Learning Prediction Intervals for Regression: Generalization and Calibration

Haoxian Chen
IEOR
Columbia University

Ziyi Huang
EE
Columbia University

Henry Lam
IEOR
Columbia University

Huajie Qian
Alibaba Group

Haofeng Zhang
IEOR
Columbia University

Abstract

We study the generation of prediction intervals in regression for uncertainty quantification. This task can be formalized as an empirical constrained optimization problem that minimizes the average interval width while maintaining the coverage accuracy across data. We strengthen the existing literature by studying two aspects of this empirical optimization. First is a general learning theory to characterize the optimality-feasibility tradeoff that encompasses Lipschitz continuity and VC-subgraph classes, which are exemplified in regression trees and neural networks. Second is a calibration machinery and the corresponding statistical theory to optimally select the regularization parameter that manages this tradeoff, which bypasses the overfitting issues in previous approaches in coverage attainment. We empirically demonstrate the strengths of our interval generation and calibration algorithms in terms of testing performances compared to existing benchmarks.

1 Introduction

While most literature in machine learning focuses on point prediction, uncertainty quantification plays, arguably, an equally important role in reliability assessment and risk-based decision-making. In regression, a natural approach to quantify uncertainty is the prediction interval (PI), namely an upper and lower limit for a given feature value X that covers the corresponding outcome Y with high probability. The interval center represents the expected outcome, whereas the width represents the uncertainty. A test point with a high expected outcome, but wide PI width, means that the outcome can still be low with a significant chance, thus signifies a downside risk that should not be overlooked.

In this paper, we study the construction of PIs that satisfy an overall target prediction level across data, known as the expected coverage rate (Rosenfeld et al., 2018). Compared to widely used conditional (on X) coverage rate, this notion is advantageously more tangible to measure and easier to achieve. This means that a much wider class of models can be trained to build PIs, as less conditions are needed to impose on the true relation and the model class to obtain satisfactory guarantees. In general, constructing a good PI in this framework requires balancing a tradeoff between the expected interval width and coverage maintenance, which can be formalized as an empirical constrained optimization. This viewpoint has been used in Khosravi et al. (2010) and Pearce et al. (2018) that focus on neural networks, Rosenfeld et al. (2018) that studies a dual formulation, and Galván et al. (2017) that uses multi-objective evolutionary optimization. It also relates to the learning of minimum volume sets (Polonik, 1997; Scott and Nowak, 2006) in which a similar tradeoff between volume and probability content appears. Building on these works, our goal in this paper is to study two key inter-related statistical aspects of this empirical constrained optimization that enhances previous results both in theory and in practice:

Feasibility-Optimality Tradeoff for Interval Models. We develop a learning theory for the PIs constructed from empirical constrained optimization that statistically achieves both feasibility (coverage) and optimality (interval width). Methodologically, we build a general "sensitivity measure" that controls this tradeoff, which in turn requires developing deviation bounds for simultaneous empirical processes. Our theory in particular covers the Lipschitz continuous model class (in parameter) and finite Vapnik–Chervonenkis (VC)-subgraph class, exemplified by a wide class of

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

neural networks and regression trees. Such type of joint coverage-width learning guarantees appears the first in the literature. It expands the coverage-only results and the considered model classes in Rosenfeld et al. (2018). It also generalizes Scott and Nowak (2006) as both our constraints and objectives possess extra sophistication related to the shape requirement of the set as an interval, and also we characterize feasibility and optimality attainments explicitly instead of the implicit metric in Scott and Nowak (2006).

Calibration Method and Performance Guarantees. We propose a general-purpose, ready-to-implement calibration methodology to guarantee overall PI coverage with prefixed confidence. This approach is guided by a novel utilization of the high-dimensional Berry-Esseen theorem (Chernozhukov et al., 2017). It is designed to combat the overfitting issue of interval models and perform accurately on the *test* set. We demonstrate empirically how our approach either outperforms other methods in terms of achieving correct coverages or, for those methods with comparable coverages, we attain shorter interval widths. Moreover, our approach applies, with little adjustment, to accurately construct multiple PIs at different prediction levels simultaneously. This adds extra flexibility for decision-makers to construct PIs without needing to pre-select the prediction level in advance.

2 Related Work

We first review two most closely related methods, and then move on to other works.

Conformal Learning (CL). First proposed in Vovk et al. (2005), conformal learning (CL) is a class of methods that leverage data exchangeability to constructs PIs with finite-sample and distribution-free coverage guarantees. The original CL requires retraining for each possible test point and is therefore computationally prohibitive in general. Split/inductive CL (Papadopoulos, 2008; Lei et al., 2015, 2018) improves the computational efficiency based on a holdout validation that avoids retraining, but at the cost of higher variability and wider intervals due to less efficient data use. Lying in between are variants based on more efficient cross-validation schemes, including leave-one-out (or the Jackknife; Barber et al. (2019); Alaa and van der Schaar (2020); Steinberger and Leeb (2016)), K-fold (Vovk, 2015) and ensemble methods (Gupta et al., 2019; Kim et al., 2020). Recently, quantile regression are combined with CL (Kivaranovic et al., 2020; Romano et al., 2019) to take into account the heterogeneity of uncertainties across feature values. Despite its generality, the coverage guarantees from CL are only marginal with respect to the training data (except split CL (Vovk, 2012)), whereas our proposed calibration method provides a stronger high confidence guarantee. Moreover, our approach explicitly optimizes the interval width, therefore typically generates shorter PIs than CL.

Quantile Regression (QR). Quantile regression (QR) estimates the conditional quantiles of Y that can be used to construct PIs. Classical QR methods require strong assumptions (e.g., linearity or other parametric forms; Chapter 4 in Koenker and Hallock (2001)). Approaches that relax these assumptions include quantile regression forests (QRF) (Meinshausen, 2006) and kernel support vector machine (SVM) (Steinwart et al., 2011). However, little is known about their finite-sample coverage performance because of estimation errors in the quantiles. Recently, Kivaranovic et al. (2020) proposes calibrating the weight parameter in the pinball loss on a holdout data set to enhance PI coverage. This calibration scheme, however, does not address overfitting on the holdout set, thus could fall short of providing correct coverages on test data as our experiments show.

Other Approaches. PI construction has been substantially studied in classical statistics. To understand this construction, the error of a (point) prediction can typically be decomposed into two components: model uncertainty, which comes from the variability in the training data or method, and outcome uncertainty, which comes from the noise of Y conditional on X. The classical literature often assumes well-defined and simple forms on the relation between X and Y (e.g., linear model, Gaussian; Seber and Lee 2012). In this case, model uncertainty reduces to parameter estimation errors. Outcome uncertainty, on the other hand, is intrinsic in the population distribution but not the training, i.e. it arises even if the model is perfectly trained. An array of methods account for both sources of variability, which utilize approaches ranging from asymptotic normality (Seber and Lee, 2012), deconvolution (Schmoyer, 1992), and resampling schemes such as the bootstrap (Stine, 1985) and jackknife (Steinberger and Leeb, 2016).

To overcome the strong assumptions in classical statistical models, several model-free approaches have been developed. Nonparametric regression, such as spline or kernel-based methods (Doksum and Koo, 2000; Olive, 2007), removes rigid model assumptions but at the expense of strong dimension dependence. Gaussian processes or kriging-based methods (Sacks et al., 1989), popular in the areas of metamodeling and computer emulation, regard outcomes as a response surface and perform Gaussian posterior updates. In particular, stochastic kriging (SK) model (Ankenman et al., 2010) constructs PIs that account for both model and output variabilities by using a prior correlation structure that entails both. However, SK does not deliver a frequentist coverage guarantee nor convergence rate. More recently, uncertainty quantification of neural networks regarding their model and output variabilities are studied, via methods such as the delta method and the bootstrap (Papadopoulos et al., 2001; Khosravi et al., 2011). Nonetheless, like in classical statistical models, these approaches can only capture variability due to data and training noises, but not the bias against the true relation.

