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 24[th] 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 $\mathcal{D} := \{(X_i, Y_i)\}_{i=1,\ldots,n}$ 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:

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

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

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

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 $\tilde{\mathcal{D}} := \{(X_i, Y_i - f(X_i))\}_{i=1,\ldots,n}$.

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

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

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 $\Gamma(x,\beta) := \inf\{p(y_1|x) + p(y_2|x) : \mathbb{P}_\pi(y_1 \le Y \le y_2 | X = x) \le 1-\beta\}$ 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*

$$\phi_1(n,\epsilon,\mathcal{H}) \le 2\exp\Big(-\frac{n\epsilon^2}{C\|H\|_{\psi_2}^2 \mathrm{vc}(\mathcal{H}_+)}\Big),$$
$$\phi_2(n,t,\mathcal{H}) \le \begin{cases} 4^{n+1}\exp(-t^2 n) & \text{if } n < \frac{C\mathrm{vc}(\mathcal{H})}{2} \\ 4\Big(\frac{2en}{C\mathrm{vc}(\mathcal{H})}\Big)^{C\mathrm{vc}(\mathcal{H})}\exp(-t^2 n) & \text{if } n \ge \frac{C\mathrm{vc}(\mathcal{H})}{2} \end{cases} \tag{4.2}$$

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

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

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 $\{\{(x,z) : L(x) \le z \le U(x)\} : L, U \in \mathcal{H} \text{ and } L \le U\}$ 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., $|h(x,\theta_1) - h(x,\theta_2)| \le \mathcal{L}(x)\|\theta_1 - \theta_2\|_2$, 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*

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

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

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 $t := \sqrt{\frac{1}{n}\log\frac{2}{\eta} + \frac{l}{n}\cdot 4C\log(C_{\mathcal{H}})\log\log(C_{\mathcal{H}})} \le \frac{\alpha}{4}$, 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

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

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., $\{(x,z) \in \mathcal{X} \times \mathbb{R} : h(x) > z\} = \cup_{s=1}^{S+1} \big(R_s \cup (-\infty, c_s)\big)$. 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 $\{\text{PI}_j(x)=[L_j(x),U_j(x)]: j=1,\dots,m\}$. 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 $\hat{\text{CR}}(\text{PI}_j) := (1/n_v)\sum_{i=1}^{n_v} I_{Y_i'\in \text{PI}_j(X_i')} \geq 1-\alpha_k+\epsilon_j$ for some selected margins ϵ_j.

Algorithm 1: Normalized PI Calibration

Input: Candidate PIs $\{\text{PI}_j=[L_j,U_j]: j=1,\dots,m\}$, 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 PI_j, $j=1,\dots,m$, compute its empirical coverage rate on $\mathcal{D}_v$, $\hat{\text{CR}}(\text{PI}_j) := \frac{1}{n_v}\sum_{i=1}^{n_v} I_{Y_i'\in\text{PI}_j(X_i')}$. Compute the sample covariance matrix $\hat{\Sigma}\in\mathbb{R}^{m\times m}$ with $\hat{\Sigma}_{j_1,j_2} = \frac{1}{n_v}\sum_{i=1}^{n_v}\big(I_{Y_i'\in\text{PI}_{j_1}(X_i')} - \hat{\text{CR}}(\text{PI}_{j_1})\big)\big(I_{Y_i'\in\text{PI}_{j_2}(X_i')} - \hat{\text{CR}}(\text{PI}_{j_2})\big)$.

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

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

where $|\text{PI}_j(\cdot)| := U_j(\cdot)-L_j(\cdot)$ is the width, and $\{X_i\}_{i=1}^n$ is the training data set.

Output: $\text{PI}_{j^*_{1-\alpha_k}}$ 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 $\text{CR}(\text{PI}_j) := \mathbb{P}_\pi(Y\in\text{PI}_j(X))$ 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 $\text{CR}(\text{PI}_j)\geq\hat{\text{CR}}(\text{PI}_j)-q_{1-\beta}\hat{\sigma}_j/\sqrt{n_v}$ 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* $1-\underline{\alpha} := \max_{j=1,\dots,m}\text{CR}(\text{PI}_j)$, $1-\alpha_{\min} := 1-\min_{k=1,\dots,K}\alpha_k$, *and* $\tilde{\alpha} := \min\{\alpha_{\min}, 1-\max_{k=1,\dots,K}\alpha_k\}$. *For every collection of interval models* $\{\text{PI}_j : 1\leq j\leq m\}$, *every* n_v, *and* $\beta\in(0,\frac{1}{2})$, *the PIs output by Algorithm 1 satisfy*

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

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:

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

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. Stochas-

tic 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.

Learning Prediction Intervals for Regression: Generalization and Calibration: Supplementary Materials

We provide further results and discussions in this supplemental material. Appendix A presents results on the consistency of the obtained PI from the empirical constrained optimization. Appendix B provides additional discussion on Theorem 4.3. Appendix C shows the joint coverage-width guarantee for the linear hypothesis class. Appendix D shows an alternate analysis for neural networks using Pollard's pseudo-dimension and our derived results for the VC-subgraph class. Appendix E further discusses the finite-sample guarantees for our coverage calibration procedure. Appendix F presents and explains an alternate calibration procedure. Appendix G discusses a Lagrangian formulation to train neural networks that construct PIs. Appendix H reviews some background in empirical processes. Appendix I illustrates experimental details and additional experimental results. Finally, Appendix J shows all technical proofs.

A Results on Basic Consistency

This section presents our results regarding asymptotic consistency in using $\widehat{\text{opt}}(t)$ to approximate the PI rendered by (3.1). Assuming the weak uniform law of large numbers for both the empirical interval width and coverage rate, we first show the following general result:

Theorem A.1 (A general consistency result)**.** *Denote by $(\hat{L}_t^*, \hat{U}_t^*)$ an optimal solution of $\widehat{\text{opt}}(t)$. Suppose Assumptions 1-2 hold. If the hypothesis class $\mathcal{H}$ is weak π_X-Glivenko-Cantelli (GC), and the induced set class $\{\{(x,y) \in \mathcal{X} \times \mathbb{R} : L(x) \leq y \leq U(x)\} : L, U \in \mathcal{H}, L \leq U\}$ is weak π-GC in the product space (see Section H for related definitions), then $\widehat{\text{opt}}(t)$ is consistent with respect to (3.1) in the sense that there exists a sequence $t_n \to 0$ such that, with probability tending to one, $\mathbb{P}_\pi(Y \in [\hat{L}_{t_n}^*(X), \hat{U}_{t_n}^*(X)]) \geq 1-\alpha$, and that $\mathbb{E}_{\pi_X}[\hat{U}_{t_n}^*(X) - \hat{L}_{t_n}^*(X)] \to \mathcal{R}^*(\mathcal{H})$ in probability.*

Theorem A.1 states that, if the weak uniform law of large numbers holds for both the hypothesis class and the induced set of "between"-graphs, the interval learned from $\widehat{\text{opt}}(t)$ has the desired coverage rate and a vanishing optimality gap in width by properly selecting the margin t. This general result requires a simultaneous control of the function class $\mathcal{H}$ and its induced set class. Our next result shows that, under a mild boundedness condition on the conditional density function, GC property of the function class $\mathcal{H}$ can be propagated to the induced set class, therefore it suffices to control the class $\mathcal{H}$ only.

Theorem A.2 (Consistency for strong π_X-GC hypothesis)**.** *Assume the conditional density function from Assumption 2 is bounded, i.e., $\sup_{x,y} p(y|x) < \infty$, and that $\mathbb{E}_\pi[|Y|] < \infty$, then strong π_X-GC of the hypothesis class $\mathcal{H}$ implies strong π-GC of the induced set class defined in Theorem A.1. Therefore, if Assumptions 1-2 are further assumed, the conclusion of Theorem A.1 holds for strong π_X-GC $\mathcal{H}$.*

B Further Discussion of Theorem 4.3

We provide further discussion on Theorem 4.3 regarding $\mathcal{H}_+$ versus $\mathcal{H}$ in the bound. Note that, since $\mathcal{H} \subset \mathcal{H}_+$, the augmented class $\mathcal{H}_+$ being VC-subgraph is a stronger condition than $\mathcal{H}$ being VC-subgraph. Nonetheless, we comment that this is a technical assumption used to accommodate potentially unbounded outcomes or PIs (e.g., Assumption 1 implies unboundedness of functions in $\mathcal{H}$). When Y is uniformly bounded, say within $[0,1]$, it suffices to consider bounded L, U only in PI construction. In that case, $\mathcal{H}$ being VC-subgraph already suffices to ensure similar finite-sample bounds.

C Linear Hypothesis Class

We present a joint coverage-width guarantee for PIs constructed from a linear hypothesis class. Consider the linear hypothesis $\mathcal{H} = \{a^T x + b : \|a\|_1 \le B, b \in \mathbb{R}\}$ for some $B > 0$. The l_1-norm of the coefficient is set bounded to control model complexity.

We demonstrate how Theorem 4.3 is applied to this class. First note that the augmented class $\mathcal{H}_+ = \mathcal{H}$ is the same class of the linear function class $\{a^T x + b : \|a\|_1 \le B, b \in \mathbb{R}\}$, which is VC-subgraph of dimension at most $d + 2$ (see, e.g., Theorem 2.6.7 in Van der Vaart and Wellner (1996)), and hence $\mathrm{vc}(\mathcal{H}) = \mathrm{vc}(\mathcal{H}_+) \le d + 2$. Therefore

$$\phi_1(n, \epsilon, \mathcal{H}) \le 2 \exp\Big(- \frac{n\epsilon^2}{C\|H\|_{\psi_2}^2 \mathrm{vc}(\mathcal{H}_+)} \Big) \tag{C.1}$$

and

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

in Theorem 4.3 hold with both $\mathrm{vc}(\mathcal{H})$ and $\mathrm{vc}(\mathcal{H}_+)$ replaced by $d+2$. To derive the $\|H\|_{\psi_2}$ in (C.1), we calculate $H(x) = \sup_{\|a\|_1 \le B} |a^T(x - \mathbb{E}[X])| = B\|x - \mathbb{E}[X]\|_\infty$, leading to $\|H\|_{\psi_2} = B\|\|X - \mathbb{E}[X]\|_\infty\|_{\psi_2}$.

The above analysis is a direct application of general VC theory to the linear function class, and the bound ϕ_1 exhibits a polynomial dependence on the dimension d. A finer analysis that exploits the linear structure can potentially deliver bounds with much lighter dimension dependence, e.g., Zhang (2002) provides specialized covering number bounds for linear function classes with norm-constrained coefficients which ultimately translate into tighter deviation bounds. The theory in Zhang (2002) however requires that the variable X has a bounded support, whereas here we are able to show a logarithmic dependence for unbounded X through a more elementary treatment. Specifically, the maximal deviation can be expressed as $\sup_{h \in \mathcal{H}} |\mathbb{E}_{\hat{\pi}_X}[h(X)] - \mathbb{E}_{\pi_X}[h(X)]| \le B\|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}_{\pi_X}[X]\|_\infty$, and applying the sub-Gaussian concentration inequality to the supremum norm gives rise to the following:

Theorem C.1 (Linear hypothesis class)**.** *For the linear class $\mathcal{H}$ defined as above we have*

$$\phi_1(n, \epsilon, \mathcal{H}) \le 2 \exp\Big(- \frac{\epsilon^2 n}{CB^2 \|\|X - \mathbb{E}_{\pi_X}[X]\|_\infty\|_{\psi_2}^2 \log d} \Big)$$

where C is a universal constant.

D Alternate Analysis of Neural Networks using VC Dimension

In Section 4 we have established the joint coverage-width guarantee for PIs constructed from neural networks using our Lipschitz class results (Theorem 4.4). Here we provide an alternate approach to analyze neural networks via our VC class results (Theorem 4.3). We consider the VC dimension of a real-valued neural network as a VC-subgraph class, which is also known as Pollard's pseudo-dimension Pollard (2012). Bounds for pseudo-dimension are relatively well-established for neural networks with sigmoid or piece-wise polynomial activation. For example, when all activation functions are sigmoid, the class is VC-subgraph with $\mathrm{vc}(\mathcal{H}) = O(W^2U^2)$ (Theorem 14.2 in Anthony and Bartlett (2009)). Alternatively, if all the activation functions are piece-wise polynomials with a bounded number of pieces and of bounded degrees, e.g., rectified linear unit (ReLU, see LeCun et al. (2015); Goodfellow et al. (2016)) or linear activation, then we have $\mathrm{vc}(\mathcal{H}) = O(WU)$ (Theorem 8 in Bartlett et al. (2019)) and simultaneously that $\mathrm{vc}(\mathcal{H}) = O(WS^2 + WS\log(W))$ (Theorem 8.8 in Anthony and Bartlett (2009), Theorem 6 in Bartlett et al. (2019), and Theorem 1 in Bartlett et al. (1999)). Similar bounds are also available (e.g., Theorem 8.14 in Anthony and Bartlett (2009)) when the network involves both sigmoid and piece-wise polynomial activation functions. On the other hand, the augmented class $\mathcal{H}_+$ is a subclass of an augmented neural network where the output unit of $\mathcal{H}$ serves as the last hidden layer (with a single neuron) followed by a new output unit with linear activation, i.e., the class $\{ah + b : h \in \mathcal{H}, a \in \mathbb{R}, b \in \mathbb{R}\}$, and thus its VC-subgraph property can be propagated to this augmented class.

E More Discussions of Finite-Sample Guarantees for the Coverage Calibration Procedure

We provide further interpretations on the margin $q_{1-\beta}\hat{\sigma}_j/\sqrt{n_v}$ in Algorithm 1, and the error terms of (5.1) in Theorem 5.1. The margin $q_{1-\beta}\hat{\sigma}_j/\sqrt{n_v}$ in Algorithm 1 is reasoned from the CLT that $\sqrt{n_v}\big(\hat{\mathrm{CR}}(\mathrm{PI}_1)-\mathrm{CR}(\mathrm{PI}_1),\ldots,\hat{\mathrm{CR}}(\mathrm{PI}_m)-\mathrm{CR}(\mathrm{PI}_m)\big)\xrightarrow{d}N(0,\Sigma)$ where Σ is the covariance matrix with $\Sigma_{j_1,j_2}=\mathrm{Cov}_\pi(I_{Y\in\mathrm{PI}_{j_1}(X)},I_{Y\in\mathrm{PI}_{j_2}(X)})$. Approximating Σ with the sample covariance $\hat{\Sigma}$ from Step 1 of Algorithm 1 and applying the continuous mapping theorem, we have $\sqrt{n_v}\max_j(\hat{\mathrm{CR}}(\mathrm{PI}_j)-\mathrm{CR}(\mathrm{PI}_j))/\hat{\sigma}_j\xrightarrow{d}\max_j Z_j/\hat{\sigma}_j$ where $(Z_j)_{j=1,\ldots,m}$ follows $N(0,\Sigma)$. Therefore, using the $1-\beta$ quantile of $\max_j Z_j/\hat{\sigma}_j$ in the margin leads to a uniform control of the statistical errors in $\hat{\mathrm{CR}}(\mathrm{PI}_j)$'s with probability approximately $1-\beta$. Theorem 5.1 states this approximation concretely.

The polynomial term in (5.1) corresponds to the error of the joint central limit convergence, and the exponential error term quantifies the probability of the undesirable event that none of the candidate PIs satisfies the penalized constraint in Step 3. In practice, one usually targets at relatively high coverage rates, say at least 50%, and would train the candidate PIs in such a way that the true coverage rates of some of the PIs sufficiently exceed the highest target level, e.g., by heavily penalizing the coverage error. In that case, $\alpha_{\min}=\tilde{\alpha}$, and $\underline{\alpha}<\alpha_{\min}$ with a sufficient gap, therefore using a sample size n_v of order $\Omega(\log^7(m)/\alpha_{\min})$ is enough for ensuring $\epsilon>0$ and the probability (5.1) close to $1-\beta$ so that correct coverage rates are guaranteed with high confidence. This logarithmic dependence on m allows us to advantageously use lots of candidate models in the calibration step.

Another notable feature of the finite-sample error is its independence of K, the number of target rates. This independence arises from the choice of the margin based on the Gaussian supremum that leads to a uniform control of the statistical errors in the empirical coverage rates. This provides the flexibility of constructing PIs for arbitrarily many target levels simultaneously.

Besides coverage attainment, our calibration procedure also possesses guaranteed performance regarding the other side of the feasibility-optimality tradeoff, provided that only the calibration data are used to assess the width in Step 3 of Algorithm 1. This is detailed in the following result:

Theorem E.1. *Assume all the candidate PIs in Algorithm 1 are selected from a hypothesis class $\mathcal{H}$ whose envelope $H(x):=\sup_{h\in\mathcal{H}}|h(x)-\mathbb{E}_{\pi_X}[h(X)]|$ has a finite sub-Gaussian norm $\|H\|_{\psi_2}<\infty$. If in Step 3 of Algorithm 1 each $j^*_{1-\alpha_k}$ is selected according to*

$$j^*_{1-\alpha_k}=\operatorname*{arg\,min}_{1\le j\le m}\Big\{\frac{1}{n_v}\sum_{i=1}^{n_v}|\mathrm{PI}_j(X'_i)|:\hat{\mathrm{CR}}(\mathrm{PI}_j)\ge 1-\alpha_k+\frac{q_{1-\beta}\hat{\sigma}_j}{\sqrt{n_v}}\Big\}$$

and all other steps are kept the same, then for every $\epsilon>0$ we have

$$\begin{aligned}&\mathbb{P}_{\mathcal{D}_v}\Big(\mathbb{E}_{\pi_X}[U_{j^*_{1-\alpha_k}}(X)-L_{j^*_{1-\alpha_k}}(X)]\le\min_{j:\mathrm{CR}(\mathrm{PI_j})\ge 1-\alpha_k+\epsilon}\mathbb{E}_{\pi_X}[U_j(X)-L_j(X)]+2C\epsilon\|H\|_{\psi_2}\ \text{for all } k=1,\ldots,K\Big)\\ \ge\ & 1-8m\exp\Big(-\frac{1}{4}\max\big\{\epsilon-C\sqrt{\frac{\log(m/\beta)}{n_v}},0\big\}^2 n_v\Big)\end{aligned}$$

for some universal constant C.

F Alternate Calibration Scheme

We present an alternate coverage calibration scheme than Algorithm 1 that switches the Gaussian vector used in the margin from "normalized" to "unnormalized". To explain, the margin $q_{1-\beta}\hat{\sigma}_j/\sqrt{n_v}$ used in Algorithm 1 is set proportional to the standard deviation of the empirical coverage rate for each individual PI. An alternative is to set $q'_{1-\beta}$, the $1-\beta$ quantile of $\max\{Z_j:1\le j\le m\}$, as a uniform margin for all the PIs, which also captures the uniform error in coverage rates due to the convergence $\sqrt{n_v}\max_j(\hat{\mathrm{CR}}(\mathrm{PI}_j)-\mathrm{CR}(\mathrm{PI}_j))\xrightarrow{d}\max_j Z_j$. This alternative scheme is depicted in Algorithm 2.

Algorithm 2 enjoys a similar finite-sample performance guarantee:

Algorithm 2: Unnormalized PI Calibration

Input : Same as in Algorithm 1.

Procedure:

1. Same as in Algorithm 1.

2. Compute $q'_{1-\beta}$, the $(1-\beta)$-quantile of $\max\{Z_j : 1 \le j \le m\}$ where $(Z_1, \ldots, Z_m)$ is a multivariate Gaussian with mean zero and covariance $\hat{\Sigma}$.

