0(\epsilon, \delta)$, is not too large when $\epsilon$ and $\delta$ are small. In Section 2.2, we show how to construct a PI, $\Gamma_c^\alpha(X)$, that is PAV such that $n_0(\epsilon, \delta)$ is of order $-(\log(\delta)/\epsilon^2)$. Corollary 1 shows that PAV prediction intervals can also control average with proper choice of $\alpha, \epsilon,$ and $\delta$. 

2
The two prediction intervals, \( \Gamma^\alpha_0(X) \) and \( \Gamma^\alpha_1(X) \), are the result of a PI-Network that outputs a three-dimensional vector and optimizes a loss function used in quantile regression. The next subsection describes and motivates this loss function.

### 1.2 The Loss Function of the PI-Network

Let \( N : \mathbb{R}^d \to \mathbb{R}^3 \) be a network such that \( N(x) = (l(x), m(x), u(x)) \), where \( l, m, u : \mathbb{R}^d \to \mathbb{R} \) with the restriction that \( l(x) \leq m(x) \leq u(x) \) for all \( x \in \mathbb{R} \). We use \( m(x) \) to estimate the median of \( Y \) given \( X \), while \( l(x) \) and \( u(x) \) estimate the lower and upper bounds of our prediction interval, respectively. The monotonicity, \( l(x) \leq m(x) \leq u(x) \), is easily enforced by modifying any network that outputs a triple \((z_1, z_2, z_3)\) to output \((z'_1, z'_2, z'_3)\) given by \( z'_1 = z_1, z'_2 = z'_1 + \text{ReLU}(z_2 - z'_1), \) and \( z'_3 = z'_2 + \text{ReLU}(z_3 - z'_2) \). Here, \( \text{ReLU}(\cdot) \) is the rectified linear unit, \( \max(0, \cdot) \).

For \( \tau \in [0, 1] \) and \( u \in \mathbb{R} \), let \( h_\tau(u) = (\tau - 1_{u \leq 0})u \), be the asymmetric absolute loss function. We define the level-\( \tau \) loss function evaluated on \( N \) at \((x, y)\) by

\[
\mathcal{L}_\tau(N(x), y) = h_{\tau/2}(y - l(x)) + h_{1/2}(y - m(x)) + h_{1-\tau/2}(y - u(x)).
\]

Letting \( |D'| \) denote the cardinality of a set \( D' \), the empirical risk of network \( N \) on \( D' \subseteq D \) is

\[
\mathcal{R}_{D', \tau}(N) = \frac{1}{|D'|} \sum_{(x_i, y_i) \in D'} \mathcal{L}_\tau(N(x_i), y_i).
\]

**Definition 2.** Let \( N_{D', \tau}(x) = (l_{D', \tau}(x), m_{D', \tau}(x), u_{D', \tau}(x)) \) denote a neural network fit on \( D' \subseteq D \) by minimizing the empirical risk \( \mathcal{R}_{D', \tau}(N) \). In the trivial case \( \tau = 0 \), we set \( l_0(x) = -\infty \) and \( u_0(x) = \infty \) for all \( x \in \mathbb{R}^d \).

By the strong law of large numbers, the empirical risk \( \mathcal{R}_{D', \tau}(N) \) converges almost surely to the risk \( \mathbb{E}[\mathcal{L}_\tau(N(X), Y)] \) as \( |D| \to \infty \). Define \( q_\tau(x) = \inf \{ y : \mathbb{P}(Y \leq y | X = x) \geq \tau \} \) to be the conditional \( \tau \)-quantile of \( Y \) given \( X = x \), e.g., \( q_{1/2}(x) \) is the conditional median of \( Y \) given \( X = x \). Following standard texts [14], \((q_{1/2}(x), q_1/2(x), q_{1-\tau/2}(x))\) is the minimizer of the risk \( \mathbb{E}[\mathcal{L}_\tau(N(X), Y)] \). For a given confidence level \( \alpha \in (0, 1) \), \((q_{\alpha/2}(x), q_{1/2}(x), q_{1-\alpha/2}(x))\) has the desirable properties that