Lastly, our PI construction follows the *high-quality* criterion in works including Khosravi et al. (2010, 2011); Galván et al. (2017); Pearce et al. (2018); Rosenfeld et al. (2018); Zhang et al. (2019); Zhu et al. (2019), which propose various loss functions to capture the width-coverage tradeoff. They are also related to the highest density intervals in statistics (Box and Tiao, 2011). Our investigations in this paper provide theoretical guarantees in using this criterion.

3 PI Learning as Empirical Constrained Optimization

We consider the general regression setting where $X \in \mathcal{X} \subset \mathbb{R}^d$ is the feature vector and $Y \in \mathcal{Y} \subset \mathbb{R}$ is the outcome. Given an i.i.d. data set $\mathcal{D} := \{(X_i, Y_i)\}_{i=1,\dots,n}$ each drawn from an unknown joint distribution π, our goal is to find a PI defined as:

Definition 3.1. *An interval $[L(x), U(x)]$, where both $L, U : \mathcal{X} \to \mathbb{R}$, is called a prediction interval (PI) with (overall) coverage rate $1-\alpha$ ($0 \le \alpha \le 1$) if $\mathbb{P}_\pi(Y \in [L(X), U(X)]) \ge 1-\alpha$, where $\mathbb{P}_\pi$ denotes the probability with respect to the joint distribution π.*

We aim to find two functions L and U, both in a hypothesis class $\mathcal{H}$. To learn the optimal high-quality PI that attains a given coverage rate $1-\alpha$, we consider the following constrained optimization problem:

$$\begin{aligned} &\min_{L,U\in\mathcal{H} \text{ and } L\le U} \ \mathbb{E}_{\pi_X}[U(X)-L(X)] \\ &\quad \text{subject to } \mathbb{P}_\pi(Y \in [L(X), U(X)]) \ge 1-\alpha \end{aligned} \tag{3.1}$$

where $\mathbb{E}_{\pi_X}$ denotes the expectation with respect to the marginal distribution of X. Given the data $\mathcal{D}$, we approximate (3.1) with the following empirical constrained optimization problem

$$\begin{aligned} \widehat{\text{opt}}(t): &\min_{L,U\in\mathcal{H} \text{ and } L\le U} \ \mathbb{E}_{\hat\pi_X}[U(X)-L(X)] \\ &\quad \text{subject to } \mathbb{P}_{\hat\pi}(Y \in [L(X), U(X)]) \ge 1-\alpha+t \end{aligned} \tag{3.2}$$

parameterized by $t \in [0, \alpha]$, where $\mathbb{E}_{\hat\pi_X}$, $\mathbb{P}_{\hat\pi}$ are expectation and probability with respect to the empirical distribution constructed from the data $\mathcal{D}$, e.g., $\mathbb{E}_{\hat\pi_X}[U(X)-L(X)] = \frac{1}{n}\sum_{i=1}^n (U(X_i)-L(X_i))$. The adjustment t, which can be viewed as a penalty term for the empirical coverage rate, is used to boost the generalized coverage performance for the optimal interval solved from $\widehat{\text{opt}}(t)$: If no adjustment is made in the constraint ($t=0$), then, because of noise, the true coverage can be lower than $1-\alpha$ with significant probability even if the empirical coverage is above $1-\alpha$. A positive t decreases the probability of such an event. Choosing t too large, however, would eliminate more intervals from the feasible set, leading to a deterioration in the obtained expected width (objective). One of our main investigations is to balance the coverage and width performance by judiciously choosing t.

We point out that, while our focus is on training L and U directly, our approach can also be applied naturally when we are given in advance a well-trained point predictor $f : \mathcal{X} \to \mathbb{R}$ (obtained by any means independent of $\mathcal{D}$). In this case we may seek two non-negative functions $\delta_i : \mathcal{X} \to [0, \infty)$ such that $[L(x), U(x)] = [f(x) - \delta_1(x), f(x) + \delta_2(x)]$. Our subsequent development applies by constructing lower and upper bounds for the "translational" data set $\tilde{\mathcal{D}} := \{(X_i, Y_i - f(X_i))\}_{i=1,\dots,n}$.

4 Joint Coverage-Width Learning Guarantees

We establish finite-sample generalization error bounds for two major classes: 1) finite VC dimensions, and 2) Lipschitz continuous in parameters, by building on a unified "sensitivity bound" on the oracle optimization. Corresponding results on consistency are provided in Appendix A. To begin, let $R_t^*(\mathcal{H})$ be the optimal value of

$$\begin{aligned} \text{opt}(t): &\min_{L,U\in\mathcal{H} \text{ and } L\le U} \ \mathbb{E}_{\pi_X}[U(X)-L(X)] \\ &\quad \text{subject to } \mathbb{P}_\pi(Y \in [L(X), U(X)]) \ge 1-\alpha+t \end{aligned} \tag{4.1}$$

which is (3.1) but with a higher target prediction level, and correspondingly $R^*(\mathcal{H})$ be the optimal value of (3.1). We make the following assumptions on the hypothesis class $\mathcal{H}$, and the conditional distribution of Y given X:

Assumption 1. *For every function $h \in \mathcal{H}$, we have $h + c \in \mathcal{H}$ for every constant $c \in \mathbb{R}$.*

Assumption 2. *The conditional distribution of Y given $X = x$ has a density $p(\cdot|x)$ for every $x \in \mathcal{X}$. Moreover, for every x, y such that $\mathbb{P}_\pi(Y \le y|X = x) \in (0, 1)$, we have $p(y|x) > 0$ and that $p(\cdot|x)$ is continuous at y.*

The simple closedness property in Assumption 1 turns out to allow sufficiently tight and tractable analysis for many, potentially complicated, function classes under the same roadmap. Many common classes (e.g., linear, piece-wise constant such as tree-based models, and neural networks with linear activation in the output unit) satisfy Assumption 1. It is also straightforward to enforce a class to satisfy Assumption 1 by simply adding one extra parameter. Assumption 2 is a mild non-degeneracy condition on the conditional distribution (e.g., Assumption 2 holds when $p(\cdot|x)$ is Gaussian or uniform over a certain interval for each x).

We have the following upper bound for $R_t^*(\mathcal{H}) - R^*(\mathcal{H})$ for $0 < t < \alpha$:

Theorem 4.1 (Sensitivity bound with respect to target prediction level)**.** *Suppose Assumptions 1-2 hold. For every $x \in \mathcal{X}$ and $\beta \in (0,1)$, let $\Gamma(x,\beta) := \inf\big\{p(y_1|x) + p(y_2|x) : \mathbb{P}_\pi(y_1 \le Y \le y_2 | X = x) \le 1-\beta\big\}$ and $\gamma_\beta := \sup\{z > 0 : \mathbb{P}_{\pi_X}(\Gamma(X,\beta) < z) \le \beta\}$. Then, for every $\alpha \in (0,1)$ and $t \in (0,\alpha)$, we have $R_t^*(\mathcal{H}) - R^*(\mathcal{H}) \le 6t/((\alpha - t)\gamma_{(\alpha-t)/3})$.*

We briefly explain how Theorem 4.1 helps develop generalization guarantees. Let $\mathcal{H}_t^2, \hat{\mathcal{H}}_t^2 \subset \mathcal{H}^2 := \mathcal{H} \times \mathcal{H}$ be the feasible set of $\mathrm{opt}(t)$ and $\widehat{\mathrm{opt}}(t)$ respectively. If a uniform error bound Δt over the class $\mathcal{H}^2$ can be established for the empirical coverage constraint in (3.2), then $\mathcal{H}_{t+\Delta t}^2 \subset \hat{\mathcal{H}}_t^2$, therefore the shortest interval we can potentially learn from $\widehat{\mathrm{opt}}(t)$ can only be wider than the oracle optimum from (3.1) by at most $R_{t+\Delta t}^*(\mathcal{H}) - R^*(\mathcal{H})$. The density lower bounds in Theorem 4.1 ensure a sufficient increase of the coverage probability as interval width increases, which inversely constrains the growth of interval width as the required coverage increases.