3. For each coverage rate $k = 1, \ldots, K$ compute

$$j^*_{1-\alpha_k} = \operatorname*{arg\,min}_{1\le j\le m} \Big\{ \frac{1}{n+n_v}\Big(\sum_{i=1}^{n} |\mathrm{PI}_j(X_i)| + \sum_{i=1}^{n_v} |\mathrm{PI}_j(X'_i)| \Big) : \hat{\mathrm{CR}}(\mathrm{PI}_j) \ge 1-\alpha_k + \frac{q'_{1-\beta}}{\sqrt{n_v}} \Big\}$$

where $\{X_i\}_{i=1}^n$ is the training data set.

Output: $\mathrm{PI}_{j^*_{1-\alpha_k}}$ for $k = 1, \ldots, K$.

Theorem F.1. *Under the same setting of Theorem 5.1, the finite sample error* (5.1) *continues to hold for the PIs output by Algorithm 2, but with* $\epsilon = \max\{\alpha_{\min} - \underline{\alpha} - C_1\sqrt{\log(m/\beta)/n_v}, 0\}$.

We compare Algorithms 1 and 2 in terms of statistical efficiency. Like for Algorithm 1, if the target coverage rates are above 50% and the maximal achieved coverage rate of the candidate PIs sufficiently exceeds the highest target level, then $\alpha_{\min} = \tilde{\alpha}$, and $\underline{\alpha} < \alpha_{\min}$ with a sufficient gap. The new expression for ϵ in Theorem F.1 now implies a sample size n_v of order $\Omega\big(\frac{\log m}{\alpha^2_{\min}} + \frac{\log^7 m}{\alpha_{\min}}\big)$ for Algorithm 2 to guarantee correct coverage rates with high confidence. Note that the dependence on $\alpha_{\min}$ grows from a linear one in Algorithm 1 to quadratic, suggesting that Algorithm 1 is more powerful in the case of high target coverage levels. However, when the target coverage levels are moderate (around 50%), Algorithm 2 is more efficient instead, due to a smaller margin than the one in Algorithm 1 for PIs with moderate coverage rates. To explain, denote by $\mathrm{PI}_{\bar{j}}$ the PI with the maximal sample standard deviation (coverage closest to 50%), i.e., $\hat{\sigma}_{\bar{j}} = \max_{1\le j\le m} \hat{\sigma}_j$, then the margin used for $\mathrm{PI}_{\bar{j}}$ in Algorithm 1 satisfies

$$\begin{aligned} q_{1-\beta}\hat{\sigma}_{\bar{j}} &= \hat{\sigma}_{\bar{j}} \cdot 1-\beta \text{ quantile of } \max_{1\le j\le m} Z_j/\hat{\sigma}_j \\ &= 1-\beta \text{ quantile of } \max_{1\le j\le m} \hat{\sigma}_{\bar{j}} Z_j/\hat{\sigma}_j \\ &> 1-\beta \text{ quantile of } \max_{1\le j\le m} Z_j \quad \text{if not all } \hat{\sigma}_j\text{'s are equal} \\ &= q'_{1-\beta} \end{aligned}$$

which is the margin used by Algorithm 2.

G A Lagrangian Formulation for Training Neural-Network-Based Prediction Intervals

We discuss a Lagrangian formulation of (3.2) for training neural networks to construct PIs. This formulation has the dual multiplier set as the tunable parameter to balance the tradeoff between the objective and the constraint in (3.2). Specifically, we use

$$L(\delta;\lambda) = \mathbb{E}_{\hat{\pi}_X}[U(X) - L(X)] + \lambda(1-\alpha+t-\mathbb{P}_{\hat{\pi}}(Y \in [L(X), U(X)]))$$

or

$$L(\delta;\lambda) = \frac{1}{n}\sum_{i=1}^{n}(U(x_i) - L(x_i)) + \frac{\lambda}{n}\sum_{i=1}^{n} I_{y_i \notin [L(x_i), U(x_i)]} + \text{constant}$$

where λ is the multiplier. In practice, we use a "soft" version of the Lagrangian function for gradient descent. The "soft" loss we adopt is introduced in Section 6.

We can build multiple PI models by using different parameters $\lambda > 0$. Then, these models are calibrated using Algorithm 1 or 2 so that the coverage constraint in (3.1) is satisfied. Intuitively, if λ is large, $\sum_{i=1}^{m}(U(X_i) - L(X_i))$

contributes less to the overall loss function, and hence the resulting interval tends to be wide but have a high coverage rate. On the contrary, a small λ entails a short interval with a low coverage rate. Hence, a neural network is a reasonable approach to solve (3.2) since a neural network with the above loss can directly address the tradeoff between the interval width and the coverage rate.

H Empirical Process Background

For a class $\mathcal{G}$ of measurable functions from $\mathcal{X}$ to $\mathbb{R}$ such that $\mathbb{E}_{\pi_X}[|g(X)|] < \infty$ for every $g \in \mathcal{G}$, we say it is weak (resp. strong) π_X-Glivenko–Cantelli (GC) if $\sup_{g\in\mathcal{G}} \left|\frac{1}{n}\sum_{i=1}^{n} g(X_i) - \mathbb{E}_{\pi_X}[g(X)]\right| \to 0$ in probability (resp. almost surely) as $n \to \infty$. For a class $\mathcal{S}$ of measurable subsets of $\mathcal{X}$, i.e., $S \subset \mathcal{X}$ for every $S \in \mathcal{S}$, we say it's weak (resp. strong) π_X-GC if the corresponding indicator class $\{I_{\cdot\in S} : S \in \mathcal{S}\}$ is weak (resp. strong) π_X-GC. When no ambiguity arises, we sometimes suppress the underlying distribution π_X.

A collection of k points $\{x_1, \ldots, x_k\} \subset \mathcal{X}$ is said to be shattered by a class $\mathcal{S}$ of subsets of $\mathcal{X}$ if $\operatorname{card}(\{\{x_1, \ldots, x_k\} \cap S : S \subset \mathcal{S}\}) = 2^k$, where $\operatorname{card}(\cdot)$ denotes the cardinality of a set. The VC dimension of the class $\mathcal{S}$ is defined as $\operatorname{vc}(\mathcal{S}) := \max\{k : \exists\ \{x_1, \ldots, x_k\} \subset \mathcal{X} \text{ shattered by } \mathcal{S}\}$. It is called a VC class if $\operatorname{vc}(\mathcal{S}) < \infty$. A class $\mathcal{G}$ of functions from $\mathcal{X}$ to $\mathbb{R}$ is called VC-subgraph with VC dimension d if the set of subgraphs $\mathcal{S}_{\mathcal{G}} := \{\{(x, z) \in \mathcal{X} \times \mathbb{R} : z < g(x)\} : g \in \mathcal{G}\}$ is a VC class on the product space $\mathcal{X} \times \mathbb{R}$ with $\operatorname{vc}(\mathcal{S}_{\mathcal{G}}) = d$. Without ambiguity we use the same notation $\operatorname{vc}(\mathcal{G})$ to denote the VC dimension of a VC-subgraph class $\mathcal{G}$. Note that VC or VC-subgraph classes are combinatorial in nature and distribution-independent, whereas GC classes here are with respect to a specific distribution π_X.

Given a class $\mathcal{G}$ of functions from $\mathcal{X}$ to $\mathbb{R}$, and a probability measure Q on $\mathcal{X}$, the ϵ-covering number $N(\epsilon, \mathcal{G}, L_2(Q))$ is the minimum number of $L_2(Q)$-balls of size ϵ needed to cover the whole class $\mathcal{G}$. A pair of functions $l, u : \mathcal{X} \to \mathbb{R}$ is called a bracket of size ϵ with respect to $L_2(Q)$ if $l \le u$ almost surely and $(\mathbb{E}_Q[(u - l)^2])^{1/2} \le \epsilon$, and every function g such that $l \le g \le u$ is said to be contained in the bracket. The ϵ-bracketing number $N_{[]}(\epsilon, \mathcal{G}, L_2(Q))$ is the minimum number of brackets of size ϵ needed to cover the whole class $\mathcal{G}$.

The above terminologies extend to the product space $\mathcal{X} \times \mathcal{Y}$ with the joint distribution π in a straightforward manner, i.e., by replacing each occurrence of $\mathcal{X}$ and π_X with $\mathcal{X} \times \mathcal{Y}$ and π respectively.

I Experimental Details and Additional Experiments

We illustrate additional experiments and experimental details, which are divided into two subsections. Appendix I.1 presents and visualizes different PI construction approaches on one-dimensional examples. Appendix I.2 illustrates the Pareto curves for the results in Section 6. Appendix I.3 provides details of our experimental implementations.

I.1 Illustration in One-Dimensional Examples

We conduct experiments and visualize PI construction on three univariate examples. Table 3 shows the generative distributions for the three univariate synthetic datasets. Implementation details can be found in Section I.3.

Index	Tested Function	Function Space	Variable Space	Noise Type(ϵ)
1	$f(x) = \sin(x) + x\epsilon$	$\mathbb{R} \to \mathbb{R}$	$x \sim Unif[-3, 3]$	$\epsilon \sim Unif[-2, 2]$
2	$f(x) = \frac{x^2}{2} + \cos x + x\epsilon$	$\mathbb{R} \to \mathbb{R}$	$x \sim Unif[-3, 3]$	$\epsilon \sim Unif[-2, 2]$
3	$f(x) = x^2 + \frac{\sin(x)}{8} + x\epsilon$	$\mathbb{R} \to \mathbb{R}$	$x \sim Unif[-3, 3]$	$\epsilon \sim Unif[-1, 2]$

Table 3: Tested Functions

Figure 1 illustrates the PIs. We test PIs on a testing dataset and evaluate their performances using the metrics of the coverage rate (CR) and the interval width (IW). All baselines are targeted to attain the prediction level 95%. The titles of all plots are named as Synthetic 1d-{index of synthetic dataset}. Each row shows the performances of the same approach but on different datasets, and each column shows the performances of different approaches on the same dataset. The upper and lower bounds of the PIs that **attain** the 95% target level are shown in solid and green lines, otherwise in dashed and red lines. The covered areas are shaded with their corresponding

colors. Data points are the black dots in the plot. The corresponding CR and IW are shown in the label at the upper center of each plot.

We observe that on all the datasets, our approaches NNGN and NNGU outperform other methods in terms of always attaining the 95% target prediction level and having narrow intervals at the same time. QRF, CV+, SVMQR and NNVA do not attain the 95% prediction levels, which is consistent with the observation that no finite-sample coverage guarantees are known for these approaches. SCQR and SCL attain the 95% prediction level but their intervals appear much wider than NNGN and NNGU. Also, since NNGU is designed to be more conservative than NNGN, the intervals calibrated by NNGN are generally shorter than the ones calibrated by NNGU. This observation is consistent with Table 2 in Section 6.

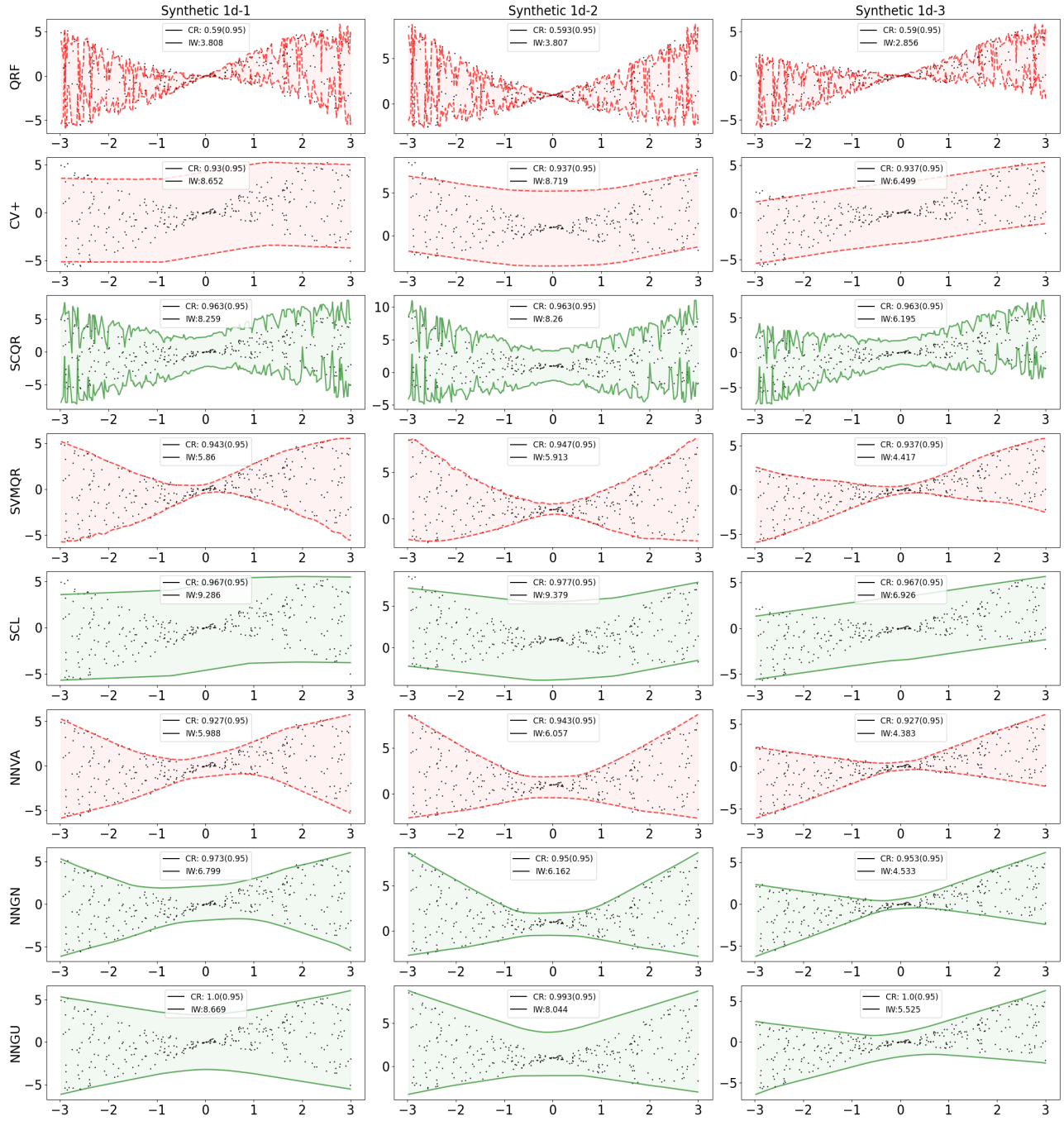

Figure 1: Comparison of single PI constructions. PIs that attain 95% target level are shown in solid and green lines, otherwise in dashed and red lines.

I.2 Pareto Curves

We illustrate the Pareto curves for the simultaneous PIs results with 19 target prediction levels in Section 6, in terms of the coverage rate (CR) (X-axis) and the interval width (IW) (Y-axis), to offer a more intuitive comparison. The titles of all plots are named as the datasets in Section 6. The arrangement of the plots are the same as the ones in Section I.1. Specifically, each subplot contains two curves, one constructed with (input coverage, width) and another constructed with (achieved coverage, width). The curve representing the input coverage and width is shown in dashed line while the one representing the achieved coverage and width is shown in solid line along with a 90% confidence interval as the shaded area. Each point in the line denotes the average

obtained from $N = 50$ repetitions of trials.

Ideally, a method performs well if in its Pareto curves, 1) the dashed line is on the left of the solid line, and 2) at the same time there is no level intersection between the solid line and the shaded area, meaning that within the same input coverage level, there is a sufficiently high probability in achieving the coverage level. Other than that, a smaller average IW at each input coverage level is better since it refers to a less conservative predictive ability. From Figures 2 and 3, we notice that CV+, NNGN, and NNGU have a lot more cases where there is no intersection between the dashed line and shaded area while SCL, QRF, SVMQR, SCQR and NNVA do not. Also the former three methods tend to generate much shorter PIs than the others. Specifically, there are 5 datasets where there is no intersection between the dashed line and the shaded area for CV+ and NNGN, and 4 for NNGU. However, NNGN and NNGU can achieve a much shorter MIW (average of IW over all input coverage levels) than the rest of the methods.

One might notice that the two Pareto curves are a bit farther away from each other for NNVA, NNGN, and NNGU when the input coverage is small. This issue can be solved by choosing the calibration parameter more appropriately by, for example, training a more continuous spectrum of candidate models.

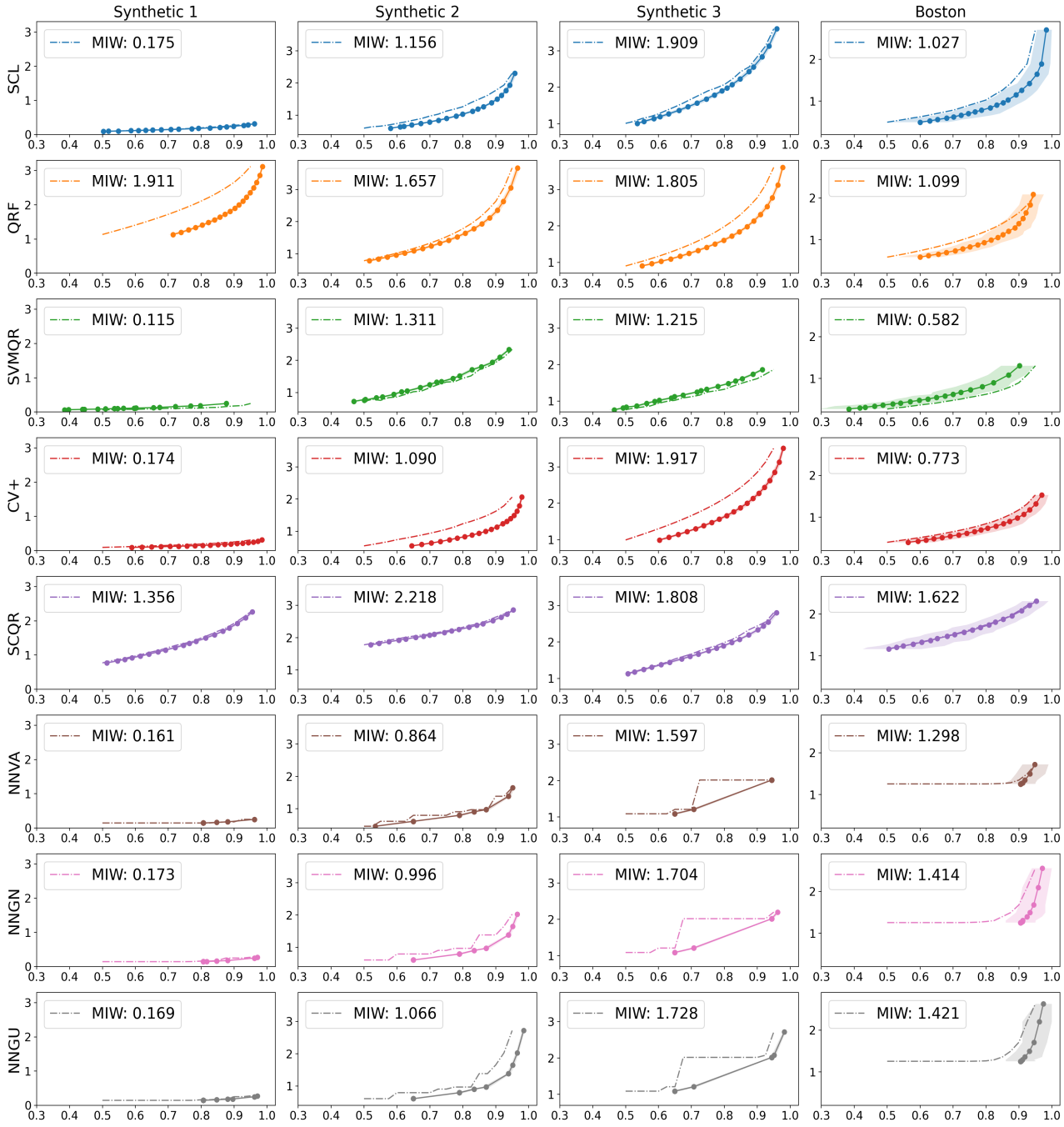

Figure 2: Comparison of simultaneous PIs constructions in synthetic datasets 1 to 3, and Boston dataset. Dashed lines are constructed with (input coverage, width), and solid lines are constructed with (achieved coverage, width). Each point in the line denotes the average obtained from $N = 50$ repetitions of trials. Shaded area is the 90% confidence interval.

