DISTRIBUTIONAL REGRESSION FOR ARBITRARY RESPONSE VARIABLES USING RELU NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper develops a unified framework for distributional regression with arbitrary response variables. Rather than relying on scalar thresholds or Euclidean structure, we model conditional event probabilities indexed by measurable sets, which extends distributional regression to general response spaces while retaining a direct connection to classical event-based formulations. Within this framework, we study two complementary estimators: a preliminary estimator based on separate regression of event indicators, and a constrained estimator that enforces inclusion-consistent predictions through a directed acyclic graph (DAG) over a finite set family. We establish general error bounds for both estimators and further specialize the results to dense ReLU networks, yielding explicit convergence rates overcoming the curse of dimensionality. Simulation studies and real-data experiments further demonstrate the practical effectiveness of the proposed framework.

1 INTRODUCTION

While regression methods are widely used in statistics and machine learning, it is well recognized that the conditional mean alone often fails to capture the full relationship between a response variable and covariates (Koenker & Bassett Jr, 1978; Shaked & Shanthikumar, 2007). In contrast, distributional regression aims to directly model the conditional distribution of the response, providing a richer characterization of the underlying stochastic relationship (Kneib et al., 2023). This has led to growing interest in distributional regression across a range of applications, such as economics (Foresi & Peracchi, 1995; Rothe, 2012; Wang et al., 2023), weather forecasting (Schlosser et al., 2019), and treatment effect estimation in clinical trials (Heller et al., 2022).

However, most existing approaches to distributional regression focus on real valued scalar (Hothorn et al., 2014; Padilla et al., 2025) or Euclidean-valued responses (Cevid et al., 2022; Klein et al., 2022; Kock & Klein, 2025), where the conditional distribution can be naturally described via cumulative distribution functions, quantiles, or threshold event probabilities. When the response variable is a more complex object, such as a network, a functional object, or more general structured data, it becomes much less clear how to define and estimate a conditional distribution in a unified and tractable way. In this paper, we consider independent observations $\{(x_i, y_i)\}_{i=1}^n$ from $(X, Y) \in [0,1]^d \times \mathcal{Y}$, and an arbitrary collection $\Lambda \subseteq \mathscr{Y}$ of measurable sets in the response space $(\mathcal{Y}, \mathscr{Y})$. Rather than modeling the conditional distribution of Y directly, we study the family of conditional event probabilities

$$F_A^*(x) = P(Y \in A \mid X = x), \qquad A \in \Lambda. \tag{1}$$

This formulation is intrinsic to $\mathcal{Y}$ and places no Euclidean structure on the response.

We propose two complementary estimators within this framework. As a natural baseline approach considered in Padilla et al. (2025), the *preliminary estimator* regresses the event indicators $\mathbb{1}_{\{y_i \in A_j\}}$ on x_i separately for each set. In our setting, this yields a simple and broadly applicable baseline for distributional regression with arbitrary response spaces, and allows us to study whether uniform statistical guarantees can be obtained over a possibly infinite family of events.

The formulation in equation 1 also induces structural constraints: whenever $A \subset A'$, we must have $F_A^*(x) \leq F_{A'}^*(x)$ for all $x \in [0,1]^d$ (a standard setting in nonparametric theory; see Györfi et al. (2002)). This monotonicity is also desirable for interpretable probability estimates. To incorporate it, we introduce a *constrained estimator*. We encode the inclusion relations among a finite subfamily

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

$\Lambda_m = \{A_j\}_{j=1}^m \subseteq \Lambda$ through a DAG and construct a joint estimator that enforces inclusion-consistent predictions by design. For these two estimators, general risk bounds are established, accommodating a variety of nonparametric estimators. Finally, motivated by the success of deep ReLU networks and their proven ability to overcome the curse of dimensionality as nonparametric estimators (Schmidt-Hieber, 2020; Kohler & Langer, 2021; Padilla et al., 2022; 2024; Ren et al., 2026), we further derive explicit convergence rates under such instantiations as an application of our general theory.

1.1 Summary of results

Framework and estimators. We develop a unified framework for distributional regression over arbitrary response spaces, where the target is a family of conditional event probabilities $F_A^*(x) = P(Y \in A \mid X = x)$, indexed by a collection of measurable sets. Within this framework, we study two complementary estimators: the preliminary estimator provides a general baseline for arbitrary-response distributional regression, accommodating collections of sets of infinite cardinality, whereas the constrained estimator delivers structurally consistent joint estimation over a finite set family.

General theory. We establish a general theory for both estimators, accommodating a broad class of nonparametric estimators under mild complexity conditions on the estimators. For the preliminary estimator, as long as the set family is totally ordered under inclusion, Theorem 1 provides uniform empirical error bounds over the set family. Corollary 1 further extends Theorem 1 to the important case of multivariate Euclidean responses with lower-orthant set families, even though such families need not satisfy the full nesting condition of Theorem 1. For the constrained estimator, Theorem 2 and bound 6 establish aggregated empirical squared-error guarantees when the number of sets forming the inclusion DAG is allowed to grow with the sample size n, with the dependence on the DAG topology characterized by the number of ancestors of each node and the lengths of its root-to-node paths.

Convergence rates for Dense ReLU Networks. We instantiate the general theory with dense ReLU network estimators. Under hierarchical composition assumptions in (Kohler & Langer, 2021), we derive explicit bounds for both estimators, with Corollary 2 covering the constrained estimator, building on a new DNN approximation theorem showing that the target probability functions can be well approximated through our DAG-constrained parametrization, and Corollaries 3 and 4 covering the preliminary estimator through specializations of Theorem 1 and Corollary 1, respectively. Appendix E further applies the same general theory to trend filtering (Mammen & Van De Geer, 1997; Kim et al., 2009; Tibshirani, 2014) and orthogonal series regression, illustrating broader applicability.

1.2 Related work

Distributional regression. Distributional regression aims to characterize the full conditional distribution of a response, rather than only its conditional mean. Existing approaches fall broadly into two categories. The first models the entire conditional distribution directly, typically in a parametric way through covariate dependent transformations of a reference distribution (Hothorn et al., 2014; Klein et al., 2022; Kock & Klein, 2025), with recent work also using generative neural network models (Shen & Meinshausen, 2025). The second, closer to our work, models conditional distributions through event indicators such as $\mathbb{1}_{\{Y \le t\}}$; this line originates from Foresi & Peracchi (1995) and has since been further developed in work on misspecification testing (Rothe & Wied, 2013), distributional regression with random forests (Schlosser et al., 2019; Cevid et al., 2022), and neural networks (Padilla et al., 2025). However, across these settings, the response is typically assumed to be scalar or Euclidean vector valued, and the target events are usually defined through threshold based sets in the response space. In contrast, our framework treats distributional regression over arbitrary response spaces via set indexed conditional event probabilities, allowing for general measurable sets.