To derive finite-sample convergence rate, we first assume the availability of certain deviation bounds for the empirical coverage rate and interval width that appear in $\widehat{\mathrm{opt}}(t)$:

Theorem 4.2 (Rate of convergence)**.** *Suppose Assumptions 1-2 hold, and the following deviation bounds hold for every $\epsilon, t > 0$: $\mathbb{P}\big(\sup_{h\in\mathcal{H}}|\mathbb{E}_{\hat{\pi}_X}[h(X)] - \mathbb{E}_{\pi_X}[h(X)]| > \epsilon\big) \le \phi_1(n,\epsilon,\mathcal{H})$ and $\mathbb{P}\big(\sup_{L,U\in\mathcal{H} \text{ and } L\le U}|\mathbb{P}_{\hat{\pi}}(Y \in [L(X),U(X)]) - \mathbb{P}_\pi(Y \in [L(X),U(X)])| > t\big) \le \phi_2(n,t,\mathcal{H})$. Then for every $t \in (0,\frac{\alpha}{2})$ and $\epsilon > 0$, with probability at least $1 - \phi_1(n,\epsilon,\mathcal{H}) - \phi_2(n,t,\mathcal{H})$, we have, for every optimal solution $(\hat{L}_t^*, \hat{U}_t^*)$ of $\widehat{\mathrm{opt}}(t)$, that $\mathbb{P}_\pi(Y \in [\hat{L}_t^*(X), \hat{U}_t^*(X)]) \ge 1 - \alpha$ and $\mathbb{E}_{\pi_X}[\hat{U}_t^*(X) - \hat{L}_t^*(X)] \le \mathcal{R}^*(\mathcal{H}) + \frac{12t}{(\alpha-2t)\gamma_{(\alpha-2t)/3}} + 4\epsilon$.*

Theorem 4.2 translates the deviation bounds of two empirical processes into the probability of jointly achieving optimality and feasibility. With this, our focus is to derive deviation bounds for important hypothesis classes. Such bounds are well-understood in the literature, e.g., Chapter 2.14 in Van der Vaart and Wellner (1996) and Chapter 3 in Vapnik (2013). However, in our case, we need to go beyond the standard theory to control two empirical processes simultaneously by choosing a single function class $\mathcal{H}$. Next we present two fairly general choices of $\mathcal{H}$ for which exponential deviation bounds can be obtained for both processes.

To present the results, we introduce some terminologies. Let $\mathrm{vc}(\mathcal{S})$ be the VC dimension of a class $\mathcal{S}$ of sets. The VC dimension $\mathrm{vc}(\mathcal{G})$ of a class $\mathcal{G}$ of functions from $\mathcal{X}$ to $\mathbb{R}$ is defined to be the VC dimension of the set of subgraphs $\mathcal{S}_\mathcal{G} := \{\{(x,z) \in \mathcal{X}\times\mathbb{R} : z < g(x)\} : g \in \mathcal{G}\}$. $\mathcal{G}$ is called VC-subgraph if $\mathrm{vc}(\mathcal{G}) < \infty$. For a vector, $\|\cdot\|_p$ represents its l_p-norm for $p \ge 1$. For a function $\xi : \mathcal{X} \to \mathbb{R}$, we denote by $\|\xi\|_{\psi_2} := \inf\{c > 0 : \mathbb{E}_{\pi_X}[\exp(\xi^2(X)/c^2)] \le 2\}$ its sub-Gaussian norm under the distribution π_X, and denote by $\|\xi\|_p := \big(\mathbb{E}_{\pi_X}[|\xi(X)|^p]\big)^{1/p}$ its L_p norm.

Theorem 4.3 (Joint coverage-width guarantee for VC-subgraph class)**.** *If the hypothesis class $\mathcal{H}$ is such that the augmented class $\mathcal{H}_+ := \{h + c : h \in \mathcal{H}, c \in \mathbb{R}\}$ is VC-subgraph, and $H(x) := \sup_{h\in\mathcal{H}}|h(x) - \mathbb{E}_{\pi_X}[h(X)]|$ satisfies $\|H\|_{\psi_2} < \infty$, then the deviation bounds in Theorem 4.2 satisfy*

$$\phi_1(n,\epsilon,\mathcal{H}) \le 2\exp\Big(-\frac{n\epsilon^2}{C\|H\|_{\psi_2}^2 \mathrm{vc}(\mathcal{H}_+)}\Big),$$

$$\phi_2(n,t,\mathcal{H}) \le \begin{cases} 4^{n+1}\exp(-t^2 n) & \textit{if}\, n < \frac{C\mathrm{vc}(\mathcal{H})}{2} \\ 4\Big(\frac{2en}{C\mathrm{vc}(\mathcal{H})}\Big)^{C\mathrm{vc}(\mathcal{H})}\exp(-t^2 n) & \textit{if}\, n \ge \frac{C\mathrm{vc}(\mathcal{H})}{2} \end{cases} \tag{4.2}$$

where C is a universal constant, and e is the base of the natural logarithm. If Assumptions 1-2 also hold (in which case $\mathcal{H} = \mathcal{H}_+$), then for every $\eta \in (0,1)$, when $n \ge \frac{C\mathrm{vc}(\mathcal{H})}{2}$ and we set $t := \sqrt{\frac{1}{n}\log\frac{8}{\eta} + \frac{C\mathrm{vc}(\mathcal{H})}{n}\log\frac{2en}{C\mathrm{vc}(\mathcal{H})}} \le \frac{\alpha}{4}$, with probability at least $1-\eta$, we have, for every optimal solution $(\hat{L}_t^*, \hat{U}_t^*)$ of $\widehat{\mathrm{opt}}(t)$, that $\mathbb{P}_\pi(Y \in [\hat{L}_t^*(X), \hat{U}_t^*(X)]) \ge 1 - \alpha$ and

$$\begin{aligned}\mathbb{E}_{\pi_X}[\hat{U}_t^*(X) - \hat{L}_t^*(X)] &\le \mathcal{R}^*(\mathcal{H}) \\ &+ \frac{24}{\alpha\gamma_{\frac{\alpha}{6}}}\sqrt{\frac{1}{n}\log\frac{8}{\eta} + \frac{C\mathrm{vc}(\mathcal{H})}{n}\log\frac{2en}{C\mathrm{vc}(\mathcal{H})}} \\ &+ \sqrt{\frac{\mathrm{vc}(\mathcal{H}_+)}{n}\cdot 16C\|H\|_{\psi_2}^2\log\frac{4}{\eta}}.\end{aligned}$$

Theorem 4.3 reveals that, after ignoring logarithmic factors, the sample size n needed to learn a good PI with guaranteed coverage from $\widehat{\mathrm{opt}}(t)$ is of order $\Omega(\mathrm{vc}(\mathcal{H}))$, if t of order $O\big(\sqrt{\mathrm{vc}(\mathcal{H})/n}\big)$ is adopted. The corresponding optimality gap (in width) is $O\big(\sqrt{\mathrm{vc}(\mathcal{H}_+)/n}\big)$. Appendix B provides further discussion regarding the use of $\mathcal{H}_+$ versus $\mathcal{H}$ in the bound.

Similar sample complexities for VC-major $\mathcal{H}$ have been proposed in Rosenfeld et al. (2018) for a formulation that can be viewed as the dual of (3.1), where the coverage rate is maximized under a mean width constraint. Comparing VC-major and VC-subgraph classes, they both cover many hypothesis classes commonly used in practice, e.g., linear functions, regression trees, and neural networks that are to be discussed momentarily. Nonetheless, in general neither VC-major nor VC-subgraph implies the other, and therefore our VC-subgraph results are in parallel to those in Rosenfeld et al. (2018). More importantly, their results provide finite-sample errors only for the coverage rate, but not the interval width, whereas we characterize performances *jointly* in coverage and width, thanks to our novel sensitivity measure from Theorem 4.1. Moreover, our results appear to bypass a technical issue of Rosenfeld et al. (2018). A key result there is that for a VC-major class $\mathcal{H}$, the induced set of between-graphs $\{\{(x,z) : L(x) \leq z \leq U(x)\} : L, U \in \mathcal{H} \text{ and } L \leq U\}$ is a VC class. When the class $\mathcal{H}$ is uniformly bounded from below, say 0, then this is equivalent to $\mathcal{H}$ being VC-subgraph. However, a counter-example for this conclusion is constructed in Theorem 2.1, statement f, in Dudley (1987). Our approach overcomes such technical ambiguities.