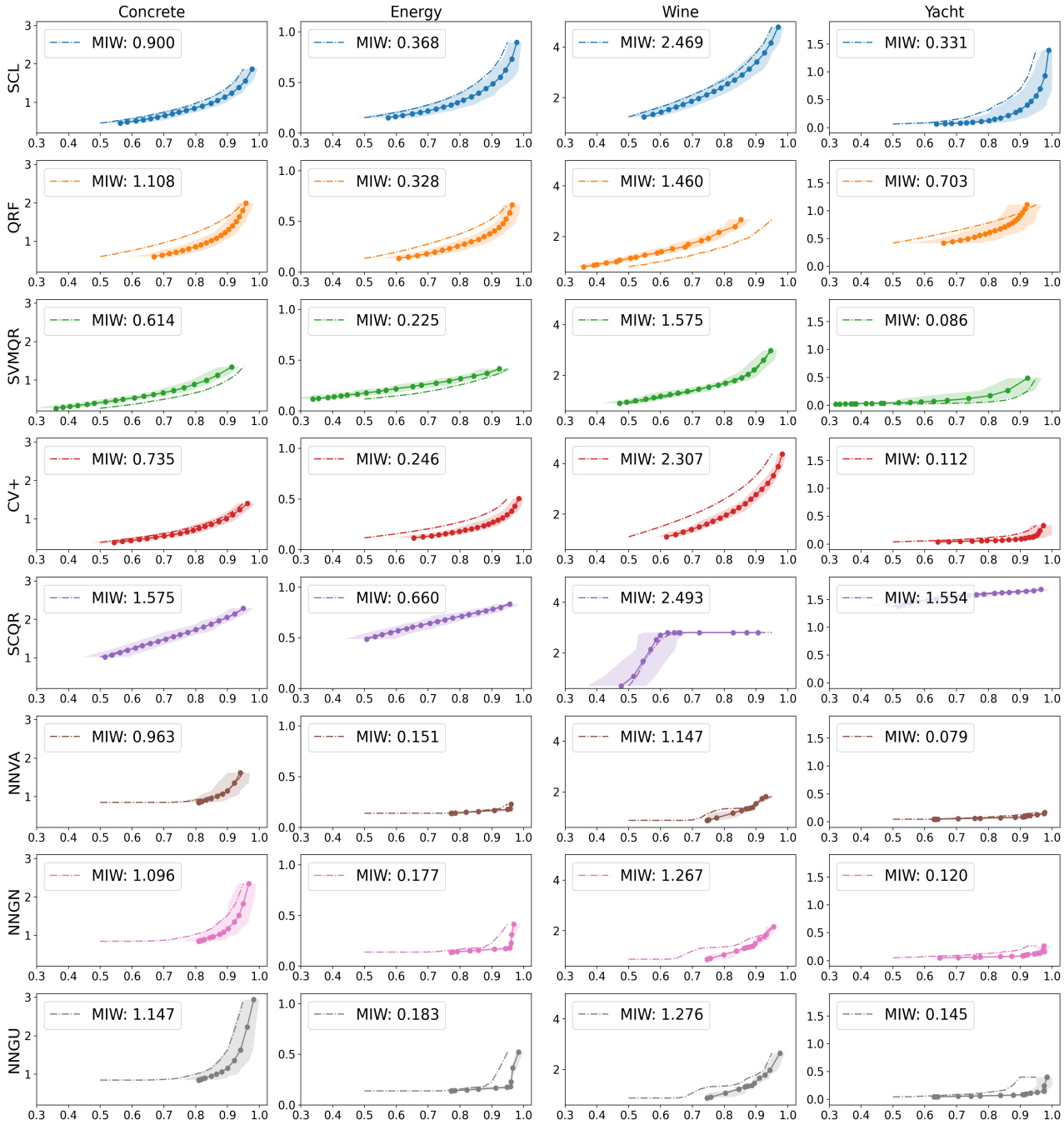

Figure 3: Comparison of simultaneous PIs constructions in Concrete, Energy, Wine and Yacht dataset. Dashed lines are constructed with (input coverage, width), and solid lines are constructed with (achieved coverage, width). Each point in the line denotes the average obtained from $N = 50$ repetitions of trials. Shaded area is the 90% confidence interval.

I.3 Implementation Details

We elaborate more details about our experimental implementations in Sections 6 and I.1.

Datasets. Three synthetic datasets and five real-world benchmark datasets have been shown in Section 6. The real-world datasets are the open-access datasets "Boston", "Concrete", "Energy", "Wine" and "Yacht" that have been widely used in previous studies (Hernández-Lobato and Adams, 2015; Gal and Ghahramani, 2016;

Lakshminarayanan et al., 2017) for regression tasks. Table 4 shows their details.

Dataset	N	d	Open-access Link
Boston: Boston Housing	506	13	kaggle.com/c/boston-housing
Concrete: Concrete Strength	1030	8	kaggle.com/aakashphadtare/concrete-data
Energy: Energy Efficiency	768	8	kaggle.com/elikplim/eergy-efficiency-dataset
Wine: Red Wine Quality	1599	11	kaggle.com/uciml/red-wine-quality-cortez-et-al-2009
Yacht: Yacht Hydrodynamics	308	6	archive.ics.uci.edu/ml/datasets/yacht+hydrodynamics

Table 4: Full names and details of benchmarking regression datasets. N is the number of samples in the dataset and d is the dimension of the feature vector.

The data are split into training and testing sets as follows. For the methods where validation data are needed for calibration (NNVA, NNGN, NNGU), we use a proportion of training data as the validation data. In the single PI case, all of the real-world datasets have 80%/20% training/testing split. For NN methods, "Boston" has 24% of data for validation; "Concrete" has 16% of data for validation; "Energy" has 10% of data for validation; "Wine" has 12% of data for validation; "Yacht" has 15% of data for validation. The three multivariate synthetic datasets in Section 6 have 1600/3000 training/testing split. For NN methods, 350 data points are used for validation. In the simultaneous PIs case, all the real-world datasets have 80%/20% training/testing split. For NN methods, "Boston" has 24% of data for validation; "Concrete" has 16% of data for validation; "Energy" has 24% of data for validation; "Wine" has 24% of data for validation; "Yacht" has 24% of data for validation. The three multivariate synthetic datasets in Section 6 have 1600/3000 training/testing split. For NN methods, 350 data points are used for validation. In addition, for the experiments in Section I.1, the three univariate synthetic datasets have 1200/300 training/testing split. For NN methods, 60 data points are used for validation.

Implementations. We provide details of our algorithms and baseline approaches in Section 6. They appear in the same order as in Tables 1 and 2.

(1) QRF: quantile regression forests, as proposed in Meinshausen (2006). Our code is based on *RandomForestQuantileRegressor* from the package *scikit-garden* in Python.

(2) CV+: CV+ prediction interval, as proposed in Section 3 in Barber et al. (2019). In addition, the base regression algorithm is a neural network using mean square loss.

(3) SCQR: split conformalized quantile regression, as proposed in Algorithm 1 in Romano et al. (2019). The base quantile regression algorithm is exactly the QRF in (1).

(4) SVMQR: quantile regression via support vector machine, the code of which is available in Steinwart and Thomann (2017).

(5) SCL: split conformal learning with correction. The original split conformal learning is described in Algorithm 2 in Lei et al. (2018). The base regression algorithm is a neural network using mean square loss. Moreover, we apply a "correction" method to stipulate the coverage constraint with high confidence, which has been proposed in Equation 7 in Proposition 2b in Vovk (2012) to enhance the split/inductive conformal learning. Specifically, we change the prediction level $1-\alpha$ to $1-\alpha'$ by letting

$$\beta \geq bin_{n_v,\alpha}(\lfloor \alpha'(n_v+1)-1 \rfloor),$$

where $1-\beta$ is the prefixed confidence level (90% throughout our experiments), n_v is the size of the calibration set and $bin_{n_v,\alpha}$ is the cumulative binomial distribution function with n_v trials and probability of success α.

(6) NNVA: neural networks using the loss shown in Section 6 with a vanilla scheme. In the vanilla scheme, we build multiple PI models by choosing different parameters $\lambda > 0$ in the loss, and then select PIs with the smallest interval width among those whose empirical coverage rates on the validation dataset is larger than the target prediction levels, i.e., without the Gaussian margin in Algorithm 1.

(7) Ours-NNGN: neural networks using the loss shown in Section 6 with the normalized Gaussian PI calibration in Algorithm 1. We build multiple PI models by choosing different parameters $\lambda > 0$ in the loss.

(8) Ours-NNGU: neural networks using the loss shown in Section 6 with the unnormalized Gaussian PI calibration in Algorithm 2. We build multiple PI models by choosing different parameters $\lambda > 0$ in the loss.

For (6)(7)(8), in order to improve the training of the neural networks, we use the ensemble technique in Pearce et al. (2018). That is, instead of directly taking the outputs of one network, we train networks several times by using different initializations and then define the final prediction U and L as

$$U = \overline{U} + 1.96\sigma_U, \text{ where } \overline{U} = \frac{1}{e}\sum_{j=1}^{e}\hat{U}_j,\ \sigma_U = \frac{1}{e-1}\sum_{j=1}^{e}(\hat{U}_j - \overline{U})^2$$

$$L = \overline{L} - 1.96\sigma_L, \text{ where } \overline{L} = \frac{1}{e}\sum_{j=1}^{e}\hat{L}_j,\ \sigma_L = \frac{1}{e-1}\sum_{j=1}^{e}(\hat{L}_j - \overline{L})^2$$

where e denotes the ensemble size.

Architectures and Hyper-parameters. For (2)(5)(6)(7)(8), we train fully connected neural networks with ReLU activations, using the Adam method for stochastic optimization. Note that the base neural networks in (2)(5) have only one output unit (point prediction) while our neural networks in (6)(7)(8) have two output units (lower and upper bounds of PIs). Nevertheless, the neural networks in (2)(5)(6)(7)(8) have the same architecture of hidden layers on each dataset. On "Boston" and "Concrete", we have 1 hidden layer and each hidden layer has 50 neurons. On "Energy", "Wine" and "Yacht", we have 2 hidden layers and each hidden layer has 64 neurons. On three multivariate synthetic datasets in Section 6, we have 2 hidden layers and each hidden layer has 50 neurons. On the three univariate synthetic datasets in Section I.1, we have 1 hidden layer and each hidden layer has 50 neurons. In addition, $N = 50$ repetitions of trials are run on each dataset. The confidence level is $1 - \beta = 90\%$ in all experiments.

J Technical Proofs

J.1 Proofs for Results in Section 4

Proof of Theorem 4.1. We first present a lemma:

Lemma J.1. *Let ξ be a $[0,1]$-valued random variable such that $\mathbb{E}[\xi] \le 1-\beta$ for some $\beta \in [0,1]$, then we have $\mathbb{P}(\xi \le 1-\beta') \ge \beta - \beta'$ for every $\beta' \in (0,\beta)$.*

Proof. Define

$$\xi' := \begin{cases} 0 & \text{if } \xi \le 1-\beta' \\ 1-\beta' & \text{otherwise} \end{cases}.$$

Then $\xi \ge \xi'$ almost surely, hence

$$1-\beta \ge \mathbb{E}[\xi] \ge \mathbb{E}[\xi'] = (1-\beta')(1-\mathbb{P}(\xi \le 1-\beta'))$$

which gives

$$\mathbb{P}(\xi \le 1-\beta') \ge \frac{\beta-\beta'}{1-\beta'} \ge \beta - \beta'.$$

□

Now we turn to the main proof. For any $\epsilon > 0$, let $(L_\epsilon, U_\epsilon) \in \mathcal{H} \times \mathcal{H}$ be an ϵ-optimal solution of (3.1), i.e., $\mathbb{P}_\pi(Y \in [L_\epsilon(X), U_\epsilon(X)]) \ge 1-\alpha$ and $\mathbb{E}_{\pi_X}[U_\epsilon(X) - L_\epsilon(X)] \le R^*(\mathcal{H}) + \epsilon$. Consider the enlarged interval $L_\epsilon^c := L_\epsilon - c$, $U_\epsilon^c := U_\epsilon + c$, where $c \ge 0$ is a constant. Let

$$P(c) := \mathbb{P}_\pi(Y \in [L_\epsilon^c(X), U_\epsilon^c(X)])$$

be the coverage rate of the new interval. $P(c)$ satisfies $\lim_{c\to+\infty} P(c) = 1$ because of the continuity of measure, hence the smallest c such that the coverage rate is above $1-\alpha+t$ is finite, i.e.,

$$c^* := \inf\left\{c \ge 0 : P(c) \ge 1-\alpha+t\right\} < \infty.$$

We want to derive an upper bound for c^*. If $c^* = 0$, every non-negative number is a valid upper bound, therefore we focus on the non-trivial case $c^* > 0$. In this case, we must have $P(c^*) = 1 - \alpha + t$ due to continuity of the coverage probability function $P(\cdot)$. To explain the continuity of $P(\cdot)$, by conditioning on X we can rewrite

$$\begin{aligned} P(c) &= \mathbb{E}_{\pi_X}[\mathbb{P}_\pi(Y \in [L_\epsilon^c(X), U_\epsilon^c(X)]|X)] \\ &= \mathbb{E}_{\pi_X}\Big[\int_{L_\epsilon(X)-c}^{U_\epsilon(X)+c} p(y|X)dy\Big] \end{aligned} \tag{J.1}$$

and note that $\int_{L_\epsilon(X)-c}^{U_\epsilon(X)+c} p(y|X)dy$ is continuous in c and bounded by 1 almost surely, therefore the continuity follows from bounded convergence theorem. For every $c \in [0, c^*]$, we have $P(c) \le 1 - \alpha + t$ by the definition of c^*, hence applying Lemma J.1 to the conditioned form (J.1) of $P(c)$ gives

$$\mathbb{P}_{\pi_X}\Big(\mathbb{P}_\pi(Y \in [L_\epsilon^c(X), U_\epsilon^c(X)]|X) \le 1 - \frac{\alpha - t}{3}\Big) \ge \frac{2}{3}(\alpha - t).$$

Now we write

$$\begin{aligned} 1 - \alpha + t &= P(c^*) \\ &= \mathbb{P}_\pi(Y \in [L_\epsilon^{c^*}(X), L_\epsilon(X)) \cup (U_\epsilon(X), U_\epsilon^{c^*}(X)]) + \mathbb{P}_\pi(Y \in [L_\epsilon(X), U_\epsilon(X)]) \\ &\ge \mathbb{E}_{\pi_X}[\mathbb{P}_\pi(Y \in [L_\epsilon^{c^*}(X), L_\epsilon(X)) \cup (U_\epsilon(X), U_\epsilon^{c^*}(X)]|X)] + 1 - \alpha \\ &\quad \text{by conditioning on the } X \text{ and feasibility of } (L_\epsilon, U_\epsilon) \\ &= \mathbb{E}_{\pi_X}\Big[\int_0^{c^*} p(L_\epsilon^c(X)|X) + p(U_\epsilon^c(X)|X)dc\Big] + 1 - \alpha \\ &\ge \mathbb{E}_{\pi_X}\Big[\int_0^{c^*} \Gamma\big(X, 1 - \mathbb{P}_\pi(Y \in [L_\epsilon^c(X), U_\epsilon^c(X)]|X)\big)dc\Big] + 1 - \alpha \\ &\quad \text{by the definition of } \Gamma(\cdot,\cdot) \\ &= \int_0^{c^*} \mathbb{E}_{\pi_X}\big[\Gamma\big(X, 1 - \mathbb{P}_\pi(Y \in [L_\epsilon^c(X), U_\epsilon^c(X)]|X)\big)\big]dc + 1 - \alpha \\ &\ge \int_0^{c^*} \gamma_{\frac{\alpha-t}{3}} \mathbb{P}_{\pi_X}\big(\mathbb{P}_\pi(Y \in [L_\epsilon^c(X), U_\epsilon^c(X)]|X) \le 1 - \frac{\alpha - t}{3} \text{ and } \Gamma(X, \frac{\alpha - t}{3}) \ge \gamma_{\frac{\alpha-t}{3}}\big)dc + 1 - \alpha \\ &\ge \frac{\alpha - t}{3}\gamma_{\frac{\alpha-t}{3}} c^* + 1 - \alpha. \end{aligned}$$

Therefore

$$c^* \le \frac{3t}{(\alpha - t)\gamma_{\frac{\alpha-t}{3}}}.$$

Note that $(L_\epsilon^{c^*}, U_\epsilon^{c^*})$ is feasible for (4.1), therefore by optimality we have

$$\begin{aligned} R_t^*(\mathcal{H}) &\le \mathbb{E}_{\pi_X}[U_\epsilon^{c^*}(X) - L_\epsilon^{c^*}(X)] \\ &\le R^*(\mathcal{H}) + \epsilon + 2c^* \\ &\le R^*(\mathcal{H}) + \epsilon + \frac{6t}{(\alpha - t)\gamma_{\frac{\alpha-t}{3}}}. \end{aligned}$$

Since ϵ is arbitrary, sending ϵ to 0 completes the proof. □

Proof of Theorem 4.2. Let $\hat{\mathcal{H}}_t^2$ and $\mathcal{H}_t^2$ be the feasible set of (3.2) and (4.1) respectively. When the events

$$W_\epsilon := \big\{ \sup_{h \in \mathcal{H}} |\mathbb{E}_{\hat{\pi}_X}[h(X)] - \mathbb{E}_{\pi_X}[h(X)]| \le \epsilon \big\}$$

and

$$C_t := \big\{ \sup_{L,U \in \mathcal{H} \text{ and } L \le U} |\mathbb{P}_{\hat{\pi}_X}(Y \in [L(X), U(X)]) - \mathbb{P}_{\pi_X}(Y \in [L(X), U(X)])| \le t \big\}$$

occur, it holds that $\mathcal{H}_{2t}^2 \subset \hat{\mathcal{H}}_t^2 \subset \mathcal{H}_0^2$, therefore $(\hat{L}_t^*, \hat{U}_t^*) \in \hat{\mathcal{H}}_t^2 \subset \mathcal{H}_0^2$ is feasible for (3.1). We also have

$$\begin{aligned}
& \mathbb{E}_{\pi_X}[\hat{U}_t^*(X) - \hat{L}_t^*(X)] \\
\leq\ & \mathbb{E}_{\hat{\pi}_X}[\hat{U}_t^*(X) - \hat{L}_t^*(X)] + 2\epsilon \ \text{ because of } W_\epsilon \\
\leq\ & \inf_{(L,U)\in\mathcal{H}_{2t}^2 \subset \hat{\mathcal{H}}_t^2} \mathbb{E}_{\hat{\pi}_X}[U(X) - L(X)] + 2\epsilon \ \text{ by optimality of } (\hat{L}_t^*, \hat{U}_t^*) \text{ in } \hat{\mathcal{H}}_t^2 \\
\leq\ & \inf_{(L,U)\in\mathcal{H}_{2t}^2} \mathbb{E}_{\pi_X}[U(X) - L(X)] + 4\epsilon \ \text{ because of } W_\epsilon \\
=\ & \mathcal{R}_{2t}^*(\mathcal{H}) + 4\epsilon \\
\leq\ & \mathcal{R}^*(\mathcal{H}) + \frac{12t}{(\alpha-2t)\gamma_{\frac{\alpha-2t}{3}}} + 4\epsilon \ \text{ by Theorem 4.1.}
\end{aligned}$$

Note that $\mathbb{P}(W_\epsilon \cap C_t) \geq 1 - \mathbb{P}(W_\epsilon^c) - \mathbb{P}(C_t^c) \geq 1 - \phi_1(n,\epsilon,\mathcal{H}) - \phi_2(n,t,\mathcal{H})$, concluding the theorem. □

Proof of Theorem 4.3. We will need the following results:

Lemma J.2 (Adapted from Theorem 2.6.7 in Van der Vaart and Wellner (1996))**.** *Let $\|g\|_{Q,2}$ be the L_2-norm of a function g under a probability measure Q. For a VC-subgraph class $\mathcal{G}$ of functions from $\mathcal{X}$ to $\mathbb{R}$, and every probability measure Q on $\mathcal{X}$, we have for every $\epsilon \in (0,1)$*

$$N(\epsilon\|G\|_{Q,2}, \mathcal{G}, L_2(Q)) \leq C(\mathrm{vc}\,(\mathcal{G}) + 1)(16e)^{\mathrm{vc}(\mathcal{G})+1}\big(\frac{1}{\epsilon}\big)^{2\mathrm{vc}(\mathcal{G})}$$

where $G(x) := \sup_{g\in\mathcal{G}}|g(x)|$ is the envelope function of $\mathcal{G}$ and C is a universal constant.