\[
q_{1/2}(\cdot) = \arg\min_{f : \mathbb{R} \to \mathbb{R}} \mathbb{E}[|Y - f(X)|]
\]

and

\[
\mathbb{P}(q_{\alpha/2}(X) \leq Y \leq q_{1-\alpha/2}(X)) | X) = 1 - \alpha
\]

under minimal assumptions on the joint distribution of \((X, Y)\). (A sufficient condition would be the existence of a joint density with respect to Lebesgue measure.) If the problem is constrained to linear regression or M-estimation, then the estimators resulting from empirical risk minimization are consistent [23]; however, under our minimal assumptions, it is not known whether \( N_{D', \tau}(x) \) consistently estimates \((q_{\alpha/2}(x), q_1/2(x), q_{1-\alpha/2}(x)) \). That being said, we conjecture that these desirable properties are satisfied in an ad-defined asymptotic setting: First, neural networks are universal approximators (cf. Cybenko [5], Hornik [10]). Second, given a sufficiently large data set, and under the assumption that we can minimize the empirical risk, a neural network is able to learn the conditional mean of \( Y \) given \( X \) (cf. Bauer and Kohler [13]). This conjecture is supported by simulation evidence in Section [3] where we observe that \( N_{D', \tau}(x) \to (q_{\alpha/2}(x), q_1/2(x), q_{1-\alpha/2}(x)) \) as \( |D| \to \infty \).

## 2 Construction of the PI-Network

As the network \( N_{D, \alpha}(x) \) does not generally provide the desired properties in finite samples, we provide two modifications with finite sample coverage guarantees based on sample splitting. Let \( D_1 \) and \( D_2 \) be a random partition of \( D \) into two disjoint sets. \( D_1 \) is used to select and train the network and \( D_2 \) is used to adjust the resulting intervals to provide coverage guarantees.

In this paper, we focus on neural networks with a fixed architecture that has a \( d \)-dimensional input layer and a \( 3 \)-dimensional output layer. The term architecture comprises the network design (e.g., number and depth of hidden layers or dropout layers) and training parameters (e.g., number of epochs, batch size, or regularization parameters). See Goodfellow et al. [7] and citations therein.
Theorem 1. These facts can be used in a conformal inference procedure to calibrate the PI-Network. In essence, a set of \( \hat{c} \) values is selected to calibrate, we suggest setting \( \hat{c} = c(k) \) where \( k = \lceil (1-\alpha)|D_2| + 1 \rceil \) and \( c(k) \) is the \( k \)th order statistic. Set \( \hat{l}_x(\tau) = m_\tau(x) - \hat{c}(m_\tau(x) - l_\tau(x)) \) and \( \hat{u}_x(\tau) = m_\tau(x) + \hat{c}(u_\tau(x) - m_\tau(x)) \) for all \( (x,\tau) \in D_2 \). Note that Theorem 1 also holds for any data-dependent architecture, as long as the dependence is only on \( D_1 \). As all networks in this paper are fit using \( D_1 \), we simplify our notation, setting \( N_\tau(x) = N_{D_1,\tau}(x) \) and analogously for \( l_\tau(x), m_\tau(x) \) and \( u_\tau(x) \).

2.1 PI-Network with average coverage

To achieve average coverage, we use methods derived from conformal inference \cite{8,17}. In general, conformal prediction intervals require a fixed prediction procedure that is refit on an augmented data set. As it would clearly be infeasible to refit a large network many times, this process can be simplified by using classical sample splitting. Lei et al. \cite{17} refer to such methods as split-conformal. The interval \( [l_\tau(x), u_\tau(x)] \) given by \( N_\tau(x) \) will approximately capture the properties of the underlying conditional distribution of \( Y \) given \( X = x \), but the intervals may not provide average coverage. In order to achieve provably average coverage, we can use \( D_2 \) to adjust the interval using an expansion constant \( c \in (0, \infty) \).