Our next considered hypothesis class is on Lipschitz continuity with respect to the class parameter:

Theorem 4.4 (Joint coverage-width guarantee for the Lipschitz class in parameter)**.** *Suppose $\mathcal{H} = \{h(\cdot, \theta) : \theta \in \Theta\}$ where the parameter space Θ is a bounded set in $\mathbb{R}^l$. If the functions are Lipschitz continuous in the parameter, i.e., $|h(x,\theta_1) - h(x,\theta_2)| \leq \mathcal{L}(x)\|\theta_1 - \theta_2\|_2$, for all $\theta_1, \theta_2 \in \Theta, x \in \mathcal{X}$ for some $\mathcal{L} : \mathcal{X} \to \mathbb{R}$ such that $\|\mathcal{L}\|_2 < \infty$, and $H(x) := \sup_{\theta\in\Theta}|h(x,\theta) - \mathbb{E}_{\pi_X}[h(X,\theta)]|$ satisfies $\|H\|_{\psi_2} < \infty$, then the deviation bounds in Theorem 4.2 satisfy*

$$\phi_1(n,\epsilon,\mathcal{H}) \leq 2\exp\Big(-\frac{\epsilon^2 n}{C\max\{\log\frac{\mathrm{diam}(\Theta)\|\mathcal{L}\|_2}{\|H\|_2}, 1\}\|H\|_{\psi_2}^2 l}\Big),$$

$$\phi_2(n,t,\mathcal{H}) \leq \Big(C_{\mathcal{H}}^{2C\log\log C_{\mathcal{H}}} \cdot \max\{\frac{t^2 n}{2Cl}, 1\}\Big)\Big)^{2l}\exp(-2t^2 n)$$

where $\mathrm{diam}\,(\Theta) := \sup_{\theta_1,\theta_2\in\Theta}\|\theta_1 - \theta_2\|_2$ *is the diameter of Θ, $D_{Y|X} := \sup_{x,y} p(y|x)$ is the supremum of the conditional density, $C_{\mathcal{H}} := 12\mathrm{diam}\,(\Theta)\, D_{Y|X}\|\mathcal{L}\|_1$, and C is a universal constant. If Assumptions 1-2 also hold, then for every $\eta \in (0,1)$, when we set $t := \sqrt{\frac{1}{n}\log\frac{2}{\eta} + \frac{l}{n}\cdot 4C\log(C_{\mathcal{H}})\log\log(C_{\mathcal{H}})} \leq \frac{\alpha}{4}$, with probability at least $1-\eta$, we have, for every optimal solution $(\hat{L}_t^*, \hat{U}_t^*)$ of $\widehat{\mathrm{opt}}(t)$, that $\mathbb{P}_\pi(Y \in [\hat{L}_t^*(X), \hat{U}_t^*(X)]) \geq 1-\alpha$ and*

$$\begin{aligned}&\mathbb{E}_{\pi_X}[\hat{U}_t^*(X) - \hat{L}_t^*(X)] \leq \mathcal{R}^*(\mathcal{H})\\ &+ \frac{24}{\alpha\gamma_{\frac{\alpha}{6}}}\sqrt{\frac{1}{n}\log\frac{2}{\eta} + \frac{l}{n}\cdot 4C\log(C_{\mathcal{H}})\log\log(C_{\mathcal{H}})}\\ &+ \sqrt{\frac{l}{n}\cdot 16C\max\big\{\log\frac{\mathrm{diam}\,(\Theta)\,\|\mathcal{L}\|_2}{\|H\|_2}, 1\big\}\|H\|_{\psi_2}^2\log\frac{4}{\eta}}.\end{aligned} \tag{4.3}$$

The Lipschitz class is beyond the scope of Rosenfeld et al. (2018). Theorem 4.4 states that the required sample size for this class to achieve a certain learning accuracy is of order $\Omega(l\log(\|\mathcal{L}\|_1))$, which depends on the dimension of the parameter space and the Lipschitz coefficient. Correspondingly, the optimality gap on the interval width is $O(\sqrt{l\log(\|\mathcal{L}\|_2)/n})$. Here the logarithmic factor associated with the Lipschitz coefficient is significant because $\|\mathcal{L}\|_2$ can be exponential in the number of layers for deep neural networks (see more details below). Note that our Lipschitzness is in the class parameter θ rather than the input x. The latter has recently been used to regularize neural networks to improve generalization gaps (e.g., Bartlett et al. (2017); Yoshida and Miyato (2017); Gouk et al. (2018)) and robustness against adversarial attacks (e.g., Cisse et al. (2017); Hein and Andriushchenko (2017); Tsuzuku et al. (2018)), but can potentially lead to a loss in expressiveness of the network (Huster et al., 2018; Anil et al., 2019) because of the size restriction on network weights. Our result does not restrict the sensitivity of the network in the inputs, and in turn its expressiveness.

To further distinguish the technical novelty of Theorem 4.3 and Theorem 4.4, note that established results on uniform convergence bounds (UCBs) assuming VC or Lipschitzness on the hypothesis class $\mathcal{H}$ can only handle the objective (width) on $\mathcal{H}$, but not the constraint (coverage) on a different hypothesis class of indicator functions on the joint (x,y) space. Our analysis explicitly controls this constraint complexity to achieve joint optimality-feasibility guarantees. The proof of Theorem 4.3 involves bounding the VC-dimension of the indicator class through intersections of VC set classes, while that of Theorem 4.4 requires directly bounding the bracketing number and evaluating the entropy integral. Moreover, even for the objective, it appears that the explicit UCBs with potentially unbounded VC or Lipschitz classes considered here are new to the best of our knowledge.

We showcase the application of Theorems 4.3 and 4.4 in two important classes: Regression trees and neural networks. Appendix C presents an additional class of linear hypothesis.

Regression Tree. Suppose we build two binary trees to construct L, U respectively. The tree has at most $S+1$ terminal nodes, and each non-terminal node is split according to $x_{i^*} \leq q$ or $x_{i^*} > q$ for some $i^* \in \{1,\ldots,d\}$ and $q \in \mathbb{R}$. In other words, at most S splits are allowed. A regression tree constructed this way is therefore a piece-wise constant function: $h(x) = \sum_{s=1}^{S+1} c_s I_{x \in R_s}$, where each $c_s \geq 0$, $\cup_{s=1}^{S+1} R_s = \mathcal{X}$, and each R_s is a hyper-rectangle in $\mathbb{R}^d$ that takes the form

$$R_s = \left\{ x \in \mathcal{X} : \begin{array}{l} x_{i_{k,1}} \leq q_{i_{k,1}} \text{ for } k = 1,\ldots,S_1 \\ x_{i_{k,2}} > q_{i_{k,2}} \text{ for } k = 1,\ldots,S_2 \\ \text{each } q_{i_{k,1}}, q_{i_{k,2}} \in [-\infty,+\infty] \\ 0 \leq S_1 + S_2 \leq S \end{array} \right\}.$$