Lemma J.3 (Adapted from Theorem 2.14.1 in Van der Vaart and Wellner (1996))**.** *Using the notations from Lemma J.2, we define*

$$J(\mathcal{G}) := \sup_Q \int_0^1 \sqrt{1 + \log N(\epsilon\|G\|_{Q,2}, \mathcal{G}, L_2(Q))}d\epsilon$$

where the supremum is taken over all discrete probability measures Q with $\|G\|_{Q,2} < \infty$. Then we have

$$\|\sup_{g\in\mathcal{G}}|\mathbb{E}_{\hat{\pi}_X}[g(X)] - \mathbb{E}_{\pi_X}[g(X)]|\|_1 \leq \frac{1}{\sqrt{n}} \cdot CJ(\mathcal{G})\|G\|_2$$

where the L_1 norm $\|\cdot\|_1$ on the left hand size is with respect to the product measure π_X^n (i.e., the data), and C is a universal constant.

As a side note, a rigorous statement for the results in Lemma J.3 involves a so-called P-measurability condition for the class $\mathcal{G}$, but we choose not to deal with the measurability requirement here. P-measurability holds for common function classes, e.g., if there exists a countable subclass $\mathcal{G}' \subset \mathcal{G}$ such that for every $g \in \mathcal{G}$ there exists a sequence from $\mathcal{G}'$ that converges to g point-wise. We need one more result:

Lemma J.4 (Adapted from Theorem 2.14.5 in Van der Vaart and Wellner (1996))**.** *Using the notations from Lemma J.2, we have*

$$\begin{aligned}
& \|\sup_{g\in\mathcal{G}}|\mathbb{E}_{\hat{\pi}_X}[g(X)] - \mathbb{E}_{\pi_X}[g(X)]|\|_{\psi_2} \\
\leq\ & C\big(\|\sup_{g\in\mathcal{G}}|\mathbb{E}_{\hat{\pi}_X}[g(X)] - \mathbb{E}_{\pi_X}[g(X)]|\|_1 + \frac{1}{\sqrt{n}} \cdot \|G\|_{\psi_2}\big)
\end{aligned}$$

where the sub-Gaussian norm $\|\cdot\|_{\psi_2}$ on the left hand size is with respect to the product measure π_X^n (i.e., the data), and C is a universal constant.

We now turn to the main proof. We first deal with ϕ_1. Consider the centered class $\mathcal{H}_c := \{h - \mathbb{E}_{\pi_X}[h(X)] : h \in \mathcal{H}\}$, whose envelope function is H. Since $\mathcal{H}_c \subset \mathcal{H}_+$, we have $\mathrm{vc}\,(\mathcal{H}_c) \leq \mathrm{vc}\,(\mathcal{H}_+)$. We calculate the complexity measure

$J(\mathcal{H}_c)$ from Lemma J.3

$$\begin{aligned}
J(\mathcal{H}_c) &= \sup_Q \int_0^1 \sqrt{1+\log N(\epsilon\|H\|_{Q,2},\mathcal{H}_c,L_2(Q))}d\epsilon \\
&\le \int_0^1 \left(1+2(\mathrm{vc}\,(\mathcal{H}_c)+1)\log\frac{1}{\epsilon}+16(\mathrm{vc}\,(\mathcal{H}_c)+1)+\log(\mathrm{vc}\,(\mathcal{H}_c)+1)+\log C\right)^{\frac{1}{2}}d\epsilon \\
&\qquad \text{by Lemma J.2} \\
&\le \sqrt{1+16(\mathrm{vc}\,(\mathcal{H}_c)+1)+\log(\mathrm{vc}\,(\mathcal{H}_c)+1)+\log C}+\int_0^1\sqrt{2(\mathrm{vc}\,(\mathcal{H}_c)+1)\log\frac{1}{\epsilon}}d\epsilon \\
&\le C\sqrt{\mathrm{vc}\,(\mathcal{H}_c)} \quad \text{for another universal constant } C \\
&\le C\sqrt{\mathrm{vc}\,(\mathcal{H}_+)}.
\end{aligned}$$

Applying the bound in Lemma J.3 to the class $\mathcal{H}_c$ gives

$$\|\sup_{h\in\mathcal{H}_c}|\mathbb{E}_{\hat{\pi}_X}[h(X)]-\mathbb{E}_{\pi_X}[h(X)]|\|_1 \le \sqrt{\frac{\mathrm{vc}\,(\mathcal{H}_+)}{n}}\cdot C\|H\|_2.$$

Further applying Lemma J.4, and using the fact that $\|\cdot\|_2 \le C\|\cdot\|_{\psi_2}$ for some universal constant C lead to

$$\|\sup_{h\in\mathcal{H}_c}|\mathbb{E}_{\hat{\pi}_X}[h(X)]-\mathbb{E}_{\pi_X}[h(X)]|\|_{\psi_2} \le \sqrt{\frac{\mathrm{vc}\,(\mathcal{H}_+)}{n}}\cdot C\|H\|_{\psi_2}.$$

Finally, note that $\sup_{h\in\mathcal{H}}|\mathbb{E}_{\hat{\pi}_X}[h(X)]-\mathbb{E}_{\pi_X}[h(X)]| = \sup_{h\in\mathcal{H}_c}|\mathbb{E}_{\hat{\pi}_X}[h(X)]-\mathbb{E}_{\pi_X}[h(X)]|$, therefore the same bound holds for $\|\sup_{h\in\mathcal{H}}|\mathbb{E}_{\hat{\pi}_X}[h(X)]-\mathbb{E}_{\pi_X}[h(X)]|\|_{\psi_2}$. The sub-Gaussian tail bound then gives the expression for ϕ_1.

Next we analyze ϕ_2. First note that $\mathcal{H}\subset\mathcal{H}_+$, therefore $\mathrm{vc}\,(\mathcal{H}) \le \mathrm{vc}\,(\mathcal{H}_+)$. By the definition of VC-subgraph, both its closed subgraph class $\mathcal{S}_{upper} := \{\{(x,y): y\le U(x)\}: U\in\mathcal{H}\}$ and open subgraph class $\mathcal{S}'_{lower} := \{\{(x,y): y < L(x)\}: L\in\mathcal{H}\}$ have a VC dimension $\mathrm{vc}\,(\mathcal{S}_{upper}) = \mathrm{vc}\,(\mathcal{S}'_{lower}) = \mathrm{vc}\,(\mathcal{H})$ (using $\le$ or $<$ for defining subgraphs does not affect the resulting VC dimension, see Problem 10 from Section 2.6 in Van der Vaart and Wellner (1996)). To proceed, we need the following preservation result for VC classes:

Lemma J.5 (Adapted from Lemma 9.7 statements (i) and (v) in Kosorok (2007)). *Let $\mathcal{S}$ be a VC class of sets in a space $\mathbb{S}$, then*

1. *$\psi:\mathbb{S}\to\mathbb{S}$ be a one-to-one mapping, then the class $\psi(\mathcal{S}) := \{\{\psi(s): s\in S\}: S\in\mathcal{S}\}$ is also a VC class with $\mathrm{vc}\,(\psi(\mathcal{S})) = \mathrm{vc}\,(\mathcal{S})$*
2. *The complement class $\mathcal{S}^c := \{\mathbb{S}\backslash S: S\in\mathcal{S}\}$ is a VC class with $\mathrm{vc}\,(\mathcal{S}^c) = \mathrm{vc}\,(\mathcal{S})$.*

and a VC dimension bound for unions and intersections of VC classes:

Lemma J.6 (Adapted from Theorem 1.1 in Van Der Vaart and Wellner (2009)). *Suppose $\mathcal{S}_1,\dots,\mathcal{S}_K$ are VC classes of sets in a space $\mathbb{S}$. Define $\sqcup_{k=1}^K\mathcal{S}_k := \{\cup_{k=1}^K S_k : S_k\in\mathcal{S}_k \text{ for } k=1,\dots,K\}$ and $\sqcap_{k=1}^K\mathcal{S}_k := \{\cap_{k=1}^K S_k : S_k\in\mathcal{S}_k \text{ for } k=1,\dots,K\}$. We have*

$$\mathrm{vc}\left(\sqcup_{k=1}^K\mathcal{S}_k\right) \le C\log(K)\sum_{k=1}^K\mathrm{vc}\,(\mathcal{S}_k), \quad \mathrm{vc}\left(\sqcap_{k=1}^K\mathcal{S}_k\right) \le C\log(K)\sum_{k=1}^K\mathrm{vc}\,(\mathcal{S}_k)$$

for some universal constant C.

Further consider $\mathcal{S}_{lower} := \{\{(x,y): y\ge L(x)\}: L\in\mathcal{H}\}$, and $\mathcal{S}_{btw} := \{\{(x,y): L(x)\le y\le U(x)\}: L,U\in\mathcal{H} \text{ and } L\le U\}$. We observe that $\mathcal{S}_{lower} = \mathcal{S}'^c_{lower}$, and that $\mathcal{S}_{btw}\subset\mathcal{S}_{lower}\sqcap\mathcal{S}_{upper}$. Therefore by Lemma J.5 we have $\mathrm{vc}\,(\mathcal{S}_{lower}) = \mathrm{vc}\,(\mathcal{S}'_{lower})$, and applying the bound for intersection from Lemma J.6 to $\mathcal{S}_{btw}$ gives $\mathrm{vc}\,(\mathcal{S}_{btw}) \le \mathrm{vc}\,(\mathcal{S}_{lower}\sqcap\mathcal{S}_{upper}) \le C\mathrm{vc}\,(\mathcal{S}_{upper}) = C\mathrm{vc}\,(\mathcal{H})$ for some universal constant C. With this VC bound

for $\mathcal{S}_{btw}$, we are ready to use standard deviation bounds for VC set classes (see, e.g., equation (3.3) in Vapnik (2013)) to get

$$\begin{aligned} & \sup_{L,U\in\mathcal{H} \text{ and } L\le U} \mathbb{P}(|\mathbb{P}_{\hat{\pi}}(L(X)\le Y\le U(X)) - \mathbb{P}_{\pi}(L(X)\le Y\le U(X))| > t) \\ = & \sup_{S\in\mathcal{S}_{btw}} \mathbb{P}(|\mathbb{P}_{\hat{\pi}}((X,Y)\in S) - \mathbb{P}_{\pi}((X,Y)\in S)| > t) \\ \le & \ 4\text{Growth}(2n)\exp(-t^2 n) \end{aligned}$$

where $\text{Growth}(2n)$ is the growth function, or the shattering number, for the class $\mathcal{S}_{btw}$. By the Sauer–Shelah lemma we have $\text{Growth}(2n) = 2^{2n}$ if $2n < \text{vc}(\mathcal{S}_{btw})$ and $\le \left(\frac{2en}{\text{vc}(\mathcal{S}_{btw})}\right)^{\text{vc}(\mathcal{S}_{btw})}$ if $2n \ge \text{vc}(\mathcal{S}_{btw})$. With the upper bound for $\text{vc}(\mathcal{S}_{btw})$, we can bound

$$\text{Growth}(2n) \le \begin{cases} 2^{2n} & \text{if } 2n < C\text{vc}(\mathcal{H}) \\ \left(\frac{2en}{C\text{vc}(\mathcal{H})}\right)^{C\text{vc}(\mathcal{H})} & \text{if } 2n \ge C\text{vc}(\mathcal{H}) \end{cases}$$

giving rise to our formula for ϕ_2.

The feasibility and optimality bound for $\hat{\delta}_t^*$ can be obtained by solving $\phi_1(n,\epsilon,\mathcal{H}) = \frac{\eta}{2}$ and $\phi_2(n,t,\mathcal{H}) = \frac{\eta}{2}$ for ϵ and t, and then applying Theorem 4.2. □

Proof of Theorem 4.4. We first treat ϕ_1. We need a maximal inequality that is similar to Lemma J.3, but based on bracketing numbers instead:

Lemma J.7 (Adapted from Theorem 2.14.2 in Van der Vaart and Wellner (1996))**.** *Using the notations from Lemma J.2, for a function class $\mathcal{G}$ we define*

$$J_{[]}(\mathcal{G}) := \int_0^1 \sqrt{1 + \log N(\epsilon\|G\|_{\pi_X,2}, \mathcal{G}, L_2(\pi_X))}d\epsilon.$$

Then we have

$$\|\sup_{g\in\mathcal{G}}|\mathbb{E}_{\hat{\pi}_X}[g(X)] - \mathbb{E}_{\pi_X}[g(X)]|\|_1 \le \frac{1}{\sqrt{n}}\cdot CJ_{[]}(\mathcal{G})\|G\|_2$$

where the L_1 norm $\|\cdot\|_1$ on the left hand size is with respect to the product measure π_X^n (i.e., the data), and C is a universal constant.

We consider the centered class $\mathcal{H}_c := \{h - \mathbb{E}_{\pi_X}[h(X)] : h\in\mathcal{H}\}$ as in the proof of Theorem 4.3. Note that, by Jensen's inequality, the Lipschitzness condition stipulates that $|\mathbb{E}_{\pi_X}[h(X,\theta_1)] - \mathbb{E}_{\pi_X}[h(X,\theta_2)]| \le \|\mathcal{L}\|_1\|\theta_1-\theta_2\|_2$, therefore the centered class is also Lipschitz in θ, with a slightly larger coefficient

$$|h(x,\theta_1) - \mathbb{E}_{\pi_X}[h(X,\theta_1)] - (h(x,\theta_2) - \mathbb{E}_{\pi_X}[h(X,\theta_2)])| \le (\mathcal{L}(x) + \|\mathcal{L}\|_1)\|\theta_1-\theta_2\|_2.$$

The envelope function of $\mathcal{H}_c$ is H. We then calculate the bracketing number of $\mathcal{H}_c$. Using the Lipschitzness condition, the bracketing number of $\mathcal{H}_c$ can be bounded by the covering number of the parameter space Θ as below

$$N_{[]}(4\epsilon\|\mathcal{L}\|_2, \mathcal{H}_c, L_2(\pi_X)) \le N(\epsilon,\Theta,\|\cdot\|_2)$$

where $N(\epsilon,\Theta,\|\cdot\|_2)$ is the ϵ-covering number of Θ with respect to the l_2 norm, i.e., the minimum number of l_2-balls of size ϵ needed to cover Θ. Since Θ is bounded, its covering number is upper bounded by that of the l_2-ball of radius $\text{diam}(\Theta)$, which is further bounded by $\left(\frac{3\text{diam}(\Theta)}{\epsilon}\right)^l$ (see Problem 6 from Section 2.1 in Van der Vaart and Wellner (1996)). All these lead to

$$N_{[]}(\epsilon\|H\|_2, \mathcal{H}_c, L_2(\pi_X)) \le N(\frac{\epsilon\|H\|_2}{4\|\mathcal{L}\|_2}, \Theta, \|\cdot\|_2) \le \left(\frac{12\text{diam}(\Theta)\|\mathcal{L}\|_2}{\epsilon\|H\|_2}\right)^l.$$

Also note that when $\epsilon \geq \frac{4\text{diam}(\Theta)\|\mathcal{L}\|_2}{\|H\|_2}$ the bracketing number $N_{[]}(\epsilon\|H\|_2, \mathcal{H}_c, L_2(\pi_X)) = 1$ because $N(\text{diam}\,(\Theta), \Theta, \|\cdot\|_2) = 1$. We can now compute the complexity measure $J_{[]}(\mathcal{H}_c)$ as follows

$$\begin{aligned}
J_{[]}(\mathcal{H}_c) &\leq \int_0^{\min\{1, \frac{4\text{diam}(\Theta)\|\mathcal{L}\|_2}{\|H\|_2}\}} \sqrt{1 + l \log \frac{12\text{diam}\,(\Theta)\,\|\mathcal{L}\|_2}{\epsilon\|H\|_2}} d\epsilon + 1 - \min\{1, \frac{4\text{diam}\,(\Theta)\,\|\mathcal{L}\|_2}{\|H\|_2}\} \\
&\leq 1 + \sqrt{l} \int_0^{\min\{1, \frac{4\text{diam}(\Theta)\|\mathcal{L}\|_2}{\|H\|_2}\}} \sqrt{\log \frac{12\text{diam}\,(\Theta)\,\|\mathcal{L}\|_2}{\epsilon\|H\|_2}} d\epsilon \\
&= 1 + \sqrt{l} \cdot \frac{12\text{diam}\,(\Theta)\,\|\mathcal{L}\|_2}{\|H\|_2} \int_0^{\min\{\frac{1}{3}, \frac{\|H\|_2}{12\text{diam}(\Theta)\|\mathcal{L}\|_2}\}} \sqrt{\log \frac{1}{\epsilon}} d\epsilon \\
&= 1 + C\sqrt{l} \cdot \frac{12\text{diam}\,(\Theta)\,\|\mathcal{L}\|_2}{\|H\|_2} \sqrt{\log \max\{3, \frac{12\text{diam}\,(\Theta)\,\|\mathcal{L}\|_2}{\|H\|_2}\}} \cdot \min\{\frac{1}{3}, \frac{\|H\|_2}{12\text{diam}\,(\Theta)\,\|\mathcal{L}\|_2}\} \\
&\qquad \text{by Lemma J.8 below} \\
&\leq 1 + C\sqrt{l} \cdot \sqrt{\log \max\{3, \frac{12\text{diam}\,(\Theta)\,\|\mathcal{L}\|_2}{\|H\|_2}\}} \cdot \min\{\frac{4\text{diam}\,(\Theta)\,\|\mathcal{L}\|_2}{\|H\|_2}, 1\} \\
&\leq C\sqrt{l} \cdot \sqrt{\max\{\log \frac{\text{diam}\,(\Theta)\,\|\mathcal{L}\|_2}{\|H\|_2}, 1\}}.
\end{aligned}$$

Lemma J.8. *For every $c \in (0, \frac{1}{3}]$, we have*

$$\int_0^c \sqrt{\log \frac{1}{\epsilon}} d\epsilon \leq C \cdot c\sqrt{\log \frac{1}{c}}$$

where C is a universal constant.