Note that for a given \( x \in \mathbb{R}^d \) and \( y \in \mathbb{R} \), the interval \([l_\tau(x), u_\tau(x)]\) may not contain \( y \); however, with \( c \) given by

\[
c = \max \left( \frac{m_\tau(x) - y}{m_\tau(x) - l_\tau(x)}, \frac{y - m_\tau(x)}{u_\tau(x) - m_\tau(x)} \right),
\]

the interval \( \Gamma^c_x(x) = [m_\tau(x) - c(m_\tau(x) - l_\tau(x)), m_\tau(x) + c(u_\tau(x) - m_\tau(x))] \) does contain \( y \). These facts can be used in a conformal inference procedure to calibrate the PI-Network. In essence, a \( \hat{c} \in (0, \infty) \) is chosen such that at least \((1-\alpha)\)100% of the observations in the hold-out data, \( D_2 \), are contained in the interval \( \Gamma^c_x(X) \). Details are given in Algorithm 1. Note that if there are ties among the \( c_i \)’s defined in Algorithm 1 we use a tie breaking rule so that \( \hat{c} \) is always well-defined.

Theorem 1. Let \( N_\tau(x) = N_{D_1,\tau}(x) \) be as in Definition 2 and \( N_\tau^c(x) \) be the result of Algorithm 1. Set \( \Gamma^c_\alpha(x) = [l_\tau^c(x), u_\tau^c(x)] \). Then

\[
\mathbb{P}(Y \in \Gamma^c_\alpha(X)) \geq 1 - \alpha.
\]

Note that Theorem 1 also holds for any data-dependent \( \hat{\tau} \) as long as the dependence of \( \hat{\tau} \) on the data is only through \( D_1 \). As \( \tau \) controls the width of the estimated interval and that is precisely what \( \hat{c} \) is selected to calibrate, we suggest setting \( \tau = \alpha \) in practice. In the simulations of Section 3 we typically observe \( \hat{c} \geq 1 \) for \( N_\alpha(x) \), but \( \hat{c} \rightarrow 1 \) as \(|D_1|\) increases, suggesting that \( N_\alpha(x) \) has asymptotic, \( 1 - \alpha \) average coverage.

Proof of Theorem 1. Because the network \( N_\tau(x) \) is fit on \( D_1 \) which is independent of \( D_2 \), the statistics \( c_i \) are i.i.d. conditional on \( D_1 \). For the new observation \((X, Y)\), set

\[
c' = \max \left( \frac{m_\tau(X) - Y}{m_\tau(X) - l_\tau(X)}, \frac{Y - m_\tau(X)}{u_\tau(X) - m_\tau(X)} \right).
\]

Conditional on \( D_1 \), \( c' \) is independent of \( c_i \), has the same distribution as \( c_i \), and the rank of \( c' \) among the \( c_i \)'s is uniform over the set of integers \( \{1, 2, \ldots, |D_2|\} \). Therefore, \( \mathbb{P}(c' > c(k)) \mid D_1 \) \( \leq \alpha \), where \( k = \lceil (1-\alpha)|D_2| + 1 \rceil \) and \( c(k) \) is the \( k \)th order statistic. Therefore \( \mathbb{P}(Y \in \Gamma^c_\alpha(X) \mid D_1) \geq 1 - \alpha \), which implies the claim of the theorem. \( \square \)
If ties among \( c_i \)’s only happen on a set of measure 0, or if we use a random tie-breaking rule, the probability of coverage can be upper bounded by \( 1 - \alpha + 1/(|D_1| + 1) \), meaning that the intervals are not unnecessarily wide (cf. Theorem 2.2 of Lei et al. [17]). Algorithm 1 is a specialized version of the Split-Conformal algorithm of Lei et al. [17]. Our statistics \( c_i \) are tailored to our PI-Network, whereas the previous algorithm uses the residuals \( R_i = [Y_i - m(X_i)] \) or measures tailored to non-parametric inference. Using the residuals \( R_i \) instead of \( c_i \) in Algorithm 1 produces valid intervals, but they do not adapt to heteroskedasticity or asymmetry. This classical sample-splitting approach was identified for neural networks as early as Papadopoulos and Haralambous [20].