Structured estimation under order constraints. Our constrained estimator is related to recent work on estimation under structural or order constraints. In particular, Mösching & Dümbgen (2020); Henzi et al. (2021) study settings in which the conditional distribution or quantiles are required to vary monotonically with the covariates, while Henzi (2023) imposes stochastic order constraints directly on a collection of distribution functions. In contrast to these works, our structural constraint is imposed on the family of measurable sets itself: we encode inclusion relations through a DAG and enforce estimated probabilities that satisfy these inclusion relations across the set family. More broadly, the importance of structural constraints also arises in the deep learning literature, where efforts have been made to incorporate structural knowledge into neural networks so that the resulting

predictors satisfy prescribed monotonic or logical relations specified from prior knowledge (Sill, 1997; Hu et al., 2016; You et al., 2017; Xu et al., 2018; Imani & White, 2018); however, these works are primarily methodological and generally do not provide statistical guarantees. Our work is related in spirit, since the measurable sets and their inclusion relations can likewise be specified abstractly from prior knowledge. Building on this perspective, we further derive rigorous risk bounds when deep ReLU networks are used as the underlying estimator in our constrained estimator, thereby extending the framework of Kohler & Langer (2021).

Regression with complex responses. A related line of work considers regression problems where the response lies in a non-Euclidean or structured space, such as Fréchet regression (Petersen & Müller, 2019; Chen & Müller, 2022; Bhattacharjee & Müller, 2023) and network regression (Calissano et al., 2022; Zhou & Müller, 2022). These approaches demonstrate that regression can be extended beyond Euclidean responses, but they typically focus on estimating conditional means or barycenters of objects. In contrast, our goal is to estimate conditional distributional information through event probabilities indexed by sets, which provides a different complementary perspective.

1.3 Notation

Write $\tau(x) = 1/(1+e^{-x})$ for the sigmoid function. For $n \in \mathbb{N}$, we write $[n] = \{1, \dots, n\}$. We use $\mathbb{N} = \{1, 2, \dots\}$ and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. For a function $g : \mathcal{X} \to \mathbb{R}$, we define $\|g\|_n = (n^{-1}\sum_{i=1}^n g(x_i)^2)^{1/2}$, and $\|g\|_\infty = \sup_{x\in\mathcal{X}} |g(x)|$. For two positive sequences a_n and b_n, let $a_n = O(b_n)$ or $a_n \lesssim b_n$ when $a_n \le C b_n$ for some constant $C > 0$ which is independent of n, and $a_n = \Theta(b_n)$ or $a_n \asymp b_n$ when $a_n = O(b_n)$ and $b_n = O(a_n)$. For a metric space $(\mathcal{X}, d)$, a subset $\mathcal{C} \subset \mathcal{X}$ is an r-external cover of $K \subset \mathcal{X}$ if $K \subseteq \bigcup_{x\in\mathcal{C}} B_r(x, d)$. The external covering number $N(r, K, d)$ is the minimal cardinality among such covers. For $\eta > 0$, define $B_\eta(0) = \{f \in \mathbb{R}^{\mathcal{X}} : \|f\|_n \le \eta\}$. Finally, $X_n = O_{\mathbb{P}}(a_n)$ means that for every $\epsilon > 0$ there exist C, n_0 such that $\mathbb{P}(|X_n| > C a_n) \le \epsilon$ for all $n \ge n_0$. Throughout, $F_A^*(x)$ denotes a fixed measurable version of $\mathbb{E}[\mathbf{1}_{\{Y\in A\}} \mid X = x]$. We assume that these versions are chosen so that $F_A^*(x) \le F_B^*(x)$ whenever $A \subseteq B$, for all x; this holds, for example, when the conditional law of Y given $X = x$ admits a regular conditional probability kernel. Standard measurability conditions for the empirical-process suprema and estimator selections are assumed.

1.4 Outline

The remainder of this paper is organized as follows. Section 2 introduces the model setup and the two estimators studied in this paper. Section 3 establishes the theoretical guarantees for both estimators, with Section 3.1 covering the general theory for the preliminary estimator, Section 3.2 for the constrained estimator, and Section 3.3 specializing the constrained result to dense ReLU networks. Appendix E gives theoretical results for trend filtering and orthogonal series regression. Section 4 and Appendix I report comprehensive simulation studies. Section 5 and Appendix J.2 together present two real-data examples. Finally, Section 6 concludes the paper with potential extensions.

2 Methodology

Given the setup in Section 1, we now formalize the estimation procedure by considering an arbitrary function class $\mathcal{F}$ from which the estimator is constructed.

Preliminary estimator: For each $A \in \Lambda$, the preliminary estimator is obtained by:

$$\widehat{F}_A^{\mathrm{pre}} \in \operatorname*{arg\,min}_{f\in\mathcal{F}_A} \sum_{i=1}^n (z_i^A - f(x_i))^2, \quad z_i^A = \mathbf{1}_{\{y_i\in A\}}, \quad \mathcal{F}_A \subseteq \mathcal{F}. \tag{2}$$

In finite computational applications we take $A = A_j$, $j \in [m]$, and typically $\mathcal{F}_{A_j} = \mathcal{F}$. The estimator in equation 2 serves as a general baseline for arbitrary response spaces $\mathcal{Y}$, as it relies only on set membership evaluation. However, since the estimators in equation 2 are computed separately for each j, they do not necessarily satisfy the inclusion-induced monotonicity condition $F_A^*(x) \le F_{A'}^*(x)$ for $A \subset A'$. This limitation motivates the constrained estimator introduced next.

Constrained estimator: To enforce inclusion consistency by construction, we encode the finite set family Λ_m as a directed acyclic graph. Consider $G = (V, E)$ with $V = \{1, \dots, m\}$ and

$E \subset V \times V$ such that $(j, j') \in E$ if and only if $A_j \subset A_{j'}$ and, for all $k \neq j, j'$, it does not hold that $A_j \subset A_k \subset A_{j'}$. For $j \in V$, let $\mathcal{P}(j) = \{i : (i, j) \in E\}$ denote its parents and let $\mathcal{R} = \{j : \mathcal{P}(j) = \varnothing\}$ denote the roots. Given functions $g_j : [0,1]^d \to \mathbb{R}$, $j \in V$, let $\boldsymbol{g} = (g_1, \ldots, g_m)$ denote the collection of the functions. Define the auxiliary outputs recursively by

$$\widetilde{H}_j(\boldsymbol{g})(x) = \begin{cases} g_j(x), & j \in \mathcal{R}, \\ \max\limits_{k \in \mathcal{P}(j)} \widetilde{H}_k(\boldsymbol{g})(x) + \log(1 + e^{g_j(x)}), & j \notin \mathcal{R}, \end{cases} \qquad H_j(\boldsymbol{g}) = \tau \circ \widetilde{H}_j(\boldsymbol{g}).$$

The sigmoid τ guarantees outputs in $(0, 1)$, while the softplus increment $\log(1 + e^{g_j})$ is strictly positive, so $\widetilde{H}_{j'}(\boldsymbol{g}) \geq \widetilde{H}_j(\boldsymbol{g})$ whenever there is a path from j to j' in G. Since τ is monotone, this yields $H_{j'}(\boldsymbol{g})(x) \geq H_j(\boldsymbol{g})(x)$ for all such pairs and all $x \in [0,1]^d$. With this notation in hand, for a function class $\mathcal{F}$ we define the *constrained estimator* as

$$\widehat{\boldsymbol{g}} := \operatorname*{arg\,min}_{\boldsymbol{g} \in \mathcal{F}^m} \sum_{j=1}^{m} \sum_{i=1}^{n} \big(z_i^j - H_j(\boldsymbol{g})(x_i)\big)^2, \tag{2}$$

and set

$$\widehat{F}_{A_j} := H_j(\widehat{\boldsymbol{g}}), \qquad j = 1, \ldots, m. \tag{3}$$

so the resulting estimator is inclusion-consistent by construction. In practice, the function class $\mathcal{F}$ can be chosen flexibly. For example, one may take the DNN class $\mathcal{F} = \{f \in \mathcal{F}(L, w) : \|f\|_\infty \leq B\}$, where $\mathcal{F}(L, w)$ denotes the class of ReLU neural networks with L hidden layers and w neurons per hidden layer (see Appendix A.1 for the formal definition).