Proof of Lemma J.8. By a change of variable $t = \sqrt{\log \frac{1}{\epsilon}}$, we write

$$\begin{aligned}
\int_0^c \sqrt{\log \frac{1}{\epsilon}} d\epsilon &= \int_{\sqrt{\log \frac{1}{c}}}^{\infty} 2t^2 \exp(-t^2) dt \\
&= -t\exp(-t^2)\Big|_{t=\sqrt{\log \frac{1}{c}}}^{t=\infty} + \int_{\sqrt{\log \frac{1}{c}}}^{\infty} \exp(-t^2) dt \\
&\leq c\sqrt{\log \frac{1}{c}} + \int_{\sqrt{\log \frac{1}{c}}}^{\infty} \frac{t}{\sqrt{\log \frac{1}{c}}} \exp(-t^2) dt \\
&= c\sqrt{\log \frac{1}{c}} + \frac{c}{2\sqrt{\log \frac{1}{c}}} \leq \big(1 + \frac{1}{2\log 3}\big) c\sqrt{\log \frac{1}{c}}
\end{aligned}$$

where in the last line we use $c \leq \frac{1}{3}$. □

Lemma J.7 then entails the following maximal inequality for the centered class $\mathcal{H}_c$

$$\| \sup_{h \in \mathcal{H}_c} |\mathbb{E}_{\hat{\pi}_X}[h(X)] - \mathbb{E}_{\pi_X}[h(X)]| \|_1 \leq \sqrt{\frac{l}{n}} \cdot C\sqrt{\max\{\log \frac{\text{diam}\,(\Theta)\,\|\mathcal{L}\|_2}{\|H\|_2}, 1\}} \|H\|_2.$$

Further applying Lemma J.4 gives

$$\| \sup_{h \in \mathcal{H}_c} |\mathbb{E}_{\hat{\pi}_X}[h(X)] - \mathbb{E}_{\pi_X}[h(X)]| \|_{\psi_2} \leq \sqrt{\frac{l}{n}} \cdot C\sqrt{\max\{\log \frac{\text{diam}\,(\Theta)\,\|\mathcal{L}\|_2}{\|H\|_2}, 1\}} \|H\|_{\psi_2}.$$

Finally, note that $\sup_{h \in \mathcal{H}} |\mathbb{E}_{\hat{\pi}_X}[h(X)] - \mathbb{E}_{\pi_X}[h(X)]| = \sup_{h \in \mathcal{H}_c} |\mathbb{E}_{\hat{\pi}_X}[h(X)] - \mathbb{E}_{\pi_X}[h(X)]|$, hence the same sub-Gaussian norm holds for the original class $\mathcal{H}$ too. The tail bound ϕ_1 follows from the sub-Gaussian tail bound.

Secondly, we analyze ϕ_2. We first calculate the bracketing number of the corresponding indicator class

$$\mathcal{H}_{ind} := \{(x,y) \to I_{h(x,\theta_l)\le y\le h(x,\theta_u)} : \theta_l, \theta_u \in \Theta, \text{ and } h(\cdot,\theta_l) \le h(\cdot,\theta_u)\}.$$

For fixed $\theta_l^o, \theta_u^o \in \Theta$ and $\epsilon > 0$, consider a bracket enclosed by

$$\begin{aligned} l^o(x,y) &:= I_{h(x,\theta_l^o)+\mathcal{L}(x)\epsilon\le y\le h(x,\theta_u^o)-\mathcal{L}(x)\epsilon} \\ u^o(x,y) &:= I_{h(x,\theta_l^o)-\mathcal{L}(x)\epsilon\le y\le h(x,\theta_u^o)+\mathcal{L}(x)\epsilon} \end{aligned}$$

where $\mathcal{L}(x)$ is the Lipschitz coefficient. It is clear that $l^o \le u^o$. By Lipschitzness, for all θ_l, θ_u such that $\|\theta_l - \theta_l^o\|_2 \le \epsilon$ and $\|\theta_u - \theta_u^o\|_2 \le \epsilon$ we have $h(x,\theta_l^o) - \mathcal{L}(x)\epsilon \le h(x,\theta_l) \le h(x,\theta_l^o) + \mathcal{L}(x)\epsilon$ (similar for $h(x,\theta_u)$). Therefore $l^o(x,y) \le I_{h(x,\theta_l)\le y\le h(x,\theta_u)} \le u^o(x,y)$, i.e., $I_{h(x,\theta_l)\le y\le h(x,\theta_u)}$ belongs to the bracket $[l^o, u^o]$ whenever $\|\theta_l - \theta_l^o\|_2 \le \epsilon$ and $\|\theta_u - \theta_u^o\|_2 \le \epsilon$. To calculate the size of the bracket, we write

$$\begin{aligned} & \mathbb{E}_\pi[u^o(X,Y) - l^o(X,Y)] \\ = \;& \mathbb{P}_\pi(h(X,\theta_l^o) - \mathcal{L}(X)\epsilon \le Y < h(X,\theta_l^o) + \mathcal{L}(X)\epsilon) + \mathbb{P}_\pi(h(X,\theta_u^o) - \mathcal{L}(X)\epsilon < Y \le h(X,\theta_u^o) + \mathcal{L}(X)\epsilon) \\ = \;& \mathbb{E}_{\pi_X}[\mathbb{P}_\pi(h(X,\theta_l^o) - \mathcal{L}(X)\epsilon \le Y < h(X,\theta_l^o) + \mathcal{L}(X)\epsilon|X)] + \\ & \quad \mathbb{E}_{\pi_X}[\mathbb{P}_\pi(h(X,\theta_u^o) - \mathcal{L}(X)\epsilon < Y \le h(X,\theta_u^o) + \mathcal{L}(X)\epsilon|X)] \\ \le \;& 2\mathbb{E}_{\pi_X}[2D_{Y|X}\mathcal{L}(X)\epsilon] = 4D_{Y|X}\|\mathcal{L}\|_1\epsilon. \end{aligned}$$

Note that the bracket size is independent of θ_l and θ_u, therefore the bracketing number $N_{[]}(4D_{Y|X}\|\mathcal{L}\|_1\epsilon, \mathcal{H}_{ind}, L_1(\pi)) \le \big(N(\epsilon,\Theta,\|\cdot\|_2)\big)^2 \le \big(\frac{3\text{diam}(\Theta)}{\epsilon}\big)^{2l}$, i.e.,

$$N_{[]}(\epsilon, \mathcal{H}_{ind}, L_1(\pi)) \le \big(\frac{12\text{diam}\,(\Theta)\, D_{Y|X}\|\mathcal{L}\|_1}{\epsilon}\big)^{2l} = \big(\frac{C_{\mathcal{H}}}{\epsilon}\big)^{2l}. \tag{J.2}$$

To derive the deviation bound using the bracketing number (J.2), we need one more result:

Lemma J.9 (Adapted from Theorem 6.8 in Talagrand (1994)). *Suppose $\mathcal{G}$ is a class of measurable indicator functions from $\mathcal{X}\times\mathbb{R}\to\mathbb{R}$, and that its ϵ-bracketing number $N_{[]}(\epsilon,\mathcal{G},L_1(\pi)) \le \big(\frac{V}{\epsilon}\big)^\nu$ for $V,\nu>0$, then there exists a universal constant C such that for all $t \ge C\sqrt{\frac{\nu\log(V)\log\log(V)}{n}}$ we have*

$$\mathbb{P}(\sup_{g\in\mathcal{G}}|\mathbb{E}_{\hat\pi}[g(X,Y)] - \mathbb{E}_\pi[g(X,Y)]| > t) \le \frac{C}{t\sqrt{n}}\big(\frac{CVt^2n}{\nu}\big)^\nu \exp(-2t^2n).$$

Applying Lemma J.9 to the indicator class $\mathcal{H}_{ind}$, we obtain

$$\begin{aligned} & \mathbb{P}\Big(\sup_{L,U\in\mathcal{H}\text{ and }L\le U}|\mathbb{P}_{\hat\pi}(Y\in[L(X),U(X)]) - \mathbb{P}_\pi(Y\in[L(X),U(X)])| > t\Big) \\ \le \;& \frac{C}{t\sqrt{n}}\big(\frac{CC_{\mathcal{H}}t^2n}{2l}\big)^{2l}\exp(-2t^2n) \quad \text{if } t \ge C\sqrt{\frac{2l\log(C_{\mathcal{H}})\log\log(C_{\mathcal{H}})}{n}} \\ \le \;& \big(\frac{CC_{\mathcal{H}}t^2n}{2l}\big)^{2l}\exp(-2t^2n) \quad \text{assuming } 2l\log(C_{\mathcal{H}})\log\log(C_{\mathcal{H}}) \ge 1. \end{aligned} \tag{J.3}$$

To make the bound (J.3) valid for small t, we next enlarge the constant C so that the bound (J.3) is trivial. That is, we seek for a $\kappa \ge 1$ such that the bound from (J.3) satisfies

$$\big(\frac{\kappa CC_{\mathcal{H}}t^2n}{2l}\big)^{2l}\exp(-2t^2n) \ge 1 \quad \text{when } t = C\sqrt{\frac{2l\log(C_{\mathcal{H}})\log\log(C_{\mathcal{H}})}{n}}$$

which reduces to

$$\big(\kappa C^3 C_{\mathcal{H}}\log(C_{\mathcal{H}})\log\log(C_{\mathcal{H}})\big)^{2l}\exp(-4C^2 l\log(C_{\mathcal{H}})\log\log(C_{\mathcal{H}})) \ge 1.$$

Solving the above inequality for κ gives

$$\kappa \ge \frac{C_{\mathcal{H}}^{2C^2\log\log(C_{\mathcal{H}})-1}}{C^3\log(C_{\mathcal{H}})\log\log(C_{\mathcal{H}})}.$$

Using this κ in (J.3) leads to a trivial bound for $t = C\sqrt{\frac{2l\log(C_{\mathcal{H}})\log\log(C_{\mathcal{H}})}{n}}$, and for smaller t we simply use this trivial bound, therefore we obtain the following unified tail bound

$$\begin{aligned}
&\mathbb{P}(\sup_{L,U\in\mathcal{H} \text{ and } L\le U} |\mathbb{P}_{\hat{\pi}}(Y\in[L(X),U(X)]) - \mathbb{P}_{\pi}(Y\in[L(X),U(X)])| > t) \\
\le\ & \Big(\frac{C\cdot C_{\mathcal{H}}^{2C^2\log\log(C_{\mathcal{H}})}}{C^3 2l\log(C_{\mathcal{H}})\log\log(C_{\mathcal{H}})}\max\{t^2n, 2C^2l\log(C_{\mathcal{H}})\log\log(C_{\mathcal{H}})\}\Big)^{2l}\exp(-2t^2n) \\
\le\ & \Big(C_{\mathcal{H}}^{2C^2\log\log(C_{\mathcal{H}})}\max\{\frac{t^2n}{2C^2l\log(C_{\mathcal{H}})\log\log(C_{\mathcal{H}})}, 1\}\Big)^{2l}\exp(-2t^2n) \\
\le\ & \Big(C_{\mathcal{H}}^{2C^2\log\log(C_{\mathcal{H}})}\max\{\frac{t^2n}{2C^2l}, 1\}\Big)^{2l}\exp(-2t^2n) \quad \text{assuming } \log(C_{\mathcal{H}})\log\log(C_{\mathcal{H}})\ge 1.
\end{aligned}$$

Replacing C^2 with C in the above bound gives the expression for ϕ_2.

Finally, like in Theorem 4.3, we solve $\phi_1(n,\epsilon,\mathcal{H}) = \frac{\eta}{2}$ and $\phi_2(n,t,\mathcal{H}) = \frac{\eta}{2}$ for ϵ and t, and then apply Theorem 4.2 to get the feasibility and optimality errors. We briefly explain how t is derived. Consider the case $\frac{t^2n}{2Cl}\ge 1$, so we can derive the following upper bound for ϕ_2

$$\begin{aligned}
\phi_2(n,t,\mathcal{H}) &\le \Big(C_{\mathcal{H}}^{2C\log\log(C_{\mathcal{H}})}\frac{t^2n}{2Cl}\Big)^{2l}\exp(-2t^2n) \\
&= \Big(\frac{C_{\mathcal{H}}^{2C\log\log(C_{\mathcal{H}})}}{C}\Big)^{2l}\cdot\big(\frac{t^2n}{2l}\exp(-2\frac{t^2n}{2l})\big)^{2l} \\
&< \Big(\frac{C_{\mathcal{H}}^{2C\log\log(C_{\mathcal{H}})}}{C}\Big)^{2l}\cdot\big(\exp(-\frac{t^2n}{2l})\big)^{2l} \quad \text{using } \frac{t^2n}{2l} < \exp(\frac{t^2n}{2l}) \\
&\le \Big(C_{\mathcal{H}}^{2C\log\log(C_{\mathcal{H}})}\Big)^{2l}\cdot\exp(-t^2n) \quad \text{assuming } C\ge 1.
\end{aligned}$$

The choice of t presented in the theorem is obtained by equating this upper bound to $\frac{\eta}{2}$. □

Proof of Theorem 4.5. The regression tree class $\mathcal{H}$ consists of all functions that of form $h(x) = \sum_{s=1}^{S+1} c_s I_{x\in R_s}$, and note that the augmented class $\mathcal{H}_+$ consists of functions $\sum_{s=1}^{S+1}(c_s - c)I_{x\in R_s}$ which are of the same form, therefore $\mathcal{H} = \mathcal{H}_+$. As discussed before, the subgraph of a regression tree takes the form of the union of at most $S+1$ hyper-rectangles in $\mathbb{R}^{d+1}$, where each rectangle is formed by at most S axis-parallel cuts.

We first calculate the VC dimension of the set of all axis-parallel cuts. Four sets of axis-parallel cuts in $\mathbb{R}^{d+1}$ are defined as

$$\begin{aligned}
\mathcal{C}_{d+1}^{\le} &:= \{\{(x_1,\dots,x_{d+1})\in\mathbb{R}^{d+1} : x_j \le a\} : j\in\{1,2,\dots,d+1\}, a\in[-\infty,+\infty]\} \\
\mathcal{C}_{d+1}^{<} &:= \{\{(x_1,\dots,x_{d+1})\in\mathbb{R}^{d+1} : x_j < a\} : j\in\{1,2,\dots,d+1\}, a\in[-\infty,+\infty]\} \\
\mathcal{C}_{d+1}^{\ge} &:= \{\{(x_1,\dots,x_{d+1})\in\mathbb{R}^{d+1} : x_j \ge a\} : j\in\{1,2,\dots,d+1\}, a\in[-\infty,+\infty]\} \\
\mathcal{C}_{d+1}^{>} &:= \{\{(x_1,\dots,x_{d+1})\in\mathbb{R}^{d+1} : x_j > a\} : j\in\{1,2,\dots,d+1\}, a\in[-\infty,+\infty]\}.
\end{aligned}$$

Proposition 1 in Gey (2018) states that $\text{vc}\left(\mathcal{C}_{d+1}^{\le}\right) \le C\log d$ for some universal constant C. Note that $\mathcal{C}_{d+1}^{>}$ is the complement class of $\mathcal{C}_{d+1}^{\le}$, therefore by Lemma J.5 statement 2 we have the same bound for $\text{vc}\left(\mathcal{C}_{d+1}^{>}\right)$. The class $\mathcal{C}_{d+1}^{\ge}$ (resp. $\mathcal{C}_{d+1}^{<}$) can be mapped from $\mathcal{C}_{d+1}^{\le}$ (resp. $\mathcal{C}_{d+1}^{>}$) via the mapping $(x_1,\dots,x_{d+1})\to(-x_1,\dots,-x_{d+1})$, therefore by statement 1 in Lemma J.5 we have the same VC bound for them too. Using Lemma J.6 we know that the VC bound for the set of all axis-parallel cuts $\mathcal{C}_{d+1} := \mathcal{C}_{d+1}^{\le}\sqcup\mathcal{C}_{d+1}^{<}\sqcup\mathcal{C}_{d+1}^{\ge}\sqcup\mathcal{C}_{d+1}^{>}$ has VC dimension $\text{vc}(\mathcal{C}_{d+1}) \le C\log d$.

Now we can readily obtain $\text{vc}(\mathcal{H})$ via intersections and unions. Each hyper-rectangle $R_s\times(-\infty,c_s)$ is the intersection of at most $S+1$ axis-parallel cuts in $\mathbb{R}^{d+1}$. Formally, denoting by $\mathcal{R}$ as set of all such hyper-rectangles, then $\mathcal{R}\subset\sqcap_{s=1}^{S+1}\mathcal{C}_{d+1}$, and hence $\text{vc}(\mathcal{R}) \le CS\log(d)\log(S)$ by Lemma J.6. Moreover, the set of subgraphs of regression trees is a subset of $\sqcup_{s=1}^{S+1}\mathcal{R}$, therefore the VC dimension of the set of subgraphs is at most $CS^2\log(d)\log^2(S)$ again by Lemma J.6.

Finally, when $\max_x h(x) - \min_x h(x) \le M$, then the envelope $H \le 2M$ in Theorem 4.3, therefore the sub-Gaussian norm $\|H\|_{\psi_2} \le C'M$ for some universal constant C'. □

Proof of Theorem 4.6. To make the parameterizaton more instructive, we explicitly write $\theta = (W_1, b_1, \dots, W_S, b_S)$ in terms of the weights and biases. We consider a perturbation $\Delta\theta := (\Delta W_1, \Delta b_1, \dots, \Delta W_S, \Delta b_S)$, and wants to bound the difference $|h(x, \theta+\Delta\theta) - h(x, \theta)|$. Denoting $W'_s = W_s + \Delta W_s$ and $b'_s = b_s + \Delta b_s$, we define two mappings

$$\psi_{s:S} : x \in \mathbb{R}^{n_s} \to \phi_S(W_S\phi_{S-1}(\cdots \phi_{s+1}(W_{s+1}\phi_s(x) + b_{s+1})\cdots) + b_S) \in \mathbb{R}$$
$$\psi'_{0:s-1} : x \in \mathcal{X} \to \phi_{s-1}(W'_{s-1}\phi_{s-2}(\cdots \phi_1(W'_1 x + b'_1)\cdots) + b'_{s-1}) \in \mathbb{R}^{n_{s-1}}.$$

Then the difference can be expressed as

$$\begin{aligned}
& |h(x; W'_1, b'_1, \dots, W'_S, b'_S) - h(x; W_1, b_1, \dots, W_S, b_S)| \\
\le & \sum_{s=1}^{S} |h(x; W'_1, b'_1, \dots, W'_s, b'_s, W_{s+1}, b_{s+1}, \dots, W_S, b_S) - \\
& \qquad h(x; W'_1, b'_1, \dots, W'_{s-1}, b'_{s-1}, W_s, b_s, \dots, W_S, b_S)| \\
\le & \sum_{s=1}^{S} |\psi_{s:S}(W'_s\psi'_{1:s-1}(x) + b'_s) - \psi_{s:S}(W_s\psi'_{1:s-1}(x) + b_s)| \\
\le & \sum_{s=1}^{S} L_{\psi_{s:S}} \|\Delta W_s\psi'_{0:s-1}(x) + \Delta b_s\|_2 \quad \text{where } L_{\psi_{s:S}} \text{ is the Lipschitz constant of } \psi_{s:S} \\
\le & \sum_{s=1}^{S} L_{\psi_{s:S}} (\|\psi'_{0:s-1}(x)\|_2 + 1) \|\left[\Delta W_s, \Delta b_s\right]\|_2 \\
\le & \sum_{s=1}^{S} L_{\psi_{s:S}} (\|\psi'_{0:s-1}(x)\|_2 + 1) \|\left[\Delta W_s, \Delta b_s\right]\|_F
\end{aligned} \tag{J.4}$$

where in the last two lines $\|\cdot\|_2$ and $\|\cdot\|_F$ respectively denote the spectral and Frobenius norms of a matrix. Therefore the problem boils down to bounding the Lipschitz constant $L_{\psi_{s:S}}$ and the norm of the intermediate output $\|\psi'_{0:s-1}(x)\|_2$.