Let hypothesis class $\mathcal{H}$ be the collection of all such regression trees. We use Theorem 4.3. Note that the augmented class $\mathcal{H}_+ = \mathcal{H}$. Since the regression tree takes constant values on each rectangle, its subgraph in the space $\mathcal{X} \times \mathbb{R}$ is a union of hyper-rectangles, i.e., $\big\{(x,z) \in \mathcal{X} \times \mathbb{R} : h(x) > z\big\} = \cup_{s=1}^{S+1}\big(R_s \cup (-\infty, c_s)\big)$. Note that each $R_s \cup (-\infty, c_s)$ is an intersection of at most $S+1$ axis-parallel cuts in $\mathbb{R}^{d+1}$, and the set of all axis-parallel cuts is shown to have a VC dimension $O(\log(d))$ (Gey, 2018). $\mathrm{vc}\,(\mathcal{H})$ can therefore be obtained by applying VC bounds for unions and intersections of VC classes of sets (Van Der Vaart and Wellner, 2009):

Theorem 4.5 (Regression tree)**.** *The class $\mathcal{H}$ of regression trees described as above is the same class as its augmentation $\mathcal{H}_+$, and is VC-subgraph with $\mathrm{vc}\,(\mathcal{H}) = \mathrm{vc}\,(\mathcal{H}_+) \leq CS^2(\log(S))^2\log(d)$ for some universal constant C. If the trees are constructed in such a way that $\max_x h(x) - \min_x h(x) \leq M < \infty$, then Theorem 4.3 can be applied with $\|H\|_{\psi_2} \leq C'M$ for another universal constant C'.*

We note that, although upper bounds of VC dimension have been available for classification trees with binary features, and bounds for classification trees with continuous features appeared very recently (Leboeuf et al., 2020), here we consider continuous-valued regression trees with continuous features whose VC dimensions have not been addressed by previous works. Our proof of Theorem 4.5 is based on recently developed VC results for axis-parallel cuts (Gey, 2018).

Neural Network. Consider the class $\mathcal{H}$ of feed-forward neural networks with a fixed architecture and fixed activation functions, indexed by the weights and biases. Suppose the network has $S-1$ hidden layers, one input layer, and two output units. Denote by W the total number of parameters for weights and biases, and by U the total number of computation units (neurons). Let $O_s \in \mathbb{R}^{n_s}, s = 0,\ldots,S$ be the output of the s-th layer. Then $O_s = \phi_s(W_s O_{s-1} + b_s)$ where $W_s \in \mathbb{R}^{n_s \times n_{s-1}}$ is a matrix of weights, $b_s \in \mathbb{R}^{n_s}$ is a vector of biases, and $\phi_s = (\phi_{s,1},\ldots,\phi_{s,n_s})$ is a vector of activation functions. n_s denotes the number of neurons in the s-th layer. In particular $O_0 = x$ is the input vector and $O_S = (L(x), U(x))$ is the final output vector. We aim to characterize the class of one output unit $L(x)$ since the class of $U(x)$ is the same.

We utilize Theorem 4.4. This approach can advantageously handle activation functions beyond sigmoid and piece-wise polynomial (An alternate approach, which we present in Appendix D, uses Theorem 4.3 and Pollard's pseudo-dimension (Pollard, 2012) that applies to sigmoid and piece-wise polynomial). Assume that each activation function $\phi_{s,k}$ is globally M-Lipschitz so that for some constant M_0 the growth condition $|\phi_{s,k}(z)| \leq M_0 + M|z|$ holds for all $z \in \mathbb{R}$, and that each weight or bias parameter is restricted to the bounded interval $[-B,B]$ for some $B > 0$. To apply Theorem 4.4, we show the following Lipschitz property by a backpropagation-like calculation:

Theorem 4.6 (Neural network)**.** *The neural network class $\mathcal{H} = \{h(\cdot,\theta) : \theta \in [-B,B]^W\}$ defined above satisfies the Lipschitzness condition with $\mathcal{L}(x) = C\sqrt{S}(BM\sqrt{W})^S(\|x\|_2 + M_0\sqrt{U} + BM\sqrt{W})$ where C is a universal constant. Therefore Theorem 4.4 can be applied with $l = W$, $\mathrm{diam}\,(\Theta) = 2B\sqrt{W}$, and $\|\mathcal{L}\|_2 \leq C\sqrt{S}(BM\sqrt{W})^S(\|\|X\|_2\|_2 + M_0\sqrt{U} + BM\sqrt{W})$.*

We make two remarks. First, the sample size required in Theorem 4.4 to achieve a certain learning accuracy is of order $l\log(C_{\mathcal{H}})\log\log(C_{\mathcal{H}})$. Applying the Lipschitz constants from Theorem 4.6 to evaluate the $C_{\mathcal{H}}$ reveals a required sample size of order WS, up to logarithmic factors, for neural networks. Second, the size restriction B on the weights and biases enters into the error bounds in a logarithmic manner. Therefore B is allowed to be (exponentially) large and exerts little impact on the training of the network.

5 Data-Driven Coverage Calibration

In this section, we propose a general-purpose PI calibration method to balance coverage and width performances in practice. On a high level, our proposal selects the margin t in (3.2) in a data-driven manner to guarantee a coverage maintenance.

More precisely, we recall that standard practice in validation requires 1) training models multiple times each with different hyperparameters, and then 2) evaluating the trained models on a validation set to select the optimal one. Our proposal aims to select the optimal PI model in 2). In the following, we thus assume multiple "candidate" models are already available.

Algorithm 1 shows our procedure, which simultaneously outputs K PIs, each at a given prediction level

$1-\alpha_k$ $(k=1,...,K)$. It starts from a candidate set of PI models, called $\{\mathrm{PI}_j(x)=[L_j(x),U_j(x)] : j=1,\dots,m\}$. These models can be obtained from setting m different values at a “tradeoff” parameter (e.g., the dual multiplier in a Lagrangian formulation of the empirical constrained optimization; see Appendix G for a neural net example), but can also be a more general collection of PI models. We then use a validation data set $\mathcal{D}_v := \{(X_i',Y_i') : i=1,\dots,n_v\}$, independent of the PI training, to check the feasibility of each candidate PI using the criterion $\hat{\mathrm{CR}}(\mathrm{PI}_j) := (1/n_v)\sum_{i=1}^{n_v} I_{Y_i'\in\mathrm{PI}_j(X_i')} \geq 1-\alpha_k+\epsilon_j$ for some selected margins ϵ_j.

```
Algorithm 1: Normalized PI Calibration
Input: Candidate PIs
  $\{\mathrm{PI}_j=[L_j,U_j] : j=1,\dots,m\}$, target coverage
  rates $\{1-\alpha_k\in(0,1) : k=1,\dots,K\}$, calibration
  data $\mathcal{D}_v=\{(X_i',Y_i') : i=1,\dots,n_v\}$, and
  confidence level $1-\beta\in(0,1)$.
Procedure:
1. For each $\mathrm{PI}_j$, $j=1,\dots,m$, compute its
  empirical coverage rate on $\mathcal{D}_v$,
  $\hat{\mathrm{CR}}(\mathrm{PI}_j) := \frac{1}{n_v}\sum_{i=1}^{n_v} I_{Y_i'\in\mathrm{PI}_j(X_i')}$. Compute the
  sample covariance matrix $\hat\Sigma\in\mathbb{R}^{m\times m}$ with
  $\hat\Sigma_{j_1,j_2}=\frac{1}{n_v}\sum_{i=1}^{n_v}\big(I_{Y_i'\in\mathrm{PI}_{j_1}(X_i')} -$
  $\hat{\mathrm{CR}}(\mathrm{PI}_{j_1})\big)\big(I_{Y_i'\in\mathrm{PI}_{j_2}(X_i')}-\hat{\mathrm{CR}}(\mathrm{PI}_{j_2})\big)$.
2. Let $\hat\sigma_j^2=\hat\Sigma_{j,j}$, and compute $q_{1-\beta}$, the
  $(1-\beta)$-quantile of $\max_{1\leq j\leq m}\{Z_j/\hat\sigma_j : \hat\sigma_j>0\}$
  where $(Z_1,\dots,Z_m)$ is a multivariate Gaussian
  with mean zero and covariance $\hat\Sigma$.
3. For each coverage rate $1-\alpha_k$, $k=1,\dots,K$,
  compute

  $j^*_{1-\alpha_k}=\arg\min_{1\leq j\leq m}\Big\{\frac{1}{n+n_v}\Big(\sum_{i=1}^{n}|\mathrm{PI}_j(X_i)|+\sum_{i=1}^{n_v}|\mathrm{PI}_j(X_i')|\Big)$
  $: \hat{\mathrm{CR}}(\mathrm{PI}_j)\geq 1-\alpha_k+\frac{q_{1-\beta}\hat\sigma_j}{\sqrt{n_v}}\Big\}$

  where $|\mathrm{PI}_j(\cdot)| := U_j(\cdot)-L_j(\cdot)$ is the width, and
  $\{X_i\}_{i=1}^n$ is the training data set.
Output: $\mathrm{PI}_{j^*_{1-\alpha_k}}$ for $k=1,\dots,K$.
```