### 2.2 Probably Approximately Valid PI-Network

In order to provide PAV intervals, we select \( \hat{\tau} \) using \( D_2 \) so that \( l_{\hat{\tau}}(x) \) and \( u_{\hat{\tau}}(x) \) adaptively estimate the quantiles \( q_{\alpha/2}(x) \) and \( q_{1-\alpha/2}(x) \). This adaptation is required as we are not guaranteed that \( l_{\alpha}(x) \) and \( u_{\alpha}(x) \) accurately estimate these quantities. Let \( \hat{p}(N_{\hat{\tau}}, D_2) \) be the empirical coverage probability of the neural network \( N_{\hat{\tau}} \) on the data set \( D_2 \), i.e.,

\[
\hat{p}(N_{\hat{\tau}}, D_2) = \frac{1}{|D_2|} \sum_{(X_i, Y_i) \in D_2} 1_{Y_i \in [l_{\hat{\tau}}(X_i), u_{\hat{\tau}}(X_i)]}.
\]

We choose \( \hat{\tau} \) such that the network \( N_{\hat{\tau}}(x) \) has \( 1 - \alpha \) coverage on \( D_2 \), i.e., such that \( \hat{p}(N_{\hat{\tau}}, D_2) \geq 1 - \alpha \). \( N_{\hat{\tau}}(x) \) provides PAV prediction intervals, as formalized in the following theorem.

**Theorem 2.** Let \( N_{\tau}(x) = N_{D_1, \tau}(x) \) be as in Definition 2. Let \( G_K \subseteq [0, 1] \) such that \( 0 \in G_K \) and \( |G_K| = K + 1 \). Set \( n_2 = |D_2| \). Let \( \hat{\tau} = \max\{\tau \in G_K : \hat{p}(N_{\hat{\tau}}, D_2) \geq 1 - \alpha \} \) and \( \Gamma_{\alpha}^+(x) = [l_{\hat{\tau}}(x), u_{\hat{\tau}}(x)] \). Set \( p(G_{\alpha}^+|D) = P(Y \in \Gamma_{\alpha}^+(X) \mid D) \). Then,

\[
P(\ Gamma_{\alpha}^+|D) \leq 1 - \alpha - \epsilon \leq K \exp(-2\epsilon^2 n_2).
\]

(\( \hat{\tau} \) is always well defined by definition of the network \( N_{\alpha}(x) \).

The theorem continues to hold for any data-dependent grid, \( \hat{G}_K \), as long as \( 0 \in \hat{G}_K \) and the dependence of \( \hat{G}_K \) on the data is only through \( D_1 \). Given the tendency of flexible models such as neural networks to over-fit the data, \([l_{\alpha}(x), u_{\alpha}(x)]\) typically under-covers in finite samples. As such, we suggest setting \( G_K \subseteq [0, \alpha] \), and in the simulations of Section 3 we typically observe \( \hat{\tau} \leq \alpha \) but that \( \tau \to \alpha \) from below as \( |D_1| \) increases. By solving the equation \( \delta = K \exp(-2\epsilon^2 n_2) \), we get that for all \( n_2 \geq n(\delta, \delta, K) = -\log(\delta/K)/(2\epsilon^2) \),

\[
P(\ Gamma_{\alpha}^+|D) \leq 1 - \alpha - \epsilon \leq \delta.
\]

**Proof of Theorem 2.** By definition of \( N_{\hat{\tau}}(x) \), we have

\[
P(\ Gamma_{\alpha}^+|D) \leq 1 - \alpha - \epsilon \leq P(\ Gamma_{\alpha}^+|D) \leq \hat{p}(N_{\hat{\tau}}, D_2) - \epsilon)
\]

\[
= E[\ Gamma_{\alpha}^+|D) \leq \hat{p}(N_{\hat{\tau}}, D_2) - \epsilon | D_1)]
\]

We use the union bound to bound the conditional probability within the expectation from above by

\[
\sum_{\tau \in \hat{G}_K} P(\Gamma_{\alpha}^+(x) | D) \leq \hat{p}(N_{\hat{\tau}}, D_2) - \epsilon \mid D_1).
\]