We first calculate $L_{\psi_{s:S}}$. This is relatively straightforward, since $\psi_{s:S}$ is a composition of linear mappings and activation functions. By the chain rule we have

$$L_{\psi_{s:S}} \le M \cdot \prod_{k=s+1}^{S} M\|W_k\|_2 \le M^{S-s+1} \prod_{k=s+1}^{S} \|W_k\|_F \le M^{S-s+1}(B\sqrt{W})^{S-s}$$

where the last inequality uses the fact that $\|W_k\|_F \le B\sqrt{W}$ because each entry in W_k is bounded within $[-B, B]$ and there are W parameters in total.

Then we bound $\|\psi'_{0:s-1}(x)\|_2$. Note that $\psi'_{0:s-1}(x) = \phi_{s-1}(W'_{s-1}\psi'_{0:s-2}(x) + b'_{s-1})$, and $\psi'_{0:0}(x) = x$, therefore we have the following recursion

$$\begin{aligned}
\|\psi'_{0:s-1}(x)\|_2 &= \|\phi_{s-1}(W'_{s-1}\psi'_{0:s-2}(x) + b'_{s-1})\|_2 \\
&\le M_0\sqrt{U} + M\|\left[W'_{s-1}, b'_{s-1}\right]\|_2(\|\psi'_{0:s-2}(x)\|_2 + 1) \\
&\le M_0\sqrt{U} + M\|\left[W'_{s-1}, b'_{s-1}\right]\|_F(\|\psi'_{0:s-2}(x)\|_2 + 1) \\
&\le M_0\sqrt{U} + MB\sqrt{W}(\|\psi'_{0:s-2}(x)\|_2 + 1)
\end{aligned}$$

where U is the total number of neurons. Expanding the above recursion we get

$$\begin{aligned}
\|\psi'_{0:s-1}(x)\|_2 &\le (MB\sqrt{W})^{s-1}\|x\|_2 + (MB\sqrt{W} + M_0\sqrt{U})\frac{(MB\sqrt{W})^{s-1} - 1}{MB\sqrt{W} - 1} \\
&\le (MB\sqrt{W})^{s-1}(\|x\|_2 + MB\sqrt{W} + M_0\sqrt{U}).
\end{aligned}$$

Finally, we substitute the upper bounds in (J.4) to obtain the final bound

$$\begin{aligned}
&|h(x;W_1',b_1',\ldots,W_S',b_S') - h(x;W_1,b_1,\ldots,W_S,b_S)| \\
\leq\ & \sum_{s=1}^{S} M(MB\sqrt{W})^{S-s}((MB\sqrt{W})^{s-1}(\|x\|_2 + MB\sqrt{W} + M_0\sqrt{U}) + 1)\|\big[\Delta W_s, \Delta b_s\big]\|_F \\
\leq\ & \sum_{s=1}^{S} (MB\sqrt{W})^{S-s}(MB\sqrt{W})^{s}(\|x\|_2 + MB\sqrt{W} + M_0\sqrt{U})\|\big[\Delta W_s, \Delta b_s\big]\|_F \\
\leq\ & (MB\sqrt{W})^{S}(\|x\|_2 + MB\sqrt{W} + M_0\sqrt{U}) \sum_{s=1}^{S} \|\big[\Delta W_s, \Delta b_s\big]\|_F \\
\leq\ & (MB\sqrt{W})^{S}(\|x\|_2 + MB\sqrt{W} + M_0\sqrt{U}) \cdot \sqrt{S}\|\Delta\theta\|_2
\end{aligned}$$

giving rise to the Lipschitz constant. □

J.2 Proofs for Results in Section 5

We first introduce several Berry-Esseen theorems in high dimensions that serve as the main tools of the proofs. Let $\{X_i = (X_i^{(1)}, \ldots, X_i^{(m)}) : i = 1, \ldots, n\}$ be an i.i.d. data set from $\mathbb{R}^m$. Let $\bar{X}^{(j)} = (1/n)\sum_{i=1}^n X_i^{(j)}$ be the sample mean of the j-th component, and $\hat{\Sigma}$ be the sample covariance matrix formed from the data. We denote by $\hat{Z} := (\hat{Z}^{(1)}, \ldots, \hat{Z}^{(m)})$ an m-dimensional multivariate Gaussian with mean zero and covariance $\hat{\Sigma}$. Then under several light-tail conditions:

Assumption 3. $\mathrm{Var}[X_1^{(j)}] > 0$ *for all* $j = 1, \ldots, m$ *and there exists some constant* $D \geq 1$ *such that*

$$\mathbb{E}\big[\exp\Big(\frac{|X_1^{(j)} - \mathbb{E}[X_1^{(j)}]|^2}{D^2\mathrm{Var}[X_1^{(j)}]}\Big)\big] \leq 2 \text{ for all } j = 1, \ldots, m$$
$$\mathbb{E}\big[\Big(\frac{|X_1^{(j)} - \mathbb{E}[X_1^{(j)}]|}{\sqrt{\mathrm{Var}[X_1^{(j)}]}}\Big)^{2+k}\big] \leq D^k \text{ for all } j = 1, \ldots, m \text{ and } k = 1, 2.$$

Assumption 4. *Each* $X_1^{(j)}$ *is* $[0,1]$*-valued and* $\mathrm{Var}[X_1^{(j)}] \geq \eta$ *for all* $j = 1, \ldots, m$ *and some constant* $\eta > 0$.

we have the following Berry Esseen theorems (Chernozhukov et al. (2017)):

Lemma J.10 (Unnormalized supremum, adopted from Theorem EC.9 in Lam and Qian (2019))**.** *Under Assumption 3, for every* $0 < \beta < 1$ *we have*

$$|\mathbb{P}(\sqrt{n}(\bar{X}^{(j)} - \mathbb{E}[X_1^{(j)}]) \leq q_{1-\beta} \text{ for all } j = 1, \ldots, m) - (1-\beta)| \leq C\Big(\frac{D^2\log^7(mn)}{n}\Big)^{\frac{1}{6}}$$

where $q_{1-\beta}$ *is the* $1-\beta$ *quantile of* $\max_{1\leq j\leq m}\hat{Z}^{(j)}$*, i.e.*

$$\mathbb{P}(\hat{Z}^{(j)} \leq q_{1-\beta} \text{ for all } j = 1, \ldots, m | \{X_i : i = 1, \ldots, n\}) = 1 - \beta$$

and C *is a universal constant.*

Lemma J.11 (Normalized supremum, adopted from Theorem EC.10 in Lam and Qian (2019))**.** *Let* $\hat{\sigma}_j^2 = \hat{\Sigma}_{j,j}$*. Under Assumptions 3 and 4, for every* $0 < \beta < 1$ *we have*

$$\begin{aligned}
&|\mathbb{P}(\sqrt{n}(\bar{X}_{(j)} - \mathbb{E}[X_1^{(j)}]) \leq \hat{\sigma}_j q_{1-\beta} \text{ for all } j = 1, \ldots, m) - (1-\beta)| \\
\leq& C\Big(\Big(\frac{D^2\log^7(mn)}{n}\Big)^{\frac{1}{6}} + \frac{\log^2(mn)}{\sqrt{n\eta}} + m\exp\big(-c\eta D^{2/3}n^{2/3}\big)\Big).
\end{aligned}$$

Here $q_{1-\beta}$ *is such that*

$$\mathbb{P}(\hat{Z}^{(j)} \leq \hat{\sigma}_j q_{1-\beta} \text{ for all } j = 1, \ldots, m | \{X_i\}_{i=1}^n) = 1 - \beta$$

and C, c *are universal constants.*

We are now ready to prove Theorems F.1 and 5.1:

Proof of Theorem F.1. Define events

$$E_1 = \big\{\hat{\mathrm{CR}}(\mathrm{PI}_j) \geq 1-\alpha_{\min} + \frac{q'_{1-\beta}}{\sqrt{n_2}} \text{ for some } j = 1,\dots,m\big\}$$
$$E_2 = \big\{\mathrm{CR}(\mathrm{PI}_j) \geq \hat{\mathrm{CR}}(\mathrm{PI}_j) - \frac{q'_{1-\beta}}{\sqrt{n_2}} \text{ for all } j \text{ such that } \mathrm{CR}(\mathrm{PI}_j) \in (\tilde{\alpha}/2, 1-\alpha_{\min})\big\}$$
$$E_3 = \big\{\hat{\mathrm{CR}}(\mathrm{PI}_j) < \tilde{\alpha} + \frac{q'_{1-\beta}}{\sqrt{n_2}} \text{ for all } j \text{ such that } \mathrm{CR}(\mathrm{PI}_j) \leq \tilde{\alpha}/2\big\}.$$

We claim that if $E_1 \cap E_2 \cap E_3$ happens then we must have that $\mathrm{CR}(\mathrm{PI}_{j^*_{1-\alpha_k}}) \geq 1-\alpha_k$ for all $k = 1,\dots,K$. To explain, for each k, E_1 entails that the optimization problem in Step 3 of Algorithm 2 has at least one feasible solution, and E_2 and E_3 further imply that every interval with a true coverage level strictly less than $1-\alpha_k$ violates the margin constraint, hence the selected interval must have a true coverage level of at least $1-\alpha_k$. Therefore we have

$$\begin{aligned}
\mathbb{P}_{\mathcal{D}_v}(\mathrm{CR}(\mathrm{PI}_{j^*_{1-\alpha_k}}) \geq 1-\alpha_k \text{ for all } k=1,\dots,K) &\geq \mathbb{P}_{\mathcal{D}_v}(E_1 \cap E_2 \cap E_3) \\
&\geq 1 - \mathbb{P}_{\mathcal{D}_v}(E_1^c) - \mathbb{P}_{\mathcal{D}_v}(E_2^c) - \mathbb{P}_{\mathcal{D}_v}(E_3^c) \\
&= \mathbb{P}_{\mathcal{D}_v}(E_2) - \mathbb{P}_{\mathcal{D}_v}(E_1^c) - \mathbb{P}_{\mathcal{D}_v}(E_3^c). \qquad \text{(J.5)}
\end{aligned}$$

We derive bounds for the three probabilities. Let $\tilde{q}_{1-\beta}$ be the $1-\beta$ quantile of $\max\{Z_j : \mathrm{CR}(\mathrm{PI}_j) \in (\tilde{\alpha}/2, 1-\alpha_{\min}), 1 \leq j \leq m\}$ where $(Z_1,\dots,Z_m) \sim N_m(0,\hat{\Sigma})$. By stochastic dominance it is clear that $\tilde{q}_{1-\beta} \leq q'_{1-\beta}$ almost surely, therefore

$$\begin{aligned}
\mathbb{P}_{\mathcal{D}_v}(E_2) &\geq \mathbb{P}_{\mathcal{D}_v}\big(\mathrm{CR}(\mathrm{PI}_j) \geq \hat{\mathrm{CR}}(\mathrm{PI}_j) - \frac{\tilde{q}_{1-\beta}}{\sqrt{n_v}} \text{ for all } j \text{ such that } \mathrm{CR}(\mathrm{PI}_j) \in (\tilde{\alpha}/2, 1-\alpha_{\min})\big) \\
&\geq 1 - \beta - C\Big(\frac{\log^7(mn_v)}{n_v\tilde{\alpha}}\Big)^{\frac{1}{6}}
\end{aligned}$$

by applying Lemma J.10 to $\{I_{Y\in\mathrm{PI}_j(X)} : \mathrm{CR}(\mathrm{PI}_j) \in (\tilde{\alpha}/2, 1-\alpha_{\min}), 1 \leq j \leq m\}$ and noticing that Assumption 3 is satisfied with $D = \frac{C}{\sqrt{\tilde{\alpha}}}$ for some universal constant C.

We then bound the second probability

$$\begin{aligned}
\mathbb{P}_{\mathcal{D}_v}(E_1^c) &= \mathbb{P}_{\mathcal{D}_v}(\hat{\mathrm{CR}}(\mathrm{PI}_j) < 1-\alpha_{\min} + \frac{q'_{1-\beta}}{\sqrt{n_v}} \text{ for all } j = 1,\dots,m) \\
&\leq \mathbb{P}_{\mathcal{D}_v}(\hat{\mathrm{CR}}(\mathrm{PI}_{\bar{j}}) < 1-\alpha_{\min} + \frac{q'_{1-\beta}}{\sqrt{n_v}}) \text{ where } \bar{j} \text{ is the index such that } \mathrm{CR}(\mathrm{PI}_{\bar{j}}) = 1-\underline{\alpha} \\
&\leq \mathbb{P}_{\mathcal{D}_v}(\hat{\mathrm{CR}}(\mathrm{PI}_{\bar{j}}) < 1-\alpha_{\min} + \frac{C\sqrt{\log(m/\beta)}}{\sqrt{n_v}}) \\
&\quad \text{because } q'_{1-\beta} \leq C\max_j \hat{\sigma}_j\sqrt{\log(m/\beta)} \leq C\sqrt{\log(m/\beta)} \\
&\leq \exp\big(-\frac{n_v\epsilon^2}{2(\underline{\alpha}(1-\underline{\alpha})+\epsilon/3)}\big)
\end{aligned}$$

where in the last line we use Bennett's inequality (e.g., equation (2.10) in Boucheron et al. (2013)). Note that this is further bounded by $\exp\big(-C_2 n_v \min\{\epsilon, \frac{\epsilon^2}{\underline{\alpha}(1-\underline{\alpha})}\}\big)$ with another universal constant C_2.

The third probability can be bounded as

$$\begin{aligned}
\mathbb{P}_{\mathcal{D}_v}(E_3^c) &\leq \mathbb{P}_{\mathcal{D}_v}\big(\hat{\mathrm{CR}}(\mathrm{PI}_j) \geq \tilde{\alpha} \text{ for some } j \text{ such that } \mathrm{CR}(\mathrm{PI}_j) \leq \tilde{\alpha}/2\big) \\
&\leq \sum_{j:\mathrm{CR}(\mathrm{PI}_j)\leq\tilde{\alpha}/2} \mathbb{P}_{\mathcal{D}_v}(\hat{\mathrm{CR}}(\mathrm{PI}_j) \geq \tilde{\alpha}) \\
&\leq \sum_{j:\mathrm{CR}(\mathrm{PI}_j)\leq\tilde{\alpha}/2} \exp\big(-\frac{n_v(\tilde{\alpha}/2)^2}{2(\mathrm{CR}(\mathrm{PI}_j)(1-\mathrm{CR}(\mathrm{PI}_j))+\tilde{\alpha}/6)}\big) \text{ by Bennett's inequality} \\
&\leq m\exp\big(-\frac{n_v(\tilde{\alpha}/2)^2}{\tilde{\alpha}(1-\tilde{\alpha}/2)+\tilde{\alpha}/3}\big) \leq m\exp(-C_3 n_v\tilde{\alpha})
\end{aligned}$$

where C_3 is a universal constant. Substituting the bounds into (J.5) leads to the overall probability bound

$$1-\beta-C\Big(\frac{\log^7(mn_v)}{n_v\tilde{\alpha}}\Big)^{\frac{1}{6}} - \exp\big(-C_2 n_v \min\{\epsilon, \frac{\epsilon^2}{\underline{\alpha}(1-\underline{\alpha})}\}\big) - m\exp(-C_3 n_v\tilde{\alpha}).$$

It remains to show that $m\exp(-C_3 n_v\tilde{\alpha})$ is negligible relative to other error terms. Since $\tilde{\alpha} < 1$ it is clear that $\big(\frac{1}{n_v}\big)^{1/6} \leq \big(\frac{\log^7(mn_v)}{n_v\tilde{\alpha}}\big)^{1/6}$, and we argue that $\big(\frac{1}{n_v}\big)^{1/6} \geq m\exp(-C_3 n_v\tilde{\alpha})$ can be assumed so that $m\exp(-C_3 n_v\tilde{\alpha}) \leq \big(\frac{\log^7(mn_v)}{n_v\tilde{\alpha}}\big)^{1/6}$. If $\big(\frac{1}{n_v}\big)^{1/6} < m\exp(-C_3 n_v\tilde{\alpha})$, then $m > \exp(C_3 n_v\tilde{\alpha}) n_v^{-1/6}$, hence $\frac{\log^7(mn_v)}{n_v\tilde{\alpha}} \geq \frac{(C_3 n_v\tilde{\alpha})^7}{n_v\tilde{\alpha}} \geq C_3^7 (n_v\tilde{\alpha})^6$, which ultimately leads to $n_v\tilde{\alpha} \leq \frac{\log(mn_v)}{C_3}$ and $\frac{\log^7(mn_v)}{n_v\tilde{\alpha}} \geq C_3 \log^6(mn_v)$. Note that in this case the first error term already exceeds 1 (by enlarging the universal constant C if necessary) and the error bound holds true trivially. □

Proof of Theorem 5.1. The proof follows the one for Theorem F.1, and we focus on the modifications. The events are now defined as

$$\begin{aligned}
E_1 &= \big\{\hat{\mathrm{CR}}(\mathrm{PI}_j) \geq 1-\alpha_{\min} + \frac{q_{1-\beta}\hat{\sigma}_j}{\sqrt{n_v}} \text{ for some } j = 1,\dots,m\big\} \\
E_2 &= \big\{\mathrm{CR}(\mathrm{PI}_j) \geq \hat{\mathrm{CR}}(\mathrm{PI}_j) - \frac{q_{1-\beta}\hat{\sigma}_j}{\sqrt{n_v}} \text{ for all } j \text{ such that } \mathrm{CR}(\mathrm{PI}_j) \in (\tilde{\alpha}/2, 1-\alpha_{\min})\big\} \\
E_3 &= \big\{\hat{\mathrm{CR}}(\mathrm{PI}_j) < \tilde{\alpha} + \frac{q_{1-\beta}\hat{\sigma}_j}{\sqrt{n_v}} \text{ for all } j \text{ such that } \mathrm{CR}(\mathrm{PI}_j) \leq \tilde{\alpha}/2\big\}.
\end{aligned}$$

Again we have $\mathbb{P}_{\mathcal{D}_v}(\mathrm{CR}(\mathrm{PI}_{j^*_{1-\alpha_k}}) \geq 1-\alpha_k \text{ for all } k = 1,\dots,K) \geq \mathbb{P}_{\mathcal{D}_v}(E_2) - \mathbb{P}_{\mathcal{D}_v}(E_1^c) - \mathbb{P}_{\mathcal{D}_v}(E_3^c)$.