The key of our procedure is to tune ϵ_j based on a uniform central limit theorem (CLT) that captures the overall errors incurred in the empirical coverage rates. Denoting by $\mathrm{CR}(\mathrm{PI}_j) := \mathbb{P}_\pi(Y\in\mathrm{PI}_j(X))$ the true coverage rate of PI_j, this CLT implies that, setting $q_{1-\beta}=(1-\beta)$-quantile of $\max_j Z_j/\hat\sigma_j$ for some properly chosen Gaussian vector $(Z_j)_{j=1,\dots,m}$, we have $\mathrm{CR}(\mathrm{PI}_j)\geq\hat{\mathrm{CR}}(\mathrm{PI}_j)-q_{1-\beta}\hat\sigma_j/\sqrt{n_v}$ for all $j=1,\dots,m$ uniformly with probability $\approx 1-\beta$. The uniformity over j ensures the solution in Step 3, which optimizes the interval width, indeed attains feasibility (target coverage) with $1-\beta$ confidence. In this “meta-optimization”, we pool the training and validation sets together in the objective to improve the width performance. We have the following finite-sample guarantee:

Theorem 5.1. *Let $1-\underline{\alpha} := \max_{j=1,\dots,m}\mathrm{CR}(\mathrm{PI}_j)$, $1-\alpha_{\min} := 1-\min_{k=1,\dots,K}\alpha_k$, and $\tilde\alpha := \min\{\alpha_{\min},1-\max_{k=1,\dots,K}\alpha_k\}$. For every collection of interval models $\{\mathrm{PI}_j : 1\leq j\leq m\}$, every n_v, and $\beta\in(0,\frac{1}{2})$, the PIs output by Algorithm 1 satisfy*

$$\begin{aligned}&\mathbb{P}_{\mathcal{D}_v}(\mathrm{CR}(\mathrm{PI}_{j^*_{1-\alpha_k}})\geq 1-\alpha_k \text{ for all } k=1,\dots,K)\\ \geq\;& 1-\beta-C_1\Big(\Big(\frac{\log^7(mn_v)}{n_v\tilde\alpha}\Big)^{\frac{1}{6}}+\\ &\exp\big(-C_2 n_v\min\big\{\epsilon,\frac{\epsilon^2}{\underline{\alpha}(1-\underline{\alpha})}\big\}\big)\Big)\end{aligned}\tag{5.1}$$

with $\epsilon=\max\big\{\alpha_{\min}-\underline{\alpha}-C_1\big((\underline{\alpha}(1-\underline{\alpha})/n_v+\log(n_v\alpha_{\min})/n_v^2)\log(m/\beta)\big)^{1/2},0\big\}$, where $\mathbb{P}_{\mathcal{D}_v}$ denotes the probability with respect to the calibration data, and C_1,C_2 are universal constants.

The most important implication of Theorem 5.1 is that the finite-sample deterioration in the confidence level using our procedure depends only *logarithmically* on m and is *independent* of K. These enable both the use of many candidate models and the output of many PIs at different prediction levels. The latter implies that our algorithm can advantageously generate all PIs for arbitrarily many target levels simultaneously with a single validation exercise, thus is computationally cheap and comprises a strength. We provide further interpretation on the error terms of (5.1) in Appendix E. Besides coverage attainment guarantee in Theorem 5.1, our calibration procedure also possesses guaranteed performance regarding the optimality of the width, provided that only the calibration data are used to assess the width in Step 3 of Algorithm 1. More details can be found in Appendix E.

In addition, we provide and compare an alternate calibration scheme, viewed as an “unnormalized” (as opposed to “normalized”) version of Algorithm 1 when handling the standard error $\hat\sigma_j$, in Appendix F.

We discuss Algorithm 1 in relation to a naive, but natural approach that simply selects the model with the shortest average interval width, among candidate models with empirical coverage rates on the validation dataset reaching the target levels (i.e., $t=0$ in (3.2)). This latter approach is also known as the PAV validation scheme in Kivaranovic et al. (2020) (which we call NNVA in Section 6). Our algorithms improve PAV in two aspects. First, our proposal guarantees with high probability that the target level will be achieved by the test coverage thanks to a corrective margin, while PAV does not offer such a guarantee and tends to fall short. Second, our proposal enjoys a higher statistical power than PAV in that unlike PAV whose analysis is based on concentration bounds, Algorithm 1 is analyzed via an asymptotically tight joint CLT. These guarantees build on recent high-dimensional Berry-Esseen bounds (Chernozhukov et al., 2017) (which notably does not require functional complexity measures but only the geometry of "hit sets"). Moreover, we will observe in the experiments in Section 6 that our proposal empirically performs better than PAV.

6 Experiments

Datasets. We evaluate our approaches on both synthetic datasets and real-world benchmark datasets through comparisons to the state-of-the-arts. The real-world datasets ("Boston", "Concrete", "Wine", "Energy" and "Yacht") have been widely used in previous studies (Hernández-Lobato and Adams, 2015; Gal and Ghahramani, 2016; Lakshminarayanan et al., 2017) for regression tasks. The generative distributions for the three synthetic datasets are:

$$\begin{aligned}
(1) &: f(x) = \tfrac{c^T x}{2} + 10\sin(\tfrac{c^T x}{8}) + \tfrac{||x||_2}{10}\epsilon, x \sim N(0, I_{10}),\\
(2) &: f(x) = \tfrac{1}{8}(c^T x)^2 \sin(c^T x) + \tfrac{||x||_2}{10}\epsilon, x \sim N(0, I_7),\\
(3) &: f(x) = \tfrac{1}{2}c^T x \cdot \cos(c^T x)^2 + \tfrac{||x||_2}{10}\epsilon, x \sim N(0, I_9).
\end{aligned}$$

where $\epsilon \sim N(0,1)$, and c is a constant in $[-2,2]^{10}$, $[-2,2]^7$, $[-2,2]^9$ respectively.

Experimental Setup. We conduct experiments under two scenarios: the **single PI case**, where one PI at a single prediction level $1-\alpha$ are constructed, and the **simultaneous PI case**, where K PIs at K different prediction levels $1-\alpha_1, \cdots, 1-\alpha_K$ are constructed. Each trial is repeated for N times to estimate the confidence of coverage attainment. We adopt neural networks as our PI models with the following loss function: $l(x, y, L, U) := (U(x) - L(x))^2 + \lambda(\max\{L(x) - y, 0\} + \max\{y - U(x), 0\})^2$, where $\lambda > 0$ is a penalty for miscoverage and $(L(x), U(x))$ is the output vector containing the lower and upper bounds. By adjusting λ, PI models with different coverage levels can be trained, which are then fed into our calibration algorithms to obtain the final PI. We implement two calibration strategies: the normalized Gaussian PI calibration in Algorithm 1 (NNGN), and the unnormalized version in Algorithm 2 in Appendix F (NNGU). In addition, we also test the calibration scheme (NNVA) that directly compares the empirical coverage rates on the validation dataset to the target levels, without the Gaussian margin. Note that this is the PAV validation scheme in Kivaranovic et al. (2020).