3 Theory

3.1 General theory for preliminary estimator

Recall the motivation behind the preliminary estimator: we aim to obtain a *uniform* estimation error bound for $\{F_A^*\}_{A \in \Lambda}$ over a set family Λ that may be uncountable, on an arbitrary response space $\mathcal{Y}$. To make such uniform control possible, some structure on the set family is required. The following assumption imposes that Λ be totally ordered under inclusion.

Assumption 1. *Let $\Lambda \subset 2^{\mathcal{Y}}$ be a collection of measurable sets. Suppose for any $A, B \in \Lambda$ with $A \neq B$ either $A \subsetneq B$ or $B \subsetneq A$.*

Although Assumption 1 may appear strong at first sight, it captures many naturally arising set families. As a broadly applicable illustration:

Example 1. *Suppose that $(\mathcal{Y}, d_{\mathcal{Y}})$ is a metric space and let $y_0 \in \mathcal{Y}$ be a fixed element of $\mathcal{Y}$. Let*

$$B(y_0, t) := \{y \in \mathcal{Y} : d_{\mathcal{Y}}(y, y_0) \leq t\}$$

be the ball of radius t centered at y_0. Then, $\Lambda := \{B(y_0, t) : t > 0\}$ satisfies Assumption 1.

Example 1 shows that Assumption 1 allows us to define distributional regression problems on *any* metric space—including networks, function spaces, and manifolds—so that the distribution of the response's distance to a reference point y_0 can be uniformly estimated. This is precisely what our goal of distributional regression over arbitrary response spaces is meant to achieve. Under this assumption, we obtain the following result for the preliminary estimator.

Theorem 1. *Suppose Assumption 1 and the measurability conventions in Section 1.3 hold. Consider $z_i^A = \mathbb{1}_{\{y_i \in A\}}$ for $i = 1, \ldots, n$ and $A \in \Lambda$. Let $\widehat{F}_A^{pre} \in \operatorname*{arg\,min}_{f \in \mathcal{F}_A} \sum_{i=1}^n (z_i^A - f(x_i))^2$, where $\mathcal{F}_A \subset \mathcal{F}$. For each A, choose $G_A \in \mathcal{F}_A$ and put $b = \sup_{A \in \Lambda} \|G_A - F_A^*\|_\infty$. Suppose $\|F_A\|_\infty \leq U$, for all $F_A \in \mathcal{F}_A$ and $A \in \Lambda$, for some $U > 0$. For $1 < \eta < 2U\sqrt{n}$, set $J = \lceil \log_2(4nU^2/\eta^2) \rceil$,*

$$\rho_j = \frac{2^{j/2}\eta}{\sqrt{n}}, \quad \mathcal{D}(\rho) = (\mathcal{F} - \mathcal{F}) \cap B_\rho(0), \quad \mathcal{J}(\rho) = \int_0^\rho \sqrt{\log N(u, \mathcal{D}(\rho), \|\cdot\|_n)}\, du.$$

Then, conditionally on the design,

$$\mathbb{P}\left\{ \sup_{A \in \Lambda} \|\widehat{F}_A^{\mathrm{pre}} - F_A^*\|_n > \frac{\eta}{\sqrt{n}} + b \,\middle|\, \{x_i\}_{i=1}^n \right\} \lesssim \frac{\sqrt{n}}{\eta^2} \sum_{j=1}^{J} \frac{\mathcal{J}(\rho_j)}{2^j} + \frac{\sqrt{\log(2n+2)}}{\eta} + \frac{\sqrt{n} b}{\eta}. \tag{4}$$

Theorem 1 provides uniform empirical error control over the totally ordered family Λ for arbitrary estimator classes $\mathcal{F}_A$. The threshold contains the approximation error b and a term $\eta/\sqrt{n}$, while the tail bound is governed by the entropy integral $\mathcal{J}(\rho_j)$ over the shells ρ_j. For concrete classes, evaluating $\mathcal{J}(\rho_j)$ yields an explicit stochastic bound.

The supremum in Theorem 1 is taken over totally ordered set families. In many applications, such as $Y \in \mathbb{R}^u$, one instead wishes to control the estimation error uniformly over lower-orthant rectangles, which need not be totally ordered. Nevertheless, the indicator vectors induced by lower-orthant rectangles still take at most $(n+1)^u$ distinct values, and a similar argument yields a uniform bound.

Corollary 1. *Consider the above notation of Theorem 1 where $\mathcal{Y} = \mathbb{R}^u$ for some $u \in \mathbb{N}$, and let $A(t) = \{y \in \mathbb{R}^u : y_j \le t_j \ \forall j \in \{1, \dots, u\}\}$ for $t \in \mathbb{R}^u$. Then, the conclusion of Theorem 1 holds for $\Lambda := \{A(t) : t \in \mathbb{R}^u\}$, with $\sqrt{\log(2n+2)}$ replaced by $\sqrt{\log 2 + u \log(n+1)}$.*

3.2 General theory for constrained estimator

While the preliminary estimator enjoys broad generality with uniform convergence over infinite set families on arbitrary response spaces, it does not guarantee inclusion consistency: for $A \subset A'$, one may have $\widehat{F}^{\mathrm{pre}}_A(x) > \widehat{F}^{\mathrm{pre}}_{A'}(x)$. For a collection of sets encoded by a DAG, the constrained estimator in Section 2 enforces this monotonicity and estimates all set probabilities jointly. We next establish its statistical guarantees.

Theorem 2. *For a node j, let $\mathcal{A}(j)$ denote its set of ancestors, let $\overline{\mathcal{A}}(j) = \mathcal{A}(j) \cup \{j\}$, and let $\mathcal{PR}(j)$ denote the collection of paths from a root to j. Define $a_j = |\overline{\mathcal{A}}(j)|$ and $\kappa_j = \max_{p \in \mathcal{PR}(j)} |p|$, where a_j is the number of nodes that may affect the jth output and κ_j is the largest number of vertices on a root-to-j path. In particular, a root has $a_j = \kappa_j = 1$. Further, write $\bar{a}_m = m^{-1} \sum_{j=1}^m a_j$, $a_{\max,m} = \max_{j \le m} a_j$, and $\kappa_m = \max_{j \le m} \kappa_j$. For an estimator function class $\mathcal{F}$ and $\widetilde{\boldsymbol{g}} = (\widetilde{g}_1, \dots, \widetilde{g}_m) \in \mathcal{F}^m$, define $\Delta_{n,m}(\widetilde{\boldsymbol{g}}) = \max_{j \le m} \|F^*_{A_j} - H_j(\widetilde{\boldsymbol{g}})\|_\infty$. Also let $\mathcal{I}_{j,n}(\widetilde{\boldsymbol{g}}) = \{H_j(\boldsymbol{g}) - H_j(\widetilde{\boldsymbol{g}}) : \boldsymbol{g} \in \mathcal{F}^m\}$. For any η satisfying $nm > \eta^2 > C_1 nm \{\Delta_{n,m}(\widetilde{\boldsymbol{g}})\}^2$, let $K = \lceil \log(nm/\eta^2)/\log 2 \rceil$. Then,*

$$\mathbb{P}\left\{ \sum_{j=1}^m \sum_{i=1}^n \big(F^*_{A_j}(x_i) - \widehat{F}_{A_j}(x_i)\big)^2 > \eta^2 \right\} \le C_2 \sum_{j=1}^m \sum_{k=1}^K \frac{1}{2^{k-2}\eta^2} \left[1 + \sqrt{n} \int_{1/n}^{C_3 2^{k/2} \eta/\sqrt{n}} \sqrt{\log N(\epsilon, \mathcal{I}_{j,n}(\widetilde{\boldsymbol{g}}), \|\cdot\|_n)}\, d\epsilon \right], \tag{5}$$

where C_1, C_2, C_3 are positive universal constants.