The first probability bound becomes

$$\mathbb{P}_{\mathcal{D}_v}(E_2) \geq 1-\beta-C\Big(\Big(\frac{\log^7(mn_v)}{n_v\tilde{\alpha}}\Big)^{\frac{1}{6}} + \frac{\log^2(mn_v)}{\sqrt{n_v\tilde{\alpha}}} + m\exp\big(-c(n_v\tilde{\alpha})^{2/3}\big)\Big)$$

by applying Lemma J.11 and noting that Assumption 4 holds with $\eta = \tilde{\alpha}/2 \cdot (1-\tilde{\alpha}/2) \geq \frac{1}{4}\tilde{\alpha}$ and $D = \frac{C}{\sqrt{\tilde{\alpha}}}$ in Assumption 3. Here C, c are universal constants.

For the second probability we have

$$
\begin{aligned}
\mathbb{P}_{\mathcal{D}_v}(E_1^c) &\le \mathbb{P}_{\mathcal{D}_v}(\hat{\mathrm{CR}}(\mathrm{PI}_{\bar{j}}) < 1-\alpha_{\min} + \frac{q_{1-\beta}\hat{\sigma}_{\bar{j}}}{\sqrt{n_v}}) \ \text{ where } \bar{j} \text{ is the index such that } \mathrm{CR}(\mathrm{PI}_{\bar{j}}) = 1-\underline{\alpha} \\
&\le \mathbb{P}_{\mathcal{D}_v}(\hat{\mathrm{CR}}(\mathrm{PI}_{\bar{j}}) < 1-\alpha_{\min} + \frac{q_{1-\beta}t}{\sqrt{n_v}}) + \mathbb{P}_{\mathcal{D}_v}(\hat{\sigma}_{\bar{j}} > t) \\
&\quad \text{where } t = \sqrt{\underline{\alpha}(1-\underline{\alpha})} + \sqrt{2\log(n_v\alpha_{\min})/n_v} \\
&\le \mathbb{P}_{\mathcal{D}_v}(\hat{\mathrm{CR}}(\mathrm{PI}_{\bar{j}}) < 1-\alpha_{\min} + \frac{q_{1-\beta}t}{\sqrt{n_v}}) + \frac{1}{n_v\alpha_{\min}} \\
&\quad \text{where the bound } 1/(n_v\alpha_{\min}) \text{ follows from Theorem 10 in Maurer and Pontil (2009)} \\
&\le \mathbb{P}_{\mathcal{D}_v}(\hat{\mathrm{CR}}(\mathrm{PI}_{\bar{j}}) < 1-\alpha_{\min} + \frac{C\sqrt{(\underline{\alpha}(1-\underline{\alpha}) + \log(n_v\alpha_{\min})/n_v)\log(m/\beta)}}{\sqrt{n_v}}) + \frac{1}{n_v\alpha_{\min}} \\
&\quad \text{because } q_{1-\beta} \le C\sqrt{\log(m/\beta)} \qquad \text{(J.6)} \\
&\le \exp\big(-\frac{n_v\epsilon^2}{2(\underline{\alpha}(1-\underline{\alpha}) + \epsilon/3)}\big) + \frac{1}{n_v\alpha_{\min}} \ \text{ by Bennett's inequality} \\
&\le \exp\big(-C_2 n_v \min\{\epsilon, \frac{\epsilon^2}{\underline{\alpha}(1-\underline{\alpha})}\}\big) + \frac{1}{n_v\alpha_{\min}}.
\end{aligned}
$$

As for the third probability, by repeating the same analysis in Theorem F.1 we see that the bound $\mathbb{P}_{\mathcal{D}_v}(E_3^c) \le m\exp(-C_3 n_v\tilde{\alpha})$ remains valid.

Finally, using a similar argument in the proof of Theorem F.1, we can show that $\frac{1}{n_v\alpha_{\min}}$, $m\exp(-C_3 n_v\tilde{\alpha})$, and $m\exp(-c(n_v\tilde{\alpha})^{2/3})$ are all dominated by $\big(\frac{\log^7(mn_v)}{n_v\tilde{\alpha}}\big)^{1/6}$ when $\big(\frac{\log^7(mn_v)}{n_v\tilde{\alpha}}\big)^{1/6} < 1$. Moreover, $\frac{\log^2(mn_v)}{\sqrt{n_v\tilde{\alpha}}}$ can also be neglected, because $\frac{\log^2(mn_v)}{\sqrt{n_v\tilde{\alpha}}} = \big(\frac{\log^{5/2}(mn_v)}{n_v\tilde{\alpha}}\big)^{\frac{1}{3}}\big(\frac{\log^7(mn_v)}{n_v\tilde{\alpha}}\big)^{\frac{1}{6}} \le \big(\frac{\log^7(mn_v)}{n_v\tilde{\alpha}}\big)^{\frac{1}{6}}$ when $\big(\frac{\log^7(mn_v)}{n_v\tilde{\alpha}}\big)^{\frac{1}{6}} < 1$. Therefore the desired conclusion follows from combining the three probability bounds. □

Proof of Theorem E.1. The proof consists of deriving concentration bounds for the empirical width and coverage rate. We first deal with the empirical width. Using standard concentration bounds for sub-Gaussian variables, we write for every interval $\mathrm{PI}_j = [L_j, U_j]$ and every $\epsilon > 0$

$$
\begin{aligned}
&\mathbb{P}_{\mathcal{D}_v}\big(|\frac{1}{n_v}\sum_{i=1}^{n_v}(U_j(X_i') - L_j(X_i')) - \mathbb{E}_{\pi_X}[U_j(X) - L_j(X)]| > \epsilon\|H\|_{\psi_2}\big) \\
\le\ &\mathbb{P}_{\mathcal{D}_v}\big(|\frac{1}{n_v}\sum_{i=1}^{n_v}U_j(X_i') - \mathbb{E}_{\pi_X}[U_j(X)]| > \frac{\epsilon\|H\|_{\psi_2}}{2}\big) + \mathbb{P}_{\mathcal{D}_v}\big(|\frac{1}{n_v}\sum_{i=1}^{n_v}L_j(X_i') - \mathbb{E}_{\pi_X}[L_j(X)]| > \frac{\epsilon\|H\|_{\psi_2}}{2}\big) \\
&\quad \text{by the union bound} \\
\le\ &2\exp\Big(-\frac{\epsilon^2\|H\|_{\psi_2}^2 n_v}{4C^2\|U_j - \mathbb{E}_{\pi_X}[U_j]\|_{\psi_2}^2}\Big) + 2\exp\Big(-\frac{\epsilon^2\|H\|_{\psi_2}^2 n_v}{4C^2\|L_j - \mathbb{E}_{\pi_X}[L_j]\|_{\psi_2}^2}\Big) \\
&\quad \text{for some universal constant } C \\
\le\ &4\exp\Big(-\frac{\epsilon^2\|H\|_{\psi_2}^2 n_v}{4C^2\|H\|_{\psi_2}^2}\Big) \quad \text{since } |L_j - \mathbb{E}_{\pi_X}[L_j]|, |U_j - \mathbb{E}_{\pi_X}[U_j]| \le H \\
=\ &4\exp\Big(-\frac{\epsilon^2 n_v}{4C^2}\Big).
\end{aligned}
$$

Applying the union bound to all the m candidate PIs, we have

$$
\mathbb{P}_{\mathcal{D}_v}\big(|\frac{1}{n_v}\sum_{i=1}^{n_v}(U_j(X_i') - L_j(X_i')) - \mathbb{E}_{\pi_X}[U_j(X) - L_j(X)]| > \epsilon\|H\|_{\psi_2} \text{ for some } j = 1,\dots,m\big) \le 4m\exp\Big(-\frac{\epsilon^2 n_v}{4C^2}\Big)
$$

or equivalently

$$\mathbb{P}_{\mathcal{D}_v}\big(|\frac{1}{n_v}\sum_{i=1}^{n_v}(U_j(X_i')-L_j(X_i'))-\mathbb{E}_{\pi_X}[U_j(X)-L_j(X)]|>C\epsilon\|H\|_{\psi_2}\text{ for some }j=1,\dots,m\big)\le 4m\exp\Big(-\frac{\epsilon^2 n_v}{4}\Big). \tag{J.7}$$

Next we handle the empirical coverage rate

$$\begin{aligned}
&\mathbb{P}_{\mathcal{D}_v}\big(\hat{\mathrm{CR}}(\mathrm{PI}_j)-\mathrm{CR}(\mathrm{PI}_j)<-\epsilon+\frac{q_{1-\beta}\hat{\sigma}_j}{\sqrt{n_v}}\big)\\
\le\ &\mathbb{P}_{\mathcal{D}_v}\big(\hat{\mathrm{CR}}(\mathrm{PI}_j)-\mathrm{CR}(\mathrm{PI}_j)<-\epsilon+\frac{C\sqrt{\log(m/\beta)}}{\sqrt{n_v}}\big)\quad\text{by (J.6) and the fact that }\hat{\sigma}_j\le\frac{1}{2}\\
&\quad\text{where } C \text{ is another universal constant}\\
\le\ &\exp\Big(-2\max\big\{\epsilon-C\sqrt{\frac{\log(m/\beta)}{n_v}},0\big\}^2 n_v\Big)\quad\text{by Hoeffding's inequality.}
\end{aligned}$$

Again applying the union bound we get for every $\epsilon>0$

$$\mathbb{P}_{\mathcal{D}_v}\big(\hat{\mathrm{CR}}(\mathrm{PI}_j)-\mathrm{CR}(\mathrm{PI}_j)<-\epsilon+\frac{q_{1-\beta}\hat{\sigma}_j}{\sqrt{n_v}}\text{ for some }j=1,\dots,m\big)\le m\exp\Big(-2\max\big\{\epsilon-C\sqrt{\frac{\log(m/\beta)}{n_v}},0\big\}^2 n_v\Big). \tag{J.8}$$

Note that by choosing both universal constants in (J.7) and (J.8) large enough, we can use the same universal constant C in both. To relate to the width performance, we observe that when $\hat{\mathrm{CR}}(\mathrm{PI}_j)-\mathrm{CR}(\mathrm{PI}_j)\ge-\epsilon+\frac{q_{1-\beta}\hat{\sigma}_j}{\sqrt{n_v}}$ for all $j=1,\dots,m$, we have that for all PI_j whose true coverage rate $\mathrm{CR}(\mathrm{PI}_j)\ge 1-\alpha_k+\epsilon$ the inequality $\hat{\mathrm{CR}}(\mathrm{PI}_j)\ge\mathrm{CR}(\mathrm{PI}_j)-\epsilon+\frac{q_{1-\beta}\hat{\sigma}_j}{\sqrt{n_v}}\ge 1-\alpha_k+\frac{q_{1-\beta}\hat{\sigma}_j}{\sqrt{n_v}}$ holds (i.e., the constraint in Step 3 of Algorithm 1 is satisfied), therefore by the optimality of each $\mathrm{PI}_{j^*_{1-\alpha_k}}$ it must hold that

$$\frac{1}{n_v}\sum_{i=1}^{n_v}|\mathrm{PI}_{j^*_{1-\alpha_k}}(X_i')|\le\min_{j:\mathrm{CR(PI_j)}\ge 1-\alpha_\mathrm{k}+\epsilon}\frac{1}{n_v}\sum_{i=1}^{n_v}|\mathrm{PI}_j(X_i')|\quad\text{for each }k=1,\dots,K.$$

If we further have $|\frac{1}{n_v}\sum_{i=1}^{n_v}(U_j(X_i')-L_j(X_i'))-\mathbb{E}_{\pi_X}[U_j(X)-L_j(X)]|\le C\epsilon\|H\|_{\psi_2}$ for all $j=1,\dots,m$, then

$$\begin{aligned}
\mathbb{E}_{\pi_X}[U_{j^*_{1-\alpha_k}}(X)-L_{j^*_{1-\alpha_k}}(X)]\ &\le\ \frac{1}{n_v}\sum_{i=1}^{n_v}|\mathrm{PI}_{j^*_{1-\alpha_k}}(X_i')|+C\epsilon\|H\|_{\psi_2}\\
&\le\ \min_{j:\mathrm{CR(PI_j)}\ge 1-\alpha_\mathrm{k}+\epsilon}\mathbb{E}_{\pi_X}[U_j(X)-L_j(X)]+2C\epsilon\|H\|_{\psi_2}
\end{aligned}$$

for every $k=1,\dots,K$. Altogether we can conclude that

$$\begin{aligned}
&\mathbb{P}_{\mathcal{D}_v}\Big(\mathbb{E}_{\pi_X}[U_{j^*_{1-\alpha_k}}(X)-L_{j^*_{1-\alpha_k}}(X)]\le\min_{j:\mathrm{CR(PI_j)}\ge 1-\alpha_\mathrm{k}+\epsilon}\mathbb{E}_{\pi_X}[U_j(X)-L_j(X)]+2C\epsilon\|H\|_{\psi_2}\text{ for all }k=1,\dots,K\Big)\\
\ge\ &\mathbb{P}_{\mathcal{D}_v}\Big(|\frac{1}{n_v}\sum_{i=1}^{n_v}(U_j(X_i')-L_j(X_i'))-\mathbb{E}_{\pi_X}[U_j(X)-L_j(X)]|\le C\epsilon\|H\|_{\psi_2}\text{ for all }j=1,\dots,m,\text{ and}\\
&\quad\hat{\mathrm{CR}}(\mathrm{PI}_j)-\mathrm{CR}(\mathrm{PI}_j)\ge-\epsilon+\frac{q_{1-\beta}\hat{\sigma}_j}{\sqrt{n_v}}\text{ for all }j=1,\dots,m\Big)\\
\ge\ &1-4m\exp\Big(-\frac{\epsilon^2 n_v}{4}\Big)-m\exp\Big(-2\max\big\{\epsilon-C\sqrt{\frac{\log(m/\beta)}{n_v}},0\big\}^2 n_v\Big)\quad\text{by (J.7) and (J.8)}\\
\ge\ &1-8m\exp\Big(-\frac{1}{4}\max\big\{\epsilon-C\sqrt{\frac{\log(m/\beta)}{n_v}},0\big\}^2 n_v\Big)
\end{aligned}$$

where the last inequality holds because $\epsilon\ge\max\big\{\epsilon-C\sqrt{\frac{\log(m/\beta)}{n_v}},0\big\}$. □

J.3 Proofs for Appendix A

We provide proofs for Theorems A.1 and A.2:

Proof of Theorem A.1. We first construct the sequence t_n as follows. For each integer $k > 0$, weak π-GC implies that $\mathbb{P}\big(\sup_{L,U\in\mathcal{H},L\le U}|\mathbb{P}_{\hat{\pi}}(Y \in [L(X),U(X)]) - \mathbb{P}_{\pi}(Y \in [L(X),U(X)]| > \frac{1}{k}\big) \to 0$ as the data size $n \to \infty$. Therefore, there exists a large enough n_k such that $\mathbb{P}\big(\sup_{L,U\in\mathcal{H},L\le U}|\mathbb{P}_{\hat{\pi}}(Y \in [L(X),U(X)]) - \mathbb{P}_{\pi}(Y \in [L(X),U(X)]| > \frac{1}{k}\big) < \frac{1}{k}$ whenever $n \ge n_k$. Moreover, the n_k can be chosen such that $n_k < n_{k+1}$ for each k. We then let $t_n = \min\{\frac{1}{k} : n \ge n_k\}$. Clearly $t_n \to 0$ as $n \to \infty$, and, by construction, we have $\mathbb{P}\big(\sup_{L,U\in\mathcal{H},L\le U}|\mathbb{P}_{\hat{\pi}}(Y \in [L(X),U(X)]) - \mathbb{P}_{\pi}(Y \in [L(X),U(X)]| > t_n\big) < t_n$.

Then we show joint optimality and feasibility for the chosen t_n. Denote by $\hat{\mathcal{H}}_t^2$ the feasible set of the optimization (3.2), and by $\mathcal{H}_t^2$ the feasible set of (4.1). In particular $\mathcal{H}_0^2$ is the feasible set of (3.1). By the construction of t_n, we have $\mathbb{P}(\mathcal{H}_{2t_n}^2 \subset \hat{\mathcal{H}}_{t_n}^2 \subset \mathcal{H}_0^2) > 1 - t_n$. Therefore $\mathbb{P}((\hat{L}_{t_n}^*, \hat{U}_{t_n}^*) \in \mathcal{H}_0^2) \ge \mathbb{P}(\hat{\mathcal{H}}_{t_n}^2 \in \mathcal{H}_0^2) \ge 1 - t_n \to 0$, concluding the asymptotic feasibility of $(\hat{L}_{t_n}^*, \hat{U}_{t_n}^*)$. To show optimality, when the events

$$W_\epsilon := \big\{ \sup_{h\in\mathcal{H}} |\mathbb{E}_{\hat{\pi}_X}[h(X)] - \mathbb{E}_{\pi_X}[h(X)]| \le \epsilon \big\}$$

and

$$C_{t_n} := \big\{ \sup_{L,U\in\mathcal{H} \text{ and } L\le U} |\mathbb{P}_{\hat{\pi}_X}(Y \in [L(X),U(X)]) - \mathbb{P}_{\pi_X}(Y \in [L(X),U(X)])| \le t_n \big\}$$

occur, it holds that $\mathcal{H}_{2t_n}^2 \subset \hat{\mathcal{H}}_{t_n}^2 \subset \mathcal{H}_0^2$, and $(\hat{L}_{t_n}^*, \hat{U}_{t_n}^*) \in \hat{\mathcal{H}}_{t_n}^2$ is feasible for (3.1). We also have

$$\begin{aligned}
& \mathbb{E}_{\pi_X}[\hat{U}_{t_n}^*(X) - \hat{L}_{t_n}^*(X)] \\
\le\ & \mathbb{E}_{\hat{\pi}_X}[\hat{U}_{t_n}^*(X) - \hat{L}_{t_n}^*(X)] + 2\epsilon \ \text{ because of } W_\epsilon \\
\le\ & \inf_{(L,U)\in\mathcal{H}_{2t_n}^2 \subset \hat{\mathcal{H}}_{t_n}^2} \mathbb{E}_{\hat{\pi}_X}[U(X) - L(X)] + 2\epsilon \ \text{ by optimality of } (\hat{L}_{t_n}^*, \hat{U}_{t_n}^*) \text{ in } \hat{\mathcal{H}}_{t_n}^2 \\
\le\ & \inf_{(L,U)\in\mathcal{H}_{2t_n}^2} \mathbb{E}_{\pi_X}[U(X) - L(X)] + 4\epsilon \ \text{ because of } W_\epsilon \\
=\ & \mathcal{R}_{2t_n}^*(\mathcal{H}) + 4\epsilon \\
\le\ & \mathcal{R}^*(\mathcal{H}) + \frac{12t_n}{(\alpha - 2t_n)\gamma_{\frac{\alpha-2t_n}{3}}} + 4\epsilon \ \text{ by Theorem 4.1.}
\end{aligned} \tag{J.9}$$

Note that $\mathbb{P}(W_\epsilon \cap C_{t_n}) \ge 1 - \mathbb{P}(W_\epsilon^c) - \mathbb{P}(C_{t_n}^c) \ge 1 - \mathbb{P}(W_\epsilon^c) - t_n$. Note that $\mathcal{H}$ being π-GC implies that $\mathbb{P}(W_\epsilon^c) \to 0$ as $n \to \infty$ for any $\epsilon > 0$, and that the term involving t_n in (J.9) goes to zero as $t_n \to 0$. On the other hand, $\mathbb{E}_{\pi_X}[\hat{U}_{t_n}^*(X) - \hat{L}_{t_n}^*(X)] \ge \mathcal{R}^*(\mathcal{H})$ when $(\hat{L}_{t_n}^*, \hat{U}_{t_n}^*) \in \mathcal{H}_0^2$ by the definition of $\mathcal{R}^*(\mathcal{H})$. Since $\mathbb{P}(\mathcal{H}_{2t_n}^2 \subset \hat{\mathcal{H}}_{t_n}^2$ and $(\hat{L}_{t_n}^*, \hat{U}_{t_n}^*) \in \mathcal{H}_0^2) \to 1$, the derived lower and upper bounds for $\mathbb{E}_{\pi_X}[\hat{U}_{t_n}^*(X) - \hat{L}_{t_n}^*(X)]$ entails that $\mathbb{E}_{\pi_X}[\hat{U}_{t_n}^*(X) - \hat{L}_{t_n}^*(X)] \to \mathcal{R}^*(\mathcal{H})$ in probability, concluding optimality.

