# 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 $\pi_X$-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 $\pi$-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 $\pi_X$-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 $\pi_X$-GC of the hypothesis class $\mathcal{H}$ implies strong $\pi$-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 $\pi_X$-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 $\phi_1$ 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 $\Sigma$ 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 $\Sigma$ 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 $\epsilon$ 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 $\lambda$ 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 $\lambda$ 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 $\lambda$ 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) $\pi_X$-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) $\pi_X$-GC if the corresponding indicator class $\{I_{\cdot\in S} : S \in \mathcal{S}\}$ is weak (resp. strong) $\pi_X$-GC. When no ambiguity arises, we sometimes suppress the underlying distribution $\pi_X$.

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 $\pi_X$.

Given a class $\mathcal{G}$ of functions from $\mathcal{X}$ to $\mathbb{R}$, and a probability measure $Q$ on $\mathcal{X}$, the $\epsilon$-covering number $N(\epsilon, \mathcal{G}, L_2(Q))$ is the minimum number of $L_2(Q)$-balls of size $\epsilon$ 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 $\epsilon$ 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 $\epsilon$-bracketing number $N_{[]}(\epsilon, \mathcal{G}, L_2(Q))$ is the minimum number of brackets of size $\epsilon$ 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 $\pi$ in a straightforward manner, i.e., by replacing each occurrence of $\mathcal{X}$ and $\pi_X$ with $\mathcal{X} \times \mathcal{Y}$ and $\pi$ 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($\epsilon$) |
|---|---|---|---|---|
| 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.