Theorem 2 provides a general tail bound for the aggregated MSE of the constrained estimator. The quantity $\Delta_{n,m}(\widetilde{\boldsymbol{g}})$ measures how well the target probability vector can be approximated by the DAG outputs given a candidate $\widetilde{\boldsymbol{g}} \in \mathcal{F}^m$, while the covering entropy of $\mathcal{I}_{j,n}(\widetilde{\boldsymbol{g}})$ controls the corresponding stochastic estimation error. Thus, the theorem applies to any base function class $\mathcal{F}$ for which these approximation and entropy terms can be controlled.

A separate, high-probability oracle inequality follows from the least-squares basic inequality and localized multiplier concentration see Lemma 3). Suppose that $\mathcal{F}$ is uniformly bounded by B_n and satisfies $\log N(\epsilon, \mathcal{F}, \|\cdot\|_n) \le v_n \log(A_n/\epsilon)$ for $0 < \epsilon \le B_n$, where $v_n \ge 1$ is a complexity quantity and $A_n \ge eB_n$. Define the approximation error $\Delta_{n,m} := \inf_{\widetilde{\boldsymbol{g}} \in \mathcal{F}^m} \Delta_{n,m}(\widetilde{\boldsymbol{g}})$. Then, for every $\delta \in (0, 1/2)$, with probability at least $1 - \delta$,

$$\frac{1}{m} \sum_{j=1}^m \big\| \widehat{F}_{A_j} - F^*_{A_j} \big\|_n^2 \lesssim \Delta^2_{n,m} + \frac{\bar{a}_m v_n}{n} \log\left(\frac{e A_n nm \sqrt{\kappa_m a_{\max,m}}}{\delta} \right). \tag{6}$$

The bound separates the direct joint approximation error from the estimation error. Specifically, the estimation error depends on the DAG structure, rather than on the number of sets m alone, primarily through the average ancestor size $\bar{a}_m$, while κ_m enters only logarithmically. Thus, DAGs with the same m can exhibit different rates depending on their structure. Appendix F illustrates this dependence for several representative DAG structures.

3.3 ReLU NEURAL NETWORKS THEORY FOR CONSTRAINED ESTIMATOR

The general theory in Section 3.2 applies to a neural network estimator once its approximation error and covering entropy are specified. In this subsection, we apply that result to the hierarchical composition class $\mathcal{H}(l, \mathcal{P})$ studied by Kohler & Langer (2021). These functions are obtained by composing multiple smooth functions, each of which may depend on an input of dimension smaller than d. Its formal definition is given in Appendix A.2.

Corollary 2. *Suppose $F^*_{A_j} \in \mathcal{H}(l_j, \mathcal{P}_j)$ for $j = 1, \ldots, m$, where $l_j \in \mathbb{N}$ and $\mathcal{P}_j \subset [1, \infty) \times \mathbb{N}$. In addition, suppose each function μ in the definition of $F^*_{A_j}$ can have different smoothness $p_\mu = q_\mu + u_\mu$, for $q_\mu \in \mathbb{N}_0$ and $u_\mu \in (0, 1]$, and possibly different input dimension d_μ, so that $(p_\mu, d_\mu) \in \mathcal{P}_j$. Let $d^j_{\max}$ be the largest input dimension and $p^j_{\max}$ the largest smoothness of any function μ associated with g^*_j. We require that all partial derivatives of order less than or equal to q_μ are uniformly bounded by a constant that does not depend on j, and that each function μ is Lipschitz continuous with Lipschitz constant $C_{\mathrm{Lip}} \geq 1$. Also assume that $\max_{j \leq m} \max\{p^j_{\max}, d^j_{\max}\} = O(1)$. Define*

$$\phi_n = \max_{j \leq m} \max_{(p,d) \in \mathcal{P}_j} n^{-2p/(2p+d)}.$$

For (L_n, w_n, R_n) and the complexity quantity v_n in Corollary 6, take $\mathcal{F}_n = \mathcal{F}^{\mathrm{cl}}(L_n, w_n, R_n)$, which is the clipped DNN class in Definition 4. Then the constrained estimator satisfies

$$\frac{1}{m} \sum_{j=1}^{m} \left\| \widehat{F}_{A_j} - F^*_{A_j} \right\|_n^2 = O_{\mathbb{P}} \left[\phi_n + \frac{\bar{a}_m v_n}{n} \log \left(C R_n n m \sqrt{\kappa_m a_{\max,m}} \right) \right].$$

If m grows no faster than a polynomial in n, then the rate simplifies to $O_{\mathbb{P}}\left(\phi_n + \frac{\bar{a}_m v_n \log n}{n}\right)$. Finally, for a fixed DAG, this further reduces to $O_{\mathbb{P}}(\phi_n \log^4 n)$.

Corollary 2 provides an explicit convergence rate for the constrained DNN estimator over inclusion DAGs of general topology and growing size. Here v_n measures the complexity of the clipped DNN class $\mathcal{F}^{\mathrm{cl}}(L_n, w_n, R_n)$ and is controlled by its depth, width, and number of trainable parameters; see Corollary 6 for its explicit form. For a fixed DAG, the rate reduces to ϕ_n up to logarithmic factors, matching the standard DNN estimation rate for the hierarchical composition class (Kohler & Langer, 2021). A key ingredient in proving Corollary 2 is the new DNN approximation result in Appendix D, which shows that dense ReLU networks composed with calibration ReLU networks approximating the logit and inverse-softplus functions can still be constructed and represented as dense ReLU networks such that $\max_{j \leq m} \left\| F^*_{A_j} - H_j(g_1, \ldots, g_m) \right\|_\infty^2 \lesssim \phi_n$. We refer the reader to Corollary 6 for the detailed statement of Corollary 2, and to Appendix D for the complete DNN approximation theory.

3.4 ReLU NEURAL NETWORKS THEORY FOR PRELIMINARY ESTIMATOR

Analogous instantiations of Theorem 1 and Corollary 1 yield explicit convergence rates for the preliminary estimator, defined in Theorem 1, when DNNs are used under hierarchical composition.