For \( \tau = 0 \), we have \([l_{\tau}(x), u_{\tau}(x)] = [-\infty, \infty] \) and the summand in the previous expression is equal to 0. For \( \tau \neq 0 \), we have \( P([l_{\tau}(x), u_{\tau}(x)] \mid D) = p([l_{\tau}(x), u_{\tau}(x)] \mid D_1) \), because the first conditional probability is independent of \( D_2 \). Observe that, conditional on \( D_1 \), \( \hat{p}(N_{\hat{\tau}}, D_2) \) is the mean of \( n_2 \) Bernoulli-trials with mean \( p([l_{\tau}(x), u_{\tau}(x)] \mid D_1) \). By Hoeffding’s inequality, the corresponding summand in the previous expression is bounded by \( \exp(-2\epsilon^2 n_2) \).}

Even though the PAV interval \( \Gamma_{\alpha}^+(x) \) does not generally provide average coverage, this can be achieved by defining a more conservative PAV interval.
We simulate covariates $X$. We compare our methods to the "fixed-width" conformal (conf-fw) method of Papadopoulos and where $f$ as well as to tune the parameters of the loss functions for high-q and neg-ll. All data experiments set $\alpha = .1$ and calculate results using $D_3$, which was unseen by the models. The $\hat{\tau}$ in $\Gamma_{\alpha}^c(X)$ is selected on the grid $G_{10} = \{1, .09, \ldots, .01, 0\}$ and we set $\tau = \alpha$ when constructing $\Gamma_{\alpha}^c(X)$. In each data example, the same network architecture is used for all methods and is trained in python using the pytorch library [21]. All computations were performed on a Google-Cloud-Platform instance with a NVIDIA Tesla K80 GPU.

The method by Nix and Weigend [19] fits a network with a two-dimensional output that estimates the mean function $\hat{\mu}(x)$ and variance function $\hat{\sigma}^2(x)$ by minimizing the negative log-likelihood of the normal distribution. The method by Pearce et al. [22] fits a network with a two-dimensional output that estimates the lower and upper bound of the PI via a loss function that penalizes in-sample mis-coverage and interval length. We use the soft loss function as described in Algorithm 1 of their paper. The midpoint of the interval is used as the prediction.

### 3.1 Artificial data

We simulate covariates $X \in [0, 1]^{100}$, where each entry is drawn i.i.d. from a standard uniform distribution. The response is given by

$$Y = f(\beta'X) + \epsilon, \quad \epsilon \sim N(0, 1 + (\beta'X)^2),$$

where $f(x) = 2\sin(\pi x) + \pi x$ and only the first 5 components of $\beta \in \{0, 1\}^{100}$ are non-zero and equal to 1. This is a challenging setting because the model is sparse, non-linear in $X$, and heteroskedastic in $Y$ given $X$. We compare all the methods to an oracle that knows this data-generating process. The oracle knows the true conditional median, $q_{1/2}(X) = f(\beta'X)$, as well as the uniformly most accurate, unbiased PI,

$$[q_{\alpha/2}(X), q_{1-\alpha/2}(X)] = [f(\beta'X) - z_{\alpha/2}\sqrt{1 + (\beta'X)^2}, f(\beta'X) + z_{\alpha/2}\sqrt{1 + (\beta'X)^2}],$$

where $z_{\alpha/2}$ is the $(1 - \alpha/2)$-quantile of the standard normal distribution. We generate $n = |D_1 \cup D_2|$ samples, where $n$ ranges between 5,000 and 100,000, and use 3/4 of the data for training and 1/4 for validation. An independent data set of size 20,000 is used for testing. For each sample size $n$, we repeat the experiment 10 times. For each method, we train a neural network with one hidden layer, 200 hidden nodes, and using between 80 and 100 passes through the data.

Simulation results are summarized in Figure 1. Panel 1 demonstrates that our methods do indeed provide average coverage in finite samples. It is unsurprising that neg-ll provides approximate coverage in this case, as it is tailored to the normal distribution. The poor performance of high-q for large samples is largely due to instability: its empirical coverage probability oscillates between approximately .8 and 1. These oscillations are smoothed in our graph.