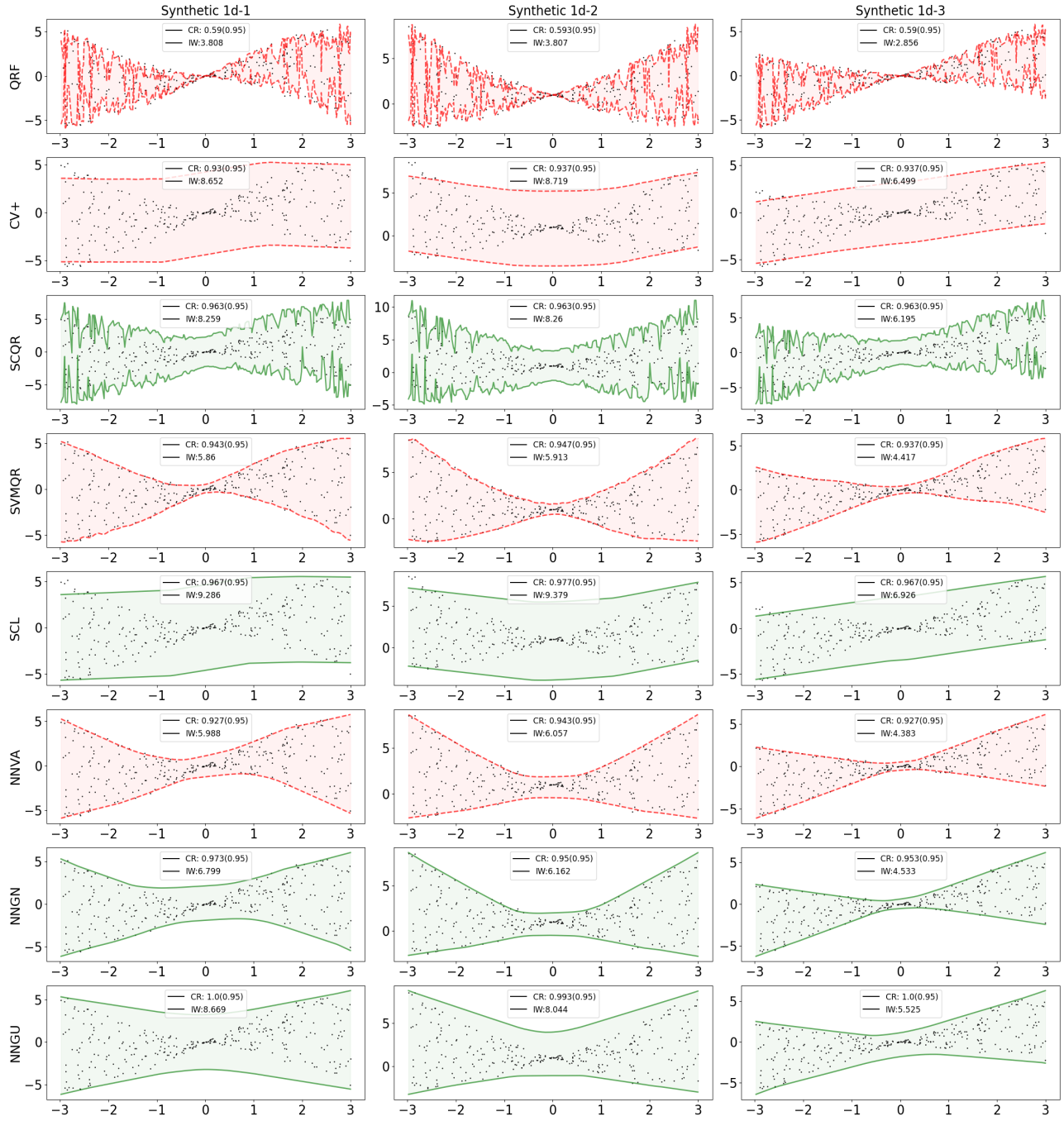


Figure 1: Comparison of single PI constructions. PIs that attain 95% target level are shown in solid and green lines, otherwise in dashed and red lines.