Corollary 3. *Consider the notation and assumptions from Theorem 1, and that for all $A \in \Lambda$ it holds that $F^*_A \in \mathcal{H}(l_A, \mathcal{P}_A)$ satisfies conditions analogous to those imposed in Corollary 2. Also, assume that $\sup_{A \in \Lambda} \max\{p^A_{\max}, d^A_{\max}\} = O(1)$. Let $\phi_n = \sup_{A \in \Lambda} \max_{(p,d) \in \mathcal{P}_A} n^{\frac{-2p}{(2p+d)}}$. If L and w are appropriately chosen (see Corollary 5 for a formal statement), then, taking the deep ReLU networks class $\mathcal{F} = \{f \in \mathcal{F}(L, w) : \|f\|_\infty \leq B\}$ for a sufficiently large constant $B = O(1)$, we have*

$$\sup_{A \in \Lambda} \frac{1}{n} \sum_{i=1}^{n} (F^*_A(x_i) - \widehat{F}^{pre}_A(x_i))^2 = O_{\mathbb{P}}\left(\phi_n \log^4 n\right).$$

As a direct application of Theorem 1, Corollary 3 extends the advantage of DNNs to the preliminary estimator: under the hierarchical composition assumption, it achieves the uniform convergence rate $\phi_n \log^4 n$ over the totally ordered family Λ, which may contain uncountably many sets.

Finally, returning to the Euclidean response setting with rectangles, we further instantiate Corollary 1:

Corollary 4. *Suppose that the conditions and notations of Corollary 3 hold and that $\mathcal{Y} = \mathbb{R}^u$. Then letting Λ be the set of rectangles as in Corollary 1, we have that*

$$\sup_{A\in\Lambda} \frac{1}{n}\sum_{i=1}^{n}(F_A^*(x_i) - \widehat{F}_A^{pre}(x_i))^2 \;=\; O_{\mathbb{P}}\left(\frac{u\log n}{n} \;+\; \phi_n \log^4 n\right).$$

Corollary 4 shows that in the multivariate Euclidean setting, the preliminary estimator satisfies a two-term uniform error bound over the lower-orthant family. The term $u\log n/n$ reflects the cost of controlling the estimator uniformly over a u-dimensional lower-orthant class, while $\phi_n \log^4 n$ is the rate induced by the dense ReLU approximation. Hence, the overall rate is determined by the slower of these two terms. In particular, when the response dimension u is fixed, the term $\phi_n \log^4 n$ continues to govern the main convergence behavior, reflecting the advantage of DNNs to overcome the curse of dimensionality. More generally, the same conclusion remains valid even when u is allowed to grow with n, as long as the dimension-dependent term $u\log n/n$ is dominated by $\phi_n \log^4 n$.

Taken together, Corollaries 2, 3, and 4 demonstrate that our general theory accommodates DNNs in distributional regression with arbitrary response variables, yielding explicit rates for both estimators under hierarchical composition of the target probabilities. We provide a comparison with other recent theoretical results in Table 4.

4 Numerical experiments

We consider eight simulation scenarios with various response types (Euclidean, manifold-valued, functional, and network responses) and different DAG structures, along with one special scenario based on real data with a simplified toy task. DNNs are used as estimators: the preliminary estimator (**Pre-NN**), which trains a separate DNN for each set independently, and the constrained estimator (**Constrain-NN**), which jointly trains all DNNs using a loss function induced by the DAG describing the sets. We also compare with a naive baseline that adds constraints to Pre-NN post hoc, where we clip Pre-NN's outputs (**Pre-NN-Clip**) according to the DAG such that if a child's predicted probability is smaller than some of its parents', we raise it to the maximum among its parents to satisfy monotonicity. As additional baselines, Distributional Random Forests (**DRF**, Cevid et al. (2022)) perform distributional regression over the sets via random forests with inclusion-consistent outputs, and **Engression** (Shen & Meinshausen, 2025) fits a generative DNN of the conditional distribution and estimates set probabilities by averaging the set indicator over sampled responses.

The sets for each scenario are predefined with a specific inclusion structure. Except for the special scenario, we generate $n_{\text{train}} \in \{1000, 2000, 4000\}$ training samples per scenario, and are repeated over 30 replications. The performance is evaluated using aggregated MSE across all sets on a held-out test set of size $n_{\text{test}} = 10000$, averaged over rounds. The true conditional probabilities $\mathbb{P}(Y \in A_j \mid X = x)$ serve as the estimation target and are computed via Monte Carlo integration with 10000 samples per data point at test time. Due to space constraints, we briefly outline the two challenging scenarios (S1, S2) and the special scenario (S3) here, and refer to Appendix I for the full details of all scenarios and the implementation settings for all methods.

S1 (Swiss roll manifold). We consider a response on a nonlinear manifold. Let $X \sim \text{Unif}([0,2]^{50})$ and generate $X_{\text{noisy}} = X + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2(X) I_{50})$ with $\sigma^2(X) = 1 + 0.25\|X\|^2/50$. We map X_{noisy} to latent coordinates $t = 1.5\pi + 3\pi\tau(w_t^\top X_{\text{noisy}} + b_t)$ and $h = \tau(w_h^\top X_{\text{noisy}} + b_h)$, where $\tau(u) = (1+e^{-u})^{-1}$, the entries of $w_t, w_h \in \mathbb{R}^{50}$ are drawn i.i.d. from $\mathcal{N}(0, 0.25)$, and $b_t, b_h \sim \mathcal{N}(0, 0.01)$. Thus, $t \in [1.5\pi, 4.5\pi]$ and $h \in [0,1]$, and we set our response $Y = (t\cos t, h, t\sin t)$. We construct $m = 20$ sets using nested marginal quantile levels of the latent coordinates t and h, forming a 5×4 inclusion grid; see Appendix I for design details.

S2 (Random dot product graph). To illustrate that our framework applies to arbitrary response spaces, we let $Y \in \{0,1\}^{N\times N}$ be the adjacency matrix with $N = 50$ nodes generated by a random dot product graph (RDPG, Athreya et al. (2018)), where each node i carries a unit-length embedding vector $z_i \in \mathbb{S}^{d-1}$ with $d = 5$, and the probability of an edge between nodes i and j is $\frac{1}{2}(\langle z_i, z_j\rangle + 1)$. The RDPG is endowed with a community structure of $K = 3$ communities, where both the community assignment probabilities $\pi_k(X)$ and the unit-length community embedding centers $a_k(X) \in \mathbb{S}^{d-1}$ are non-linearly parametrized by the covariate $X \sim \text{Unif}([0,2]^{50})$ for $k \in \{1,2,3\}$. Each node i is

Table 1: Aggregated MSE ($\times 10^2$) for S1 and S2 averaged over 30 trials. Best per column in bold.

	S1 (Swiss roll manifold)			S2 (RDPG)		
Method	n=1000	n=2000	n=4000	n=1000	n=2000	n=4000
Pre-NN	3.07 (0.15)	1.91 (0.10)	1.04 (0.05)	4.64 (0.10)	3.72 (0.07)	3.09 (0.06)
Constrain-NN	**0.96 (0.07)**	**0.72 (0.04)**	**0.57 (0.04)**	**3.48 (0.11)**	**3.05 (0.07)**	**2.81 (0.06)**
Pre-NN-Clip	2.49 (0.16)	1.56 (0.11)	0.88 (0.05)	4.39 (0.10)	3.55 (0.07)	2.99 (0.06)
DRF	1.28 (0.13)	1.22 (0.12)	1.16 (0.11)	7.13 (1.46)	6.94 (1.43)	6.30 (1.21)
Engression	4.93 (2.55)	4.66 (2.44)	4.36 (2.39)	4.20 (1.35)	3.95 (1.34)	3.83 (1.32)

Table 2: Aggregated MSE ($\times 10^2$) for S3 across STL-10 classes, averaged over 30 images per class. The "Overall" reports the average across all classes. See full results with standard deviations in Table 6.