□

Proof of Theorem A.2. It suffices to show that the induced set class is strong π-GC. Consistency then follows from Theorem A.1 and the fact that strong GC implies weak GC. We will need the following preservation results for GC classes:

Lemma J.12 (Adapted from Theorem 9.26 in Kosorok (2007))**.** *Suppose that $\mathcal{G}_1, \dots, \mathcal{G}_K$ are strong π-GC classes of functions with $\max_{1\le k\le K} \sup_{g\in\mathcal{G}_k} |\mathbb{E}_\pi[g(X,Y)]| < \infty$, and that $\psi : \mathbb{R}^K \to \mathbb{R}$ is continuous. Then the class $\psi(\mathcal{G}_1, \dots, \mathcal{G}_K) := \{\psi(g_1(\cdot), \dots, g_K(\cdot)) : g_k \in \mathcal{G}_k \text{ for } k = 1, \dots, K\}$ is strong π-GC provided that $\mathbb{E}_\pi[\sup_{g\in\psi(\mathcal{G}_1,\dots,\mathcal{G}_K)} |g(X,Y)|] < \infty$.*

We first use Lemma J.12 to simplify the problem. Denote by $\mathcal{G} := \{(x,y) \to I_{L(x)\le y\le U(x)} : L, U \in \mathcal{H}, L \le U\}$ the target indicator class for which we want to show strong π-GC. Then $\mathcal{G} \subset \psi(\mathcal{G}_l, \mathcal{G}_u)$, where $\psi(z_1, z_2) := z_1 z_2$, and

$$\begin{aligned}
\mathcal{G}_l &:= \{(x,y) \to I_{L(x)-y\le 0} : L \in \mathcal{H}\} \\
\mathcal{G}_u &:= \{(x,y) \to I_{y-U(x)\le 0} : U \in \mathcal{H}\}.
\end{aligned}$$

Note that functions in the class $\psi(\mathcal{G}_l, \mathcal{G}_u)$ are all bounded by 1 and $\psi(\cdot,\cdot)$ is obviously continuous, therefore by Lemma J.12, $\psi(\mathcal{G}_l, \mathcal{G}_u)$ (hence $\mathcal{G}$) is strong π-GC if both $\mathcal{G}_l$ and $\mathcal{G}_u$ are strong π-GC. So it suffices to show strong π-GC for $\mathcal{G}_l$ and $\mathcal{G}_u$. In the following analysis, we only show π-GC for $\mathcal{G}_u$, because the $\mathcal{G}_l$ case can be shown by the same argument.

In order to prove strong π-GC for $\mathcal{G}_u$, we first demonstrate that, without loss of generality, we can assume that the class $\mathcal{H}$ has a upper bound for $|\mathbb{E}_{\pi_X}[h(X)]|$, say $|\mathbb{E}_{\pi_X}[h(X)]| \le M$ (so that Lemma J.12 can be used to propagate the GC property from $\mathcal{H}$ to $\mathcal{G}_u$). To proceed, we write for any constant $M > 0$

$$\begin{aligned}
&\sup_{U\in\mathcal{H}} |\mathbb{P}_{\hat{\pi}}(Y - U(X) \le 0) - \mathbb{P}_{\pi}(Y - U(X) \le 0)| \\
\le\ &\sup_{U\in\mathcal{H}, |\mathbb{E}_{\pi_X}[U(X)]|\le M} |\mathbb{P}_{\hat{\pi}}(Y - U(X) \le 0) - \mathbb{P}_{\pi}(Y - U(X) \le 0)| + \qquad \text{(J.10)} \\
&\sup_{U\in\mathcal{H}, \mathbb{E}_{\pi_X}[U(X)]> M} |\mathbb{P}_{\hat{\pi}}(Y - U(X) \le 0) - \mathbb{P}_{\pi}(Y - U(X) \le 0)| + \qquad \text{(J.11)} \\
&\sup_{U\in\mathcal{H}, \mathbb{E}_{\pi_X}[U(X)]< -M} |\mathbb{P}_{\hat{\pi}}(Y - U(X) \le 0) - \mathbb{P}_{\pi}(Y - U(X) \le 0)|. \qquad \text{(J.12)}
\end{aligned}$$

We show how we control the second and third suprema in (J.11) and (J.12). Since the class $\mathcal{H}$ is strong π_X-GC, the centered function class $\{h(\cdot) - \mathbb{E}_{\pi_X}[h(X)] : h \in \mathcal{H}\}$ must have an integrable envelope (see, e.g., Lemma 8.13 in Kosorok (2007)), i.e., $F(x) := \sup_{h(x)\in\mathcal{H}} |h - \mathbb{E}_{\pi_X}[h(X)]|$ satisfies $\mathbb{E}_{\pi_X}[F(X)] < \infty$. Therefore when $\mathbb{E}_{\pi_X}[U(X)] > M$ we can bound $Y - U(X)$ almost surely as

$$Y - U(X) < Y - U(X) + \mathbb{E}_{\pi_X}[U(X)] - M \le Y + F(X) - M.$$

Similarly, when $\mathbb{E}_{\pi_X}[U(X)] < -M$ we have

$$Y - U(X) > Y - U(X) + \mathbb{E}_{\pi_X}[U(X)] + M \ge Y - F(X) + M.$$

With these bounds for $Y - U(X)$, the supremum from (J.11) can be further bounded as

$$\begin{aligned}
&\sup_{U\in\mathcal{H}, \mathbb{E}_{\pi_X}[U(X)]> M} |\mathbb{P}_{\hat{\pi}}(Y - U(X) \le 0) - \mathbb{P}_{\pi}(Y - U(X) \le 0)| \\
\le\ &\sup_{U\in\mathcal{H}, \mathbb{E}_{\pi_X}[U(X)]> M} |\mathbb{P}_{\hat{\pi}}(Y - U(X) \le 0) - 1| + \\
&\qquad \sup_{U\in\mathcal{H}, \mathbb{E}_{\pi_X}[U(X)]> M} |\mathbb{P}_{\pi}(Y - U(X) \le 0) - 1| \ \text{ by triangle inequality} \\
\le\ &|\mathbb{P}_{\hat{\pi}}(Y + F(X) \le M) - 1| + |\mathbb{P}_{\pi}(Y + F(X) \le M) - 1| \\
\le\ &|\mathbb{P}_{\hat{\pi}}(Y + F(X) \le M) - \mathbb{P}_{\pi}(Y + F(X) \le M)| + \\
&\qquad 2|\mathbb{P}_{\pi}(Y + F(X) \le M) - 1| \ \text{ again by triangle inequality} \qquad \text{(J.13)}
\end{aligned}$$

and (J.12) can be similarly bounded as

$$\begin{aligned}
&\sup_{U\in\mathcal{H}, \mathbb{E}_{\pi_X}[U(X)]< -M} |\mathbb{P}_{\hat{\pi}}(Y - U(X) \le 0) - \mathbb{P}_{\pi}(Y - U(X) \le 0)| \\
\le\ &\sup_{U\in\mathcal{H}, \mathbb{E}_{\pi_X}[U(X)]< -M} |\mathbb{P}_{\hat{\pi}}(Y - U(X) \le 0)| + \sup_{U\in\mathcal{H}, \mathbb{E}_{\pi_X}[U(X)]< -M} |\mathbb{P}_{\pi}(Y - U(X) \le 0)| \\
\le\ &|\mathbb{P}_{\hat{\pi}}(Y - F(X) \le -M)| + |\mathbb{P}_{\pi}(Y - F(X) \le -M)| \\
\le\ &|\mathbb{P}_{\hat{\pi}}(Y - F(X) \le -M) - \mathbb{P}_{\pi}(Y - F(X) \le -M)| + \\
&\qquad 2|\mathbb{P}_{\pi}(Y - F(X) \le -M)| \ \text{ by triangle inequality} \qquad \text{(J.14)}
\end{aligned}$$

Note that $\mathbb{P}_{\pi}(Y + F(X) \le M) \to 1$ as $M \to \infty$ therefore the second absolute value in (J.13) can be made arbitrarily small by choosing M sufficiently large. The first absolute value in (J.13) vanishes almost surely for every fixed M by the classical strong law of large numbers. As a result, for every $\epsilon > 0$, there exists an $M > 0$ such that almost surely there exists an n' for which the second supremum in (J.11) is less than ϵ for all sample size $n > n'$. A similar conclusion can be drawn for (J.14). Therefore it's enough to show that for every fixed M the supremum in (J.10) vanishes almost surely, or in other words, the following constrained version of $\mathcal{G}_u$

$$\mathcal{G}_u^M := \{(x, y) \to I_{y-U(x)\le 0} : U \in \mathcal{H}, |\mathbb{E}_{\pi_X}[U(X)]| \le M\}$$

is strong π-GC.

It remains to show that $\mathcal{G}_u^M$ is strong π-GC. We first note that, since Y is assumed integrable, the class $\mathcal{H}_u^M := \{(x,y) \to y - U(x) : U \in \mathcal{H}, |\mathbb{E}_{\pi_X}[U(X)]| \le M\}$ is strong π-GC, and has an integrable envelope $|Y| + M + F(X)$ where F is the envelope for $\{h(\cdot) - \mathbb{E}_{\pi_X}[h(X)] : h \in \mathcal{H}\}$. Our plan is to propagate GC from $\mathcal{H}_u^M$ to $\mathcal{G}_u^M$ using Lemma J.12. To this end, we define a function $\text{sign}(z) = 1$ if $z \le 0$ and 0 if $z > 0$, then $\mathcal{G}_u^M$ can be written as $\text{sign}(\mathcal{H}_u^M)$. We also define two auxiliary functions both parameterized by $\epsilon > 0$

$$\text{sign}_\epsilon^l(z) := \begin{cases} 1 & \text{if } z \le -\epsilon \\ -\frac{z}{\epsilon} & \text{if } \epsilon < z < 0 \\ 0 & \text{if } z \ge 0 \end{cases}$$

$$\text{sign}_\epsilon^u(z) := \begin{cases} 1 & \text{if } z \le 0 \\ 1 - \frac{z}{\epsilon} & \text{if } 0 < z < \epsilon \\ 0 & \text{if } z \ge \epsilon \end{cases}.$$

Since $\text{sign}_\epsilon^l \le \text{sign} \le \text{sign}_\epsilon^u$, we can approximate the class $\mathcal{G}_u^M$ (i.e., $\text{sign}(\mathcal{H}_u^M)$) from below by $\text{sign}_\epsilon^l(\mathcal{H}_u^M)$ and from above by $\text{sign}_\epsilon^u(\mathcal{H}_u^M)$. Moreover, we have the following approximation bound

$$\begin{aligned}
& \sup_{h \in \mathcal{H}_u^M} \left(\mathbb{E}_\pi[\text{sign}_\epsilon^u(h(X,Y))] - \mathbb{E}_\pi[\text{sign}_\epsilon^l(h(X,Y))]\right) \\
= & \sup_{h \in \mathcal{H}_u^M} \mathbb{E}_\pi[(\text{sign}_\epsilon^u - \text{sign}_\epsilon^l)(h(X,Y))] \\
\le & \sup_{U \in \mathcal{H}, |\mathbb{E}_{\pi_X}[U(X)]| \le M} \mathbb{P}_\pi(Y - U(X) \in (-\epsilon, \epsilon)) \\
& \qquad \text{because } (\text{sign}_\epsilon^u - \text{sign}_\epsilon^l)(z) \le I_{z \in (-\epsilon,\epsilon)} \\
= & \sup_{U \in \mathcal{H}, |\mathbb{E}_{\pi_X}[U(X)]| \le M} \mathbb{E}_{\pi_X}[\mathbb{P}_\pi(Y - U(X) \in (-\epsilon, \epsilon)|X)] \\
\le & \ 2\epsilon \sup_{x,y} p(y|x). \qquad \text{(J.15)}
\end{aligned}$$

On the other hand, both sign_ϵ^l and sign_ϵ^u are continuous and bounded, therefore by Lemma J.12 both $\text{sign}_\epsilon^l(\mathcal{H}_u^M)$ and $\text{sign}_\epsilon^l(\mathcal{H}_u^M)$ are strong π-GC for each fixed $\epsilon > 0$. We can then write

$$\begin{aligned}
& \sup_{h \in \mathcal{H}_u^M} |\mathbb{E}_{\hat{\pi}}[\text{sign}(h(X,Y))] - \mathbb{E}_\pi[\text{sign}(h(X,Y))]| \\
\le & \sup_{h \in \mathcal{H}_u^M} \left(|\mathbb{E}_{\hat{\pi}}[\text{sign}_\epsilon^l(h(X,Y))] - \mathbb{E}_\pi[\text{sign}(h(X,Y))]| + |\mathbb{E}_{\hat{\pi}}[\text{sign}_\epsilon^u(h(X,Y))] - \mathbb{E}_\pi[\text{sign}(h(X,Y))]|\right) \\
& \qquad \text{because } \text{sign}_\epsilon^l \le \text{sign} \le \text{sign}_\epsilon^u \\
\le & \sup_{h \in \mathcal{H}_u^M} \left(|\mathbb{E}_{\hat{\pi}}[\text{sign}_\epsilon^l(h(X,Y))] - \mathbb{E}_\pi[\text{sign}_\epsilon^l(h(X,Y))]| + |\mathbb{E}_{\hat{\pi}}[\text{sign}_\epsilon^u(h(X,Y))] - \mathbb{E}_\pi[\text{sign}_\epsilon^u(h(X,Y))]|\right. \\
& \qquad \left. + \mathbb{E}_\pi[\text{sign}_\epsilon^u(h(X,Y))] - \mathbb{E}_\pi[\text{sign}_\epsilon^l(h(X,Y))]\right) \text{ by triangle inequality} \\
\le & \sup_{h \in \mathcal{H}_u^M} |\mathbb{E}_{\hat{\pi}}[\text{sign}_\epsilon^l(h(X,Y))] - \mathbb{E}_\pi[\text{sign}_\epsilon^l(h(X,Y))]| + \\
& \qquad \sup_{h \in \mathcal{H}_u^M} |\mathbb{E}_{\hat{\pi}}[\text{sign}_\epsilon^u(h(X,Y))] - \mathbb{E}_\pi[\text{sign}_\epsilon^u(h(X,Y))]| + 2\epsilon \sup_{x,y} p(y|x) \text{ by the bound (J.15).}
\end{aligned}$$

Since ϵ is arbitrary, the strong GC of $\mathcal{G}_u^M$ then follows from the strong GC of $\text{sign}_\epsilon^l(\mathcal{H}_u^M)$ and $\text{sign}_\epsilon^l(\mathcal{H}_u^M)$, and the finiteness of $\sup_{x,y} p(y|x)$. This concludes the proof. □

J.4 Proofs for Appendix C

Proof of Theorem C.1. Since $\sup_{h \in \mathcal{H}} |\mathbb{E}_{\hat{\pi}_X}[h(X)] - \mathbb{E}_{\pi_X}[h(X)]| \le B \|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}_{\pi_X}[X]\|_\infty$, we have $\phi_1(n, \epsilon, \mathcal{H}) \le \mathbb{P}(B\|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}_{\pi_X}[X]\|_\infty > \epsilon)$. We bound the sub-Gaussian norm of $\|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}_{\pi_X}[X]\|_\infty$. Let $X^{(j)}, j = 1, \ldots, d$ be the j-th component of the random vector $X \in \mathbb{R}^d$. Since $X_i^{(j)}, i = 1, \ldots, n$ are i.i.d., we have $\|\frac{1}{n}\sum_{i=1}^n X_i^{(j)} - \mathbb{E}_{\pi_X}[X^{(j)}]\|_{\psi_2} \le \frac{C}{\sqrt{n}}\|X^{(j)} - \mathbb{E}_{\pi_X}[X^{(j)}]\|_{\psi_2}$ for some universal constant C, by general

Hoeffding's inequality (e.g., Proposition 2.6.1 in Vershynin (2018)). Now we can use the sub-Gaussian maximal inequality (e.g., Lemma 2.2.2 in Van der Vaart and Wellner (1996)) to get

$$\begin{aligned}
\|\|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}_{\pi_X}[X]\|_\infty\|_{\psi_2} &\leq C'\sqrt{\log d}\max_{1\leq j\leq d}\|\frac{1}{n}\sum_{i=1}^{n} X_i^{(j)} - \mathbb{E}_{\pi_X}[X^{(j)}]\|_{\psi_2} \\
&\qquad \text{for another universal constant } C' \\
&\leq C'C\sqrt{\frac{\log d}{n}}\max_{1\leq j\leq d}\|X^{(j)} - \mathbb{E}_{\pi_X}[X^{(j)}]\|_{\psi_2} \\
&\leq C'C\sqrt{\frac{\log d}{n}}\|\|X - \mathbb{E}_{\pi_X}[X]\|_\infty\|_{\psi_2}
\end{aligned}$$

where the last inequality holds because $|X^{(j)} - \mathbb{E}_{\pi_X}[X^{(j)}]| \leq \|X - \mathbb{E}_{\pi_X}[X]\|_\infty$ implies $\|X^{(j)} - \mathbb{E}_{\pi_X}[X^{(j)}]\|_{\psi_2} \leq \|\|X - \mathbb{E}_{\pi_X}[X]\|_\infty\|_{\psi_2}$ for all $j = 1, \ldots, d$. The expression for ϕ_1 then comes from the sub-Gaussian tail bound. □

References

M. Anthony and P. L. Bartlett. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

P. L. Bartlett, V. Maiorov, and R. Meir. Almost linear vc dimension bounds for piecewise polynomial networks. In *Advances in Neural Information Processing Systems*, pages 190–196, 1999.

P. L. Bartlett, N. Harvey, C. Liaw, and A. Mehrabian. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

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.

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

I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT press, 2016.

M. R. Kosorok. *Introduction to empirical processes and semiparametric inference*. Springer Science & Business Media, 2007.

H. Lam and H. Qian. Combating conservativeness in data-driven optimization under uncertainty: A solution path approach. *arXiv preprint arXiv:1909.06477*, 2019.

Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

A. Maurer and M. Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

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.

M. Talagrand. Sharper bounds for gaussian and empirical processes. *The Annals of Probability*, pages 28–76, 1994.

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.

R. Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

T. Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2(Mar):527–550, 2002.