Panels 2 and 3 capture the effect of heteroskedasticity, both on empirical coverage probability and quantile estimation. Panel 2 shows that all methods provide coverage conditional on $\beta'X$ at least.
in the center of the $\beta'X$ distribution. Outside this region, however, competing methods suffer considerably: none of them are wide enough to provide coverage for large values of $\beta'X$. Both pav and conf-nn provide approximate average coverage over the entire range of $\beta'X$. Panel 3 shows the total median absolute deviation for estimating conditional quantiles: $q_{1/2}(X)$, $q_{\alpha/2}(X)$, and $q_{1-\alpha/2}(X)$. Throughout the entire range of sample sizes, our methods perform as well as neg-ll, which was crafted precisely for our data generating distribution. While high-q and conf-fw can provide average coverage, they do not estimate the quantiles well.

### 3.2 Real Data

We perform similar comparisons using two public data sets from the data science platform Kaggle.com. The significantly different nature of these data sets demonstrates the flexibility of the PI-Network. The first data set is a straightforward regression example of sale prices for homes in King County, Washington, between May 2014 and May 2015. This data set includes 21,613 observations which we split s.t. $|D_1| = 15,000$, $|D_2| = 3,000$ and $|D_3| = 3,613$. There are 19 covariates describing the features of the house which are used to predict log sale price. All covariates are standardized to have mean 0 and standard deviation 1. For each method we trained a neural network with one hidden layer, with 100 hidden nodes, and between 80 and 200 passes through the data.

In the second data set, the goal is to estimate the age of a child (in months) given their sex and an X-ray of their hand. This data set includes 12,611 observations which we split s.t. $|D_1| = 7,500$, $|D_2| = 2,500$ and $|D_3| = 2,611$. We used data augmentation (random rotation and horizontal flips of the images) on $D_1$ to reduce over-fitting. We use an intermediate layer of the pretrained Inception V3 network as the feature extractor [24]. We trained a neural network on the extracted features with one hidden layer, 300 hidden nodes, and between 30 and 100 passes through the data.

The real data results are summarized in Table 1 and more details are shown in Figure 2. Table 1 shows the empirical coverage, average length, interquartile range (IQR), and MAD for each method. Observe that all methods provide average coverage at $\alpha = .1$ and the errors are comparable. On the bone-age data, both high-q and neg-ll produce extremely wide intervals. We found that training the network with the high-q and neg-ll losses was very unstable, leading either to very short or very long intervals. We assume that this instability arises because the loss functions require computing more complex derivatives. In some cases, both for conf-na and high-q, it would appear as though they achieve similar coverage and error with intervals of similar average length. These summary statistics hide the inefficiency of both methods: they have constant or nearly constant interval width. As such, they are failing to provide precise intervals when it is possible to do so.

---

1 Available at: [www.kaggle.com/harlfoxem/housesalesprediction](http://www.kaggle.com/harlfoxem/housesalesprediction), [www.kaggle.com/kmader/rsna-bone-age](http://www.kaggle.com/kmader/rsna-bone-age)
Table 1: Summary statistics of data applications

| Data Set   | Method | Ave-Coverage | Ave-Length | IQR-Length | MAD  |
|------------|--------|--------------|------------|------------|------|
|            | pav    | 0.88         | 0.54       | 0.21       | 0.10 |
| Home-Price | conf-nn| 0.91         | 0.56       | 0.16       | 0.11 |
|            | high-q | 0.92         | 0.84       | 0.04       | 0.12 |
|            | neg-ll | 1.00         | 5.15       | 1.58       | 0.14 |
|            | conf-na| 0.93         | 0.68       | 0.00       | 0.11 |
| Bone-Age   | pav    | 0.93         | 85.62      | 19.39      | 16.08|
|            | conf-nn| 0.90         | 78.12      | 20.10      | 16.26|
|            | high-q | 1.00         | 298.69     | 23.68      | 17.78|
|            | neg-ll | 1.00         | 187.36     | 13.11      | 17.33|
|            | conf-na| 0.90         | 78.88      | 0.00       | 16.62|

Figure 2: The x-axis is formed by binning the conf-nn PI lengths into 100 bins. The average coverage is computed for each bin and the results are smoothed. A scaled density plot of the lengths of the conf-nn PIs is also included.