Method / Category	Overall	Airplane	Bird	Car	Cat	Deer	Dog	Horse	Monkey	Ship	Truck
Pre-NN	3.63	2.68	3.75	3.70	3.70	4.17	3.72	4.50	3.42	2.83	3.79
Constrain-NN	**3.37**	**2.62**	**3.52**	**3.45**	**3.50**	**3.99**	**3.56**	**4.10**	**3.18**	**2.35**	**3.38**
Pre-NN-Clip	3.61	2.66	3.73	3.69	3.69	4.15	3.71	4.49	3.41	2.86	3.76
DRF	4.26	3.26	4.68	4.24	4.41	4.87	4.55	5.37	4.34	2.73	4.14
Engression	7.96	6.09	7.91	8.91	7.47	7.97	7.92	9.23	8.12	7.08	8.92

then independently assigned $c_i \sim \mathrm{Cat}(\pi(X))$, with embedding $z_i = (a_{c_i}(X) + \varepsilon_i)/\|a_{c_i}(X) + \varepsilon_i\|$ where $\varepsilon_i \sim \mathcal{N}(0, 0.3^2 I_d)$, yielding the random graph Y. We estimate the distribution of the network statistic transitivity $T(Y)$ (Graham, 2015; 2016), a standard measure of global clustering in network analysis, with $m = 10$ nested sets $A_j = \{Y : T(Y) \le q_{\alpha_j}\}$ over 10 equally spaced quantile levels α_j from the 5th to the 95th percentile of the empirical marginal distribution of $T(Y)$.

S3 (Pixel density estimation on single noisy natural image). We consider a setting where $X \in [0,1]^2$ denotes a pixel coordinate in an image, and the Euclidean response $Y = (Y_R, Y_G, Y_B) \in [0,1]^3$ is a noisy observation of the underlying clean pixel value $J(X) = (J_R, J_G, J_B)$ at coordinate X. Our goal is to estimate the conditional distribution of $Y \mid X$ for a single given noisy image. The noisy response is generated by Poisson-Gaussian noise model (Foi et al., 2008) with $Y_c = \mathrm{Poisson}(50 \cdot J_c(X))/50 + \varepsilon_c$ and $\varepsilon_c \sim \mathcal{N}(0, 0.05^2)$ i.i.d. across channels $c \in \{R, G, B\}$. The target sets are 13 hyperrectangles defined by marginal quantiles of Y. To ground the simulation in realistic images, we use the STL-10 dataset (Coates et al., 2011) as the source of clean images J, which contains 96×96 color photographs across 10 object categories (see Figure I.1.3 for illustration). For each image, we run an independent experiment: pixels on a stride-2 grid form the training set (2304 pixels) and the remaining 6912 pixels form the test set, preserving the spatial distribution of pixels across splits. Results are averaged over 30 images randomly selected per class.

Tables 1 and 2 report the performance across these simulation scenarios.[1] Specifically, **Constrain-NN** outperforms the competing methods in most settings, while **Pre-NN** also improves as n increases. This pattern is further confirmed by Table 5 with the additional scenarios reported in Appendix I, supporting the effectiveness of our proposed estimators.

5 Real data experiments

We conduct two real-data experiments: the Chicago crime dataset analyzed in Tansey et al. (2018); Padilla et al. (2025) and the Beijing Multi-Site Air Quality dataset (Zhang et al., 2017). In this section, we present partial results for the more complicated Beijing air quality dataset and refer to Appendix J for the complete results across both experiments.

Experiments on the Beijing Multi-Site Air Quality dataset. We evaluate on a multivariate distributional air quality forecasting task using the Beijing Multi-Site Air Quality dataset, which records hourly measurements of six pollutants (PM2.5, PM10, SO_2, NO_2, CO, O_3) and six meteorological covariates (temperature, pressure, etc.) at 12 monitoring stations in Beijing from March 2013 to February 2017. Under international air quality standards, the overall Air Quality Index (AQI) is

[1] We remark that for S2, **Engression** cannot learn a direct generative model of the high-dimensional RDPG response within a reasonable time; we therefore adapt it to learn the network statistics directly.

Table 3: MSE ($\times 10^2$) across 76 sets and 12 stations. Each cell reports results for **Pre-NN / Constrain-NN / Pre-NN-Clip / DRF / Engression**. The best performance is bolded. Full results with standard deviations are reported in Table 10.

	$W = 3$	$W = 6$	$W = 9$
$h = 3$	3.43 / **2.28** / 3.54 / 2.29 / 3.79	3.41 / **2.24** / 3.51 / 2.33 / 3.94	3.35 / **2.23** / 3.46 / 2.36 / 3.92
$h = 6$	3.78 / **2.84** / 3.88 / 2.90 / 4.12	3.71 / **2.81** / 3.79 / 2.91 / 4.24	3.65 / **2.79** / 3.73 / 2.92 / 4.18
$h = 9$	3.97 / **3.15** / 4.06 / 3.21 / 4.31	3.87 / **3.10** / 3.95 / 3.20 / 4.40	3.79 / **3.09** / 3.87 / 3.17 / 4.30

determined by the single pollutant with the highest individual pollution grade. Consequently, many forecasting work focuses on predicting a scalar AQI directly (Janarthanan et al., 2021; Sarkar et al., 2022). However, the same overall AQI grade can arise from very different pollutant profiles; an AQI of "moderate pollution" may be driven by elevated PM2.5 alone, by NO_2 alone, or by multiple pollutants exceeding thresholds simultaneously. Understanding which pollutants are likely to be responsible for the pollution, is essential for targeted public health interventions. We therefore reformulate the task: let $Y = (Y_1, \ldots, Y_p)$ denote the levels of p pollutants, and let $q_{j,\ell}$ denote the threshold for pollutant j at pollution grade ℓ, chosen with reference to Chinese air quality standards. We predict the joint event $\{Y_1 \geq q_{1,\ell_1}, \ldots, Y_p \geq q_{p,\ell_p}\}$, in which all pollutants exceed their respective thresholds q_{j,ℓ_j} for some specified grades $(\ell_1, \ldots, \ell_p)$. After preprocessing, this yields 76 sets whose inclusion relation induces a DAG. The prediction is conducted independently at each of the 12 monitoring stations; for each station, the covariate vector x_s at time s consists of time-cycle encodings and lagged observations of all meteorological and pollutant variables over a window of W consecutive hours ending h hours before time s. To systematically evaluate the forecasting performance, we vary the forecast horizon $h \in \{3, 6, 9\}$ and the input window length $W \in \{3, 6, 9\}$, yielding 9 experimental configurations with different difficulties. All models are trained independently per station and tested on a three-month holdout each year, with consecutive observations within each held-out month, yielding 75/25 train/test split overall. We refer the reader to Appendix J.1 for a detailed configuration, including the procedure used to define the 76 sets, the input variables, and the configuration of all estimators. More general guidelines for constructing set families are provided in Appendix J.3.

Table 3 reports the MSE averaged first over the 76 sets within each station, then across all 12 stations. The **Constrain-NN** achieves the best performance in almost all configurations. Several findings emerge. First, longer forecast horizons h yield consistently higher MSE, as prediction difficulty grows with the temporal gap between the available observations and the target time. Second, for a fixed h, the MSE of the proposed DNN-based estimators decreases as the input window length W increases, consistent with the intuition that richer historical context aids prediction; **DRF**, however, does not benefit consistently from this, potentially due to the curse of dimensionality. Third, **Constrain-NN** outperforms **Pre-NN**, which in turn outperforms **Pre-NN-Clip**, suggesting that incorporating the DAG structure improves accuracy while post-hoc clipping may substantially degrade performance in practice. These findings also hold at the per-station level (see Table 11 in Appendix J.1).