Baselines. We compare our NNGN, NNGU with the following state-of-art approaches: quantile regression forests (QRF) (Meinshausen, 2006), CV+ prediction interval (CV+) (Barber et al., 2019), split conformalized quantile regression (SCQR) (Romano et al., 2019), quantile regression via SVM (SVMQR) (Steinwart and Thomann, 2017), and split conformal learning (SCL) (Lei et al., 2018). Implementation details can be found in Appendix I.

Evaluation Metrics. For the single PI analysis, our models are evaluated on both exceedance probability (EP) and interval width (IW). EP captures the success in achieving the target confidence level, while IW indicates the average interval width. For the simultaneous PI analysis, we use multiple exceedance probability (MEP) and multiple interval width (MIW). MEP measures the proportion of trials where all PIs reach the target prediction levels simultaneously (i.e., family-wise correctness). Formally:

$$\begin{aligned}
EP &: \quad \tfrac{1}{N}\textstyle\sum_{i=1}^N \mathbb{1}\{CR_i \geq PL\}\\
IW &: \quad \tfrac{1}{Nn}\textstyle\sum_{i=1}^N \sum_{j=1}^n (U_{i,j} - L_{i,j})\\
MEP &: \quad \tfrac{1}{N}\textstyle\sum_{i=1}^N \mathbb{1}\{\bigcap_{k=1}^K \{CR_{i,k} \geq PL_k\}\}\\
MIW &: \quad \tfrac{1}{KNn}\textstyle\sum_{k=1}^K \sum_{i=1}^N \sum_{j=1}^n (U_{i,j,k} - L_{i,j,k})
\end{aligned}$$

where n is the size of testing data, CR_i ($CR_{i,k}$) is the estimated coverage rate from the i-th repetition and PL (PL_k) is the target prediction level $1-\alpha$ ($1-\alpha_k$). Throughout our experiments, the confidence level $1-\beta$ is set to 0.9 in the calibration algorithms. For both cases, the best result is achieved by the model with the smallest IW/MIW value among those with $EP/MEP \geq 0.90$. If no model achieves $EP/MEP \geq 0.90$, then the one with highest EP/MEP is the best.

Single PI Analysis. Table 1 reports the values of EP and IW for PI generation at 95% prediction level on 3 synthetic datasets and 5 real-world datasets. It is shown that QRF, CV+ and our NNGU and NNGN are the only four methods that achieve the required confidence level in all synthetic cases, and among the four our NNGN consistently generates PIs of shortest width. Moreover, NNGU attains the target confidence level on all real datasets as well. NNGN seems to fall below the target confidence for some real datasets, but is better than all other baseline methods except CV+ and SCL. Among the methods with high EP, the interval widths of our NNGN and NNGU are the smallest in 6 out of 8 datasets. In contrast, the EP values by NNVA are below the target confidence in all cases. Numerically, the averaged EP of NNGN/NNGU is 1.4/1.7 times higher than the one of NNVA. This shows in particular that NNVA may fail to ensure a correct coverage rate with high confidence on test data, which necessitates our calibration approaches.

Simultaneous PIs Analysis. Table 2 reports the values of MEP and MIW for simultaneous PIs at 19

Table 1: Single PI at the 95% prediction level. The best results are in **bold**.

	Synthetic1		Synthetic2		Synthetic3		Boston		Concrete		Energy		Wine		Yacht	
Methods	EP	IW	EP	IW	EP	IW	EP	IW	EP	IW	EP	IW	EP	IW	EP	IW
QRF	1.00	3.122	1.00	3.667	1.00	3.609	0.46	2.085	0.72	1.995	0.78	0.664	0.00	2.671	0.22	1.110
CV+	1.00	0.302	1.00	2.039	1.00	3.523	0.96	**1.538**	0.88	1.413	0.98	**0.500**	1.00	4.359	0.92	0.339
SCQR	0.90	2.264	0.78	2.863	0.98	2.804	0.66	2.309	0.62	2.289	0.68	0.837	0.30	2.810	0.86	1.681
SVMQR	0.00	0.256	0.00	2.351	0.00	1.855	0.06	1.340	0.10	1.376	0.18	0.411	0.48	2.975	0.30	0.483
SCL	0.98	0.313	0.70	2.211	1.00	3.754	0.88	2.053	0.88	1.630	0.74	0.769	0.80	4.437	0.92	0.906
NNVA	0.00	0.221	0.74	1.630	0.04	1.991	0.58	1.837	0.28	1.607	0.44	0.234	0.00	1.817	0.80	0.154
Ours-NNGN	1.00	**0.296**	1.00	**1.921**	1.00	**2.176**	0.90	2.477	0.86	2.375	0.62	0.398	0.74	2.155	0.92	**0.217**
Ours-NNGU	1.00	**0.296**	1.00	2.557	1.00	3.155	0.96	2.692	0.96	**2.643**	1.00	0.561	0.98	**2.648**	1.00	0.299

Table 2: Simultaneous PIs at 19 target prediction levels. The best results are in **bold**.

	Synthetic1		Synthetic2		Synthetic3		Boston		Concrete		Energy		Wine		Yacht	
Methods	MEP	MIW	MEP	MIW	MEP	MIW	MEP	MIW	MEP	MIW	MEP	MIW	MEP	MIW	MEP	MIW
QRF	1.00	1.911	0.92	1.657	1.00	1.805	0.46	1.099	0.72	1.108	0.78	0.328	0.00	1.460	0.22	0.703
CV+	1.00	0.176	1.00	1.078	1.00	1.935	0.66	0.780	0.60	0.736	1.00	0.245	1.00	2.289	0.86	0.114
SCQR	0.74	1.356	0.74	2.218	0.56	1.808	0.14	1.622	0.16	1.575	0.24	0.760	0.00	2.493	0.20	1.554
SVMQR	0.00	0.115	0.00	1.311	0.00	1.215	0.00	0.582	0.00	0.619	0.00	0.225	0.04	1.575	0.00	0.086
SCL	0.24	0.175	1.00	1.004	0.96	1.940	0.78	1.039	0.66	0.896	0.74	0.373	0.70	2.470	0.80	0.325
NNVA	0.00	0.160	0.28	0.969	0.04	1.606	0.26	1.086	0.16	0.956	0.00	0.151	0.00	1.147	0.04	0.079
Ours-NNGN	1.00	0.170	1.00	**1.050**	1.00	**1.727**	0.76	1.244	0.72	1.092	0.90	**0.177**	0.72	1.267	0.96	**0.120**
Ours-NNGU	1.00	**0.168**	1.00	1.083	1.00	1.734	**0.82**	1.287	**0.78**	1.123	0.60	0.183	0.98	**1.276**	0.72	0.145

target prediction levels: 50%, 52.5%, 55%, 57.5% …, 95%. Our calibration approaches NNGN and NNGU are always the best in terms of achieving the tightest interval under high MEP, or otherwise achieves the highest MEP among all. Among the 6 datasets where the target confidence level 0.9 can be attained, NNGN/NNGU yields the smallest width in 4/2 of them. In the remaining 2 datasets, NNGU achieves the highest MEP. NNGU attains the target MEP or the highest MEP in 6 out of 8 datasets. Compared to the case of single-level target, the MEP performance gaps are more significant in multi-level PI constructions. This is because the coverage rates in baseline algorithms get increasingly overfitted as more simultaneous target levels are compared against. Thanks to the use of the uniform safety margin, our NNGN and NNGU schemes are free of overfitting even in this case. These results demonstrate that our methods can accurately construct multiple PIs at different prediction levels simultaneously. Finally, compared to NNGU, NNGN tends to generate shorter PIs, and we recommend NNGN as the preferred choice.

We provide more experimental results in Appendix I.