It is instructive to consider the coverage probability of each method on the set of observations where our intervals are narrow or wide. In Figure 2, we use binned the observations based on the length of the conf-nn PIs. As there are a few outliers of the PI-width, we removed 1% of the values from each tail. For both data sets, our intervals provide coverage through most of the domain of their widths: they are narrow when the can be and are wide when they must be. Furthermore, when they are narrow they still provide coverage at close to the nominal level. This provides empirical evidence that the PI-Network is estimating conditional quantiles well. In the range where our intervals are narrow, however, the other methods have coverage near 1: the intervals are far too wide. Similarly, when our intervals are wide, high-q and conf-na fail to provide coverage as the intervals are too narrow.

4 Discussion

In this paper we demonstrate the power and ease of use of sample splitting approaches. We envision the PI-Network as a commonplace addition to many deep networks in order to express the uncertainty of their predictions without compromising predictive accuracy. As data sets grow in size, the cost of using a validation set decreases, making the observation of Barnard [2] even more accurate today:

The simple idea of splitting a sample into two and then developing the hypothesis on the basis of one part and testing it on the remainder may perhaps be said to be one of the most seriously neglected ideas in statistics, if we measure the degree of neglect by the ratio of the number of cases where a method could give help to the number of cases where it is actually used.

8
References

[1] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature Biotechnology*, 33:831–838, 2015.

[2] G. A. Barnard. Discussion of “cross-validatory choice and assessment of statistical predictions,” by M. Stone. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36:133–135, 1974.

[3] B. Bauer and M. Kohler. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, 47:2261–2285, 2019.

[4] R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, pages 160–167, 2008.

[5] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2:303–314, 1989.

[6] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059, 2016.

[7] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[8] T. Heskes. Practical confidence and prediction intervals. In *Advances in Neural Information Processing Systems*, pages 176–182, 1997.

[9] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, B. Kingsbury, et al. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29, 2012.

[10] K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4:251–257, 1991.

[11] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[12] G. Keren, N. Cummins, and B. Schuller. Calibrated prediction intervals for neural network regressors. *IEEE Access*, 6:54033–54041, 2018.

[13] A. Khosravi, S. Nahavandi, D. Creighton, and A. F. Atiya. Lower upper bound estimation method for construction of neural network-based prediction intervals. *IEEE Transactions on Neural Networks*, 22:337–346, 2011.

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

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[16] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413, 2017.

[17] J. Lei, M. G’Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113:1094–1111, 2018. doi: 10.1080/01621459.2017.1307116.

[18] D. J. MacKay. A practical bayesian framework for backpropagation networks. *Neural Computation*, 4:448–472, 1992.

[19] D. A. Nix and A. 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, 1994.
[20] H. Papadopoulos and H. Haralambous. Reliable prediction intervals with regression neural networks. *Neural Networks*, 24:842–851, 2011.

[21] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *NIPS Autodiff Workshop*, 2017.

[22] T. Pearce, M. Zaki, A. Brintrup, and A. Neely. High-quality prediction intervals for deep learning: a distribution-free, ensembled approach. *arXiv preprint arXiv:1802.07167*, 2018.

[23] G. Shafer and V. Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

[24] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[25] N. Tagasovska and D. Lopez-Paz. Frequentist uncertainty estimates for deep learning. *arXiv preprint arXiv:1811.00908*, 2018.

[26] L. G. Valiant. A theory of the learnable. In *Proceedings of the sixteenth annual ACM Symposium on Theory of Computing*, pages 436–445. ACM, 1984.

[27] V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer-Verlag, Berlin, Heidelberg, 2005.

[28] J. M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*, volume 1. The MIT Press, 2001.