6 Conclusion

This paper develops a unified framework for distributional regression with arbitrary response variables with two complementary estimators. The preliminary estimator provides a simple baseline based on separate regression of event indicators and yields uniform empirical error bounds, accommodating set families with infinitely many sets. The constrained estimator incorporates inclusion relations for the sets through a DAG and enforces inclusion-consistent probability predictions, for which we derive upper bounds on the averaged MSE that allow the number of sets to grow with the sample size. We further instantiate the general theory with DNNs and other estimators, obtaining explicit rates. Several directions remain for future work. In particular, we do not derive lower bounds for DNNs, as such bounds for the hierarchical function classes are available only in specific cases; see Remark 2 of Kohler & Langer (2021). In addition, the framework requires the target sets to be specified in advance. Appendix J.3 provides practical guidelines for their construction, while identifying scientifically meaningful sets may itself be nontrivial in some applications and constitutes an interesting future direction.

AI USE STATEMENT

In this work, we used generative AI tools to clean and reformat datasets, assist with translation, implement methods, formulate mathematical claims, and assist in the writing of proofs. Additionally, we used generative AI tools for language polishing and literature retrieval and discovery. We did not use generative AI tools to generate synthetic datasets, interpret results, design or provide feedback on the research methodology or experiments, propose or refine hypotheses, help develop theoretical models or conceptual frameworks, or provide critical ingredients for proving mathematical claims. Support for qualitative and thematic data analysis is not applicable to this work.

Specifically, the manuscript was initially written by the authors. Generative AI tools were subsequently used to assist with improving the clarity and presentation of the writing. The authors then reviewed, revised, and rephrased all AI-assisted text through multiple rounds of human review and made the final decisions regarding the content and wording of the manuscript.

For the literature review, generative AI tools were used for broad retrieval and discovery of potentially related work. The authors independently examined the relevant papers, determined their relevance, identified the pertinent information, and wrote the literature review.

For the theoretical development, the mathematical claims and proof strategies were developed by the authors. For the DNN approximation theory underlying the theoretical guarantees for the constrained estimator, primarily in Appendix D, the authors first developed detailed proof roadmaps and then used generative AI tools to assist with drafting initial proofs based on those roadmaps. The authors subsequently checked these proofs for correctness, revised and rephrased the arguments, and made the final decisions regarding their mathematical content and presentation. For all other theoretical results, the initial proofs were written entirely by the authors. Generative AI tools were subsequently used only to assist with verification, raise potential issues, and check subsequent corrections, with all revisions and final decisions made by the authors.

We have reviewed all AI-assisted work and take responsibility for the final content of this work, including all text and mathematical arguments produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the assumptions and complete proofs for all theoretical results in the appendices. The appendices also provide detailed specifications for the numerical experiments and real-data analyses, including data-generating mechanisms, set constructions, data processing, model configurations, and training settings. Computational resources are reported separately in the appendix. The implementation code will be made publicly available upon acceptance.

REFERENCES

Avanti Athreya, Donniell E Fishkind, Minh Tang, Carey E Priebe, Youngser Park, Joshua T Vogelstein, Keith Levin, Vince Lyzinski, Yichen Qin, and Daniel L Sussman. Statistical inference on random dot product graphs: a survey. *Journal of Machine Learning Research*, 18(226):1–92, 2018.

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

Satarupa Bhattacharjee and Hans-Georg Müller. Single index fréchet regression. *The Annals of Statistics*, 51(4):1770–1798, 2023.

Anna Calissano, Aasa Feragen, and Simone Vantini. Graph-valued regression: Prediction of unlabelled networks in a non-euclidean graph space. *Journal of Multivariate Analysis*, 190:104950, 2022.

Domagoj Cevid, Loris Michel, Jeffrey Näf, Peter Bühlmann, and Nicolai Meinshausen. Distributional random forests: Heterogeneity adjustment and multivariate distributional regression. *Journal of Machine Learning Research*, 23(333):1–79, 2022.

Yaqing Chen and Hans-Georg Müller. Uniform convergence of local fréchet regression with applications to locating extrema and time warping for metric space valued trajectories. *The Annals of Statistics*, 50(3):1573–1592, 2022.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Alessandro Foi, Mejdi Trimeche, Vladimir Katkovnik, and Karen Egiazarian. Practical poissonian-gaussian noise modeling and fitting for single-image raw-data. *IEEE transactions on image processing*, 17(10):1737–1754, 2008.

Silverio Foresi and Franco Peracchi. The conditional distribution of excess returns: An empirical analysis. *Journal of the American Statistical Association*, 90(430):451–466, 1995.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Bryan S Graham. Methods of identification in social networks. *Annu. Rev. Econ.*, 7(1):465–485, 2015.

Bryan S Graham. Homophily and transitivity in dynamic network formation. Technical report, National Bureau of Economic Research, 2016.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer, 2002.

Gillian Z Heller, Kristy P Robledo, and Ian C Marschner. Distributional regression in clinical trials: treatment effects on parameters other than the mean. *BMC Medical Research Methodology*, 22(1): 56, 2022.

Alexander Henzi. Consistent estimation of distribution functions under increasing concave and convex stochastic ordering. *Journal of Business & Economic Statistics*, 41(4):1203–1214, 2023.

Alexander Henzi, Johanna F Ziegel, and Tilmann Gneiting. Isotonic distributional regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):963–993, 2021.

Torsten Hothorn, Thomas Kneib, and Peter Bühlmann. Conditional transformation models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):3–27, 2014.

Zhiting Hu, Xuezhe Ma, Zhengzhong Liu, Eduard Hovy, and Eric Xing. Harnessing deep neural networks with logic rules. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2410–2420, 2016.

Ehsan Imani and Martha White. Improving regression performance with distributional losses. In *International conference on machine learning*, pp. 2157–2166. PMLR, 2018.

Ramadoss Janarthanan, Pachaivannan Partheeban, K Somasundaram, and P Navin Elamparithi. A deep learning approach for prediction of air quality index in a metropolitan city. *Sustainable Cities and Society*, 67:102720, 2021.

Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. ℓ_1 trend filtering. *SIAM review*, 51(2):339–360, 2009.

Nadja Klein, Torsten Hothorn, Luisa Barbanti, and Thomas Kneib. Multivariate conditional transformation models. *Scandinavian Journal of Statistics*, 49(1):116–142, 2022.

Thomas Kneib, Alexander Silbersdorff, and Benjamin Säfken. Rage against the mean–a review of distributional regression approaches. *Econometrics and Statistics*, 26:99–123, 2023.

Lucas Kock and Nadja Klein. Truly multivariate structured additive distributional regression. *Journal of Computational and Graphical Statistics*, 34(4):1189–1201, 2025.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pp. 33–50, 1978.

Michael Kohler and Sophie Langer. On the rate of convergence of fully connected deep neural network regression estimates. *The Annals of Statistics*, 49(4):2231–2249, 2021.

Enno Mammen and Sara Van De Geer. Locally adaptive regression splines. *The Annals of Statistics*, 25(1):387–413, 1997.