7 Conclusion

In this paper, we study the generation of PIs for regression that satisfy an expected coverage rate. This problem can be cast into an empirical constrained optimization framework that minimizes the expected interval width subject to a coverage satisfaction constraint. We develop a general learning theory to characterize the optimality-feasibility tradeoff in this optimization, in particular joint guarantees on both a short expected interval width and an attainment of the target prediction level. We also propose a readily implementable calibration procedure, constructed based on a high-dimensional Berry-Esseen Theorem, to select the best PI model among trained candidates, which offers a practical approach to build simultaneous PIs at multiple target prediction levels with statistical validity. We demonstrate the empirical strengths of our proposed approach by applying it to neural-network-based PI models with our proposed calibration procedure, and comparing them with other baselines across synthetic and real-data examples.

Acknowledgments

We gratefully acknowledge support from the National Science Foundation under grants CAREER CMMI-1834710 and IIS-1849280.

References

A. M. Alaa and M. van der Schaar. Discriminative jackknife: Quantifying uncertainty in deep learning via higher-order influence functions. *arXiv preprint arXiv:2007.13481*, 2020.

C. Anil, J. Lucas, and R. Grosse. Sorting out lipschitz function approximation. In *International Conference on Machine Learning*, pages 291–301, 2019.

B. Ankenman, B. L. Nelson, and J. Staum. Stochastic kriging for simulation metamodeling. *Operations Research*, 58(2):371–382, 2010.

R. F. Barber, E. J. Candes, A. Ramdas, and R. J. Tibshirani. Predictive inference with the jackknife+. *arXiv preprint arXiv:1905.02928*, 2019.

P. L. Bartlett, D. J. Foster, and M. J. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, pages 6240–6249, 2017.

G. E. Box and G. C. Tiao. *Bayesian inference in statistical analysis*, volume 40. John Wiley & Sons, 2011.

V. Chernozhukov, D. Chetverikov, K. Kato, et al. Central limit theorems and bootstrap in high dimensions. *The Annals of Probability*, 45(4):2309–2352, 2017.

M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin, and N. Usunier. Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 854–863. JMLR. org, 2017.

K. Doksum and J.-Y. Koo. On spline estimators and prediction intervals in nonparametric regression. *Computational Statistics & Data Analysis*, 35 (1):67–82, 2000.

R. Dudley. Universal donsker classes and metric entropy. *The Annals of Probability*, pages 1306–1326, 1987.

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.

I. M. Galván, J. M. Valls, A. Cervantes, and R. Aler. Multi-objective evolutionary optimization of prediction intervals for solar energy forecasting with neural networks. *Information Sciences*, 418:363–382, 2017.

S. Gey. Vapnik–chervonenkis dimension of axis-parallel cuts. *Communications in Statistics-Theory and Methods*, 47(9):2291–2296, 2018.

H. Gouk, E. Frank, B. Pfahringer, and M. Cree. Regularisation of neural networks by enforcing lipschitz continuity. *arXiv preprint arXiv:1804.04368*, 2018.

C. Gupta, A. K. Kuchibhotla, and A. K. Ramdas. Nested conformal prediction and quantile out-of-bag ensemble methods. *arXiv preprint arXiv:1910.10562*, 2019.

M. Hein and M. Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *Advances in Neural Information Processing Systems*, pages 2266–2276, 2017.

J. M. Hernández-Lobato and R. Adams. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International Conference on Machine Learning*, pages 1861–1869, 2015.

T. Huster, C.-Y. J. Chiang, and R. Chadha. Limitations of the lipschitz constant as a defense against adversarial examples. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 16–29. Springer, 2018.

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 (3):337–346, 2010.

A. Khosravi, S. Nahavandi, D. Creighton, and A. F. Atiya. Comprehensive review of neural network-based prediction intervals and new advances. *IEEE Transactions on Neural Networks*, 22(9):1341–1356, 2011.

B. Kim, C. Xu, and R. F. Barber. Predictive inference is free with the jackknife+-after-bootstrap. *arXiv preprint arXiv:2002.09025*, 2020.

D. Kivaranovic, K. D. Johnson, and H. Leeb. Adaptive, distribution-free prediction intervals for deep networks. In *International Conference on Artificial Intelligence and Statistics*, pages 4346–4356, 2020.

R. Koenker and K. F. Hallock. Quantile regression. *Journal of Economic Perspectives*, 15(4):143–156, 2001.

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.

J.-S. Leboeuf, F. LeBlanc, and M. Marchand. Decision trees as partitioning machines to characterize their generalization properties. *arXiv preprint arXiv:2010.07374*, 2020.

J. Lei, A. Rinaldo, and L. Wasserman. A conformal prediction approach to explore functional data. *Annals of Mathematics and Artificial Intelligence*, 74 (1-2):29–43, 2015.

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(523):1094–1111, 2018.

N. Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999, 2006.

D. J. Olive. Prediction intervals for regression models. *Computational Statistics and Data Analysis*, 51(6): 3115–3122, 2007.

G. Papadopoulos, P. J. Edwards, and A. F. Murray. Confidence estimation methods for neural networks: A practical comparison. *IEEE Transactions on Neural Networks*, 12(6):1278–1287, 2001.

H. Papadopoulos. Inductive conformal prediction: Theory and application to neural networks. In *Tools in artificial intelligence*. Citeseer, 2008.

T. Pearce, M. Zaki, A. Brintrup, and A. Neely. High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. In *International Conference on Machine Learning, PMLR: Volume 80*, 2018.

D. Pollard. *Convergence of stochastic processes*. Springer Science and Business Media, 2012.

W. Polonik. Minimum volume sets and generalized quantile processes. *Stochastic Processes and Their Applications*, 69(1):1–24, 1997.

Y. Romano, E. Patterson, and E. Candes. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, pages 3543–3553, 2019.

N. Rosenfeld, Y. Mansour, and E. Yom-Tov. Discriminative learning of prediction intervals. In *International Conference on Artificial Intelligence and Statistics*, pages 347–355, 2018.

J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn. Design and analysis of computer experiments. *Statistical Science*, pages 409–423, 1989.

R. L. Schmoyer. Asymptotically valid prediction intervals for linear models. *Technometrics*, 34(4):399–408, 1992.

C. D. Scott and R. D. Nowak. Learning minimum volume sets. *Journal of Machine Learning Research*, 7(Apr):665–704, 2006.

G. A. Seber and A. J. Lee. *Linear regression analysis*, volume 329. John Wiley & Sons, 2012.

L. Steinberger and H. Leeb. Leave-one-out prediction intervals in linear regression models with many variables. *arXiv preprint arXiv:1602.05801*, 2016.

I. Steinwart and P. Thomann. liquidsvm: A fast and versatile svm package, 2017.

I. Steinwart, A. Christmann, et al. Estimating conditional quantiles with the help of the pinball loss. *Bernoulli*, 17(1):211–225, 2011.

R. A. Stine. Bootstrap prediction intervals for regression. *Journal of the American Statistical Association*, 80(392):1026–1031, 1985.

Y. Tsuzuku, I. Sato, and M. Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. In *Advances in Neural Information Processing Systems*, pages 6541–6550, 2018.

A. Van Der Vaart and J. A. Wellner. A note on bounds for vc dimensions. *Institute of Mathematical Statistics Collections*, 5:103, 2009.

A. W. Van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes with Applications to Statistics*. Springer, 1996.

V. Vapnik. *The nature of statistical learning theory*. Springer Science and Business Media, 2013.

V. Vovk. Conditional validity of inductive conformal predictors. In *Asian Conference on Machine Learning*, pages 475–490, 2012.

V. Vovk. Cross-conformal predictors. *Annals of Mathematics and Artificial Intelligence*, 74(1-2):9–28, 2015.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic learning in a random world*. Springer Science & Business Media, 2005.

Y. Yoshida and T. Miyato. Spectral norm regularization for improving the generalizability of deep learning. *arXiv preprint arXiv:1705.10941*, 2017.

H. Zhang, J. Zimmerman, D. Nettleton, and D. J. Nordman. Random forest prediction intervals. *The American Statistician*, pages 1–15, 2019.

L. Zhu, J. Lu, and Y. Chen. HDI-forest: highest density interval regression forest. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 4468–4474. AAAI Press, 2019.