## 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.

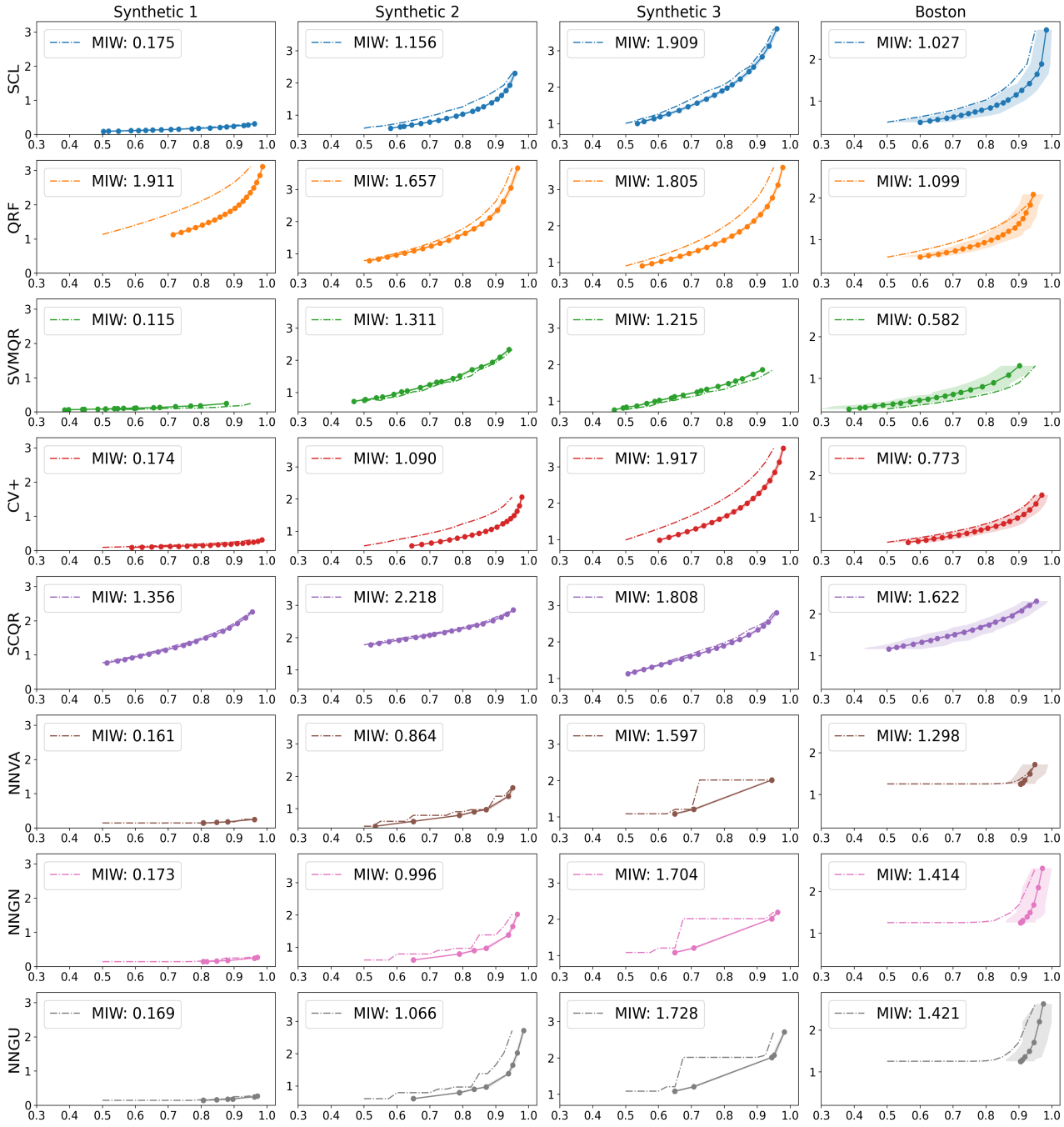


Figure 2: Comparison of simultaneous PIs constructions in synthetic datasets 1 to 3, and Boston dataset. Dashed lines are constructed with (input coverage, width), and solid lines are constructed with (achieved coverage, width). Each point in the line denotes the average obtained from $N = 50$ repetitions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

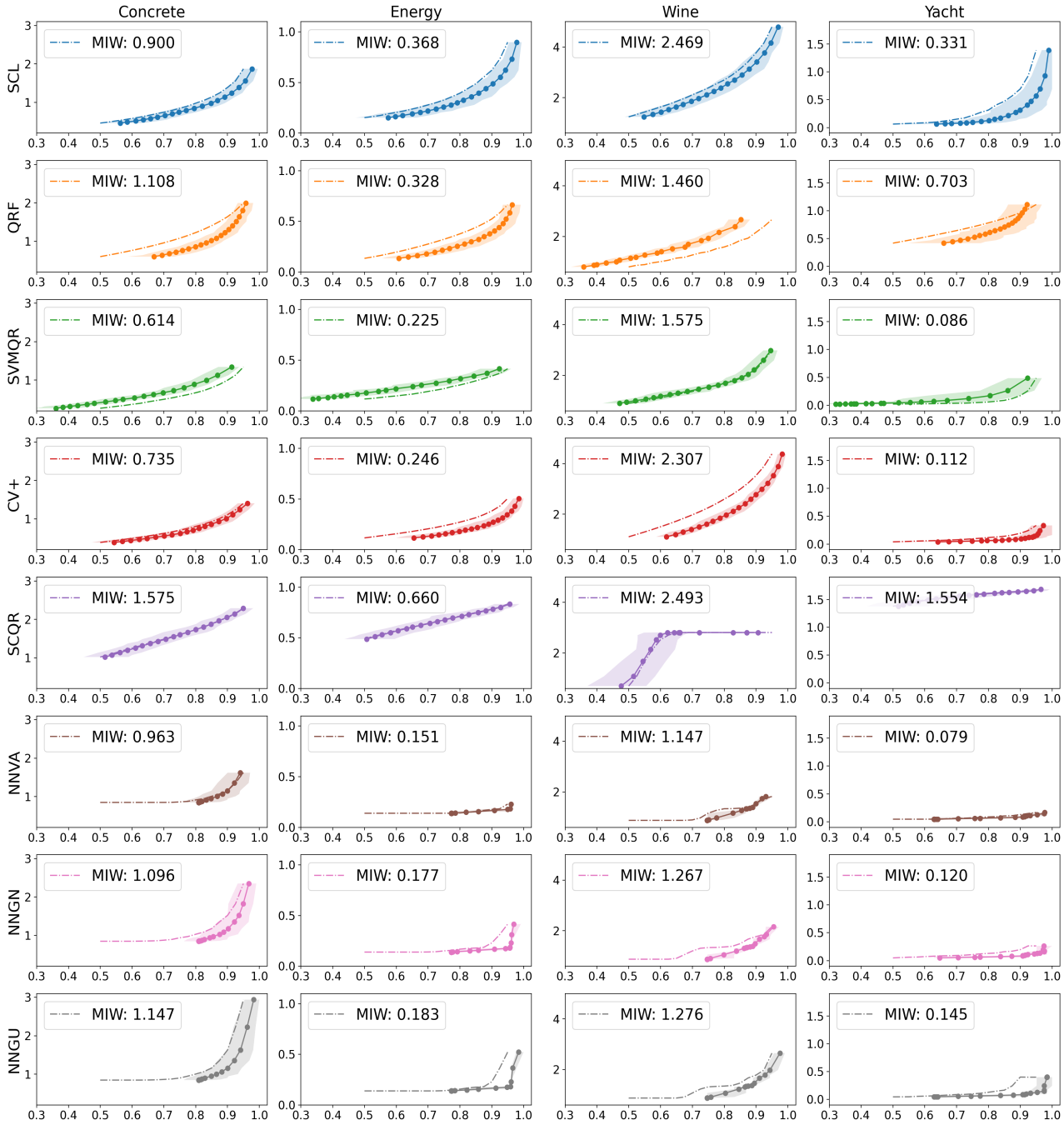


Figure 3: Comparison of simultaneous PIs constructions in Concrete, Energy, Wine and Yacht dataset. Dashed lines are constructed with (input coverage, width), and solid lines are constructed with (achieved coverage, width). Each point in the line denotes the average obtained from $N = 50$ repetitions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

(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 $\xi$ 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 $\epsilon$-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 $\epsilon$ is arbitrary, sending $\epsilon$ 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 $\pi_X^n$ (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 $\pi_X^n$ (i.e., the data), and $C$ is a universal constant.*

We now turn to the main proof. We first deal with $\phi_1$. 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 $\phi_1$.

Next we analyze $\phi_2$. 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 $\phi_2$.

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 $\epsilon$ and $t$, and then applying Theorem 4.2. □

*Proof of Theorem 4.4.* We first treat $\phi_1$. 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 $\pi_X^n$ (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 $\theta$, 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 $\Theta$ 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 $\epsilon$-covering number of $\Theta$ with respect to the $l_2$ norm, i.e., the minimum number of $l_2$-balls of size $\epsilon$ needed to cover $\Theta$. Since $\Theta$ 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 $\phi_1$ follows from the sub-Gaussian tail bound.

Secondly, we analyze $\phi_2$. 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 $\theta_l, \theta_u$ 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 $\theta_l$ and $\theta_u$, 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 $\epsilon$-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 $\kappa$ 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 $\kappa$ 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 $\phi_2$.

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 $\epsilon$ 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 $\phi_2$

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

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 $\mathrm{PI}_j$ 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 $\pi$-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 $\pi$-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 $\pi$-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 $\pi$-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 $\pi$-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 $\pi$-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 $\pi$-GC if both $\mathcal{G}_l$ and $\mathcal{G}_u$ are strong $\pi$-GC. So it suffices to show strong $\pi$-GC for $\mathcal{G}_l$ and $\mathcal{G}_u$. In the following analysis, we only show $\pi$-GC for $\mathcal{G}_u$, because the $\mathcal{G}_l$ case can be shown by the same argument.

In order to prove strong $\pi$-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 $\pi_X$-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 $\epsilon$ 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 $\pi$-GC.

It remains to show that $\mathcal{G}_u^M$ is strong $\pi$-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 $\pi$-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 $\text{sign}_\epsilon^l$ and $\text{sign}_\epsilon^u$ 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 $\pi$-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 $\epsilon$ 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 $\phi_1$ 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 & 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 & 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.