Alexandre Mösching and Lutz Dümbgen. Monotone least squares and isotonic quantiles. 2020.

Carlos Misael Madrid Padilla, Oscar Hernan Madrid Padilla, Yik Lun Kei, Zhi Zhang, and Yanzhen Chen. Confidence interval construction and conditional variance estimation with dense relu networks. *arXiv preprint arXiv:2412.20355*, 2024.

Carlos Misael Madrid Padilla, Oscar Hernan Madrid Padilla, and Sabyasachi Chatterjee. Risk bounds for distributional regression. *arXiv preprint arXiv:2505.09075*, 2025.

Oscar Hernan Madrid Padilla, Wesley Tansey, and Yanzhen Chen. Quantile regression with relu networks: Estimators and minimax rates. *Journal of Machine Learning Research*, 23(247):1–42, 2022.

Alexander Petersen and Hans-Georg Müller. Fréchet regression for random objects with euclidean predictors. 2019.

Junpeng Ren, Carlos Misael Madrid Padilla, Yanzhen Chen, and Oscar Hernan Madrid Padilla. Transfer learning in nonparametric regression with deep relu networks. *arXiv preprint arXiv:2608.20255*, 2026.

Christoph Rothe. Partial distributional policy effects. *Econometrica*, 80(5):2269–2301, 2012.

Christoph Rothe and Dominik Wied. Misspecification testing in a class of conditional distributional models. *Journal of the American Statistical Association*, 108(501):314–324, 2013.

Veeranjaneyulu Sadhanala and Ryan J Tibshirani. Additive models with trend filtering. *The Annals of Statistics*, 47(6):3032–3068, 2019a.

Veeranjaneyulu Sadhanala and Ryan J Tibshirani. Supplement to "additive models with trend filtering.", 2019b.

Nairita Sarkar, Rajan Gupta, Pankaj Kumar Keserwani, and Mahesh Chandra Govil. Air quality index prediction using an effective hybrid deep learning model. *Environmental Pollution*, 315:120404, 2022.

Lisa Schlosser, Torsten Hothorn, Reto Stauffer, and Achim Zeileis. Distributional regression forests for probabilistic precipitation forecasting in complex terrain. 2019.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. 2020.

Moshe Shaked and J George Shanthikumar. *Stochastic orders*. Springer, 2007.

Xinwei Shen and Nicolai Meinshausen. Engression: extrapolation through the lens of distributional regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(3): 653–677, 2025.

Joseph Sill. Monotonic networks. *Advances in neural information processing systems*, 10, 1997.

Wesley Tansey, Jesse Thomason, and James Scott. Maximum-variance total variation denoising for interpretable spatial smoothing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Ryan J Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. 2014.

Lloyd N Trefethen. *Approximation theory and approximation practice, extended edition*. SIAM, 2019.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, 1996. doi: 10.1007/978-1-4757-2545-2.

Yunyun Wang, Tatsushi Oka, and Dan Zhu. Bivariate distribution regression with application to insurance data. *Insurance: Mathematics and Economics*, 113:215–232, 2023.

Jingyi Xu, Zilu Zhang, Tal Friedman, Yitao Liang, and Guy Broeck. A semantic loss function for deep learning with symbolic knowledge. In *International conference on machine learning*, pp. 5502–5511. PMLR, 2018.

Seungil You, David Ding, Kevin Canini, Jan Pfeifer, and Maya Gupta. Deep lattice networks and partial monotonic functions. *Advances in neural information processing systems*, 30, 2017.

Shuyi Zhang, Bin Guo, Anlan Dong, Jing He, Ziping Xu, and Song Xi Chen. Cautionary tales on air-quality improvement in beijing. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2205), 2017.

Yidong Zhou and Hans-Georg Müller. Network regression with graph laplacians. *Journal of Machine Learning Research*, 23(320):1–41, 2022.

APPENDIX

A Neural Network Preliminaries **14**

A.1 Dense ReLU networks 14

A.2 Hierarchical composition models 14

B Proofs for the General Theory **15**

B.1 Proof of Theorem 1 15

B.2 Proof of Corollary 1 17

B.3 Proof of Theorem 2 17

B.4 Auxiliary results for Theorem 2 18

B.5 Proof of the explicit bound equation 6 21

C DNN Theory for the Preliminary Estimator **22**

C.1 Proof of Corollary 3 and 5 22

C.2 Corollary 4 24

D DNN Theory for the Constrained Estimator **24**

E Trend filtering and orthogonal series regression **48**

E.1 Trend filtering 48

E.2 Orthogonal series regression 52

F Representative DAG structures **55**

G Comparison with prior theory **55**

H Compute resources **56**

I Detailed configuration and results for numerical experiments **56**

I.1 Detailed design for the eight scenarios 56

I.2 Full experiment results 61

I.3 Configuration for estimators 62

J Real data experiments **62**

J.1 Real data: Beijing multi-Site air quality prediction 62

J.2 Real data: Chicago 2015 crime distribution estimation 67

J.3 Practical guidance for constructing a set family 67

A Neural Network Preliminaries

A.1 Dense ReLU networks

We introduce the class of deep ReLU neural networks, which serve as the base function class $\mathcal{F}$ in both estimators equation 2 and equation 2 for deriving explicit convergence rates, although our general theory applies to a broad class of nonparametric estimators. Let $\rho(x) = \max\{0, x\}$ denote the ReLU activation, applied componentwise to vectors. For integers $L \geq 1$ and $w \geq 1$, let $\mathcal{F}(L, w)$ be the class of functions

$$f = \Phi_{L+1} \circ \rho \circ \Phi_L \circ \cdots \circ \rho \circ \Phi_1,$$

where $\Phi_s(z) = \mathbf{W}_s z + \mathbf{b}_s$ is affine, $w_0 = d$, $w_1 = \cdots = w_L = w$, and $w_{L+1} = 1$; hence the network has L hidden ReLU layers of common width w and a scalar affine output layer.

Following Kohler & Langer (2021), we use the bounded dense-network class $\mathcal{F} = \{f \in \mathcal{F}(L, w) : \|f\|_\infty \leq B\}$ for a sufficiently large constant $B > 0$. For the constrained estimator, a fixed latent bound is not assumed in Corollaries 2 and 6; those results use the clipped class with the growing bound R_n.

A.2 Hierarchical composition models

Neural networks have been shown in Kohler & Langer (2021) to enjoy structural advantages over classical nonparametric methods when low-dimensional hierarchical composition is present. To this end, we impose a hierarchical composition structure on the target probability functions. Before defining the class, we recall smoothness.

Definition 1 ((p, C)-smoothness)**.** *Let $p = q + s$ for some $q \in \mathbb{N} = \mathbb{Z}^+ \cup \{0\}$ and $0 < s \leq 1$. We say that a function $g : \mathbb{R}^d \to \mathbb{R}$ is called (p, C)-smooth, if for every $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$, with $d \in \mathbb{Z}^+$, where $\sum_{j=1}^d \alpha_j = q$ the partial derivative $\partial^q g / \left(\partial a_1^{\alpha_1} \ldots \partial a_d^{\alpha_d}\right)$ exists and satisfies*

$$\left| \frac{\partial^q g}{\partial a_1^{\alpha_1} \ldots \partial a_d^{\alpha_d}}(a) - \frac{\partial^q g}{\partial a_1^{\alpha_1} \ldots \partial a_d^{\alpha_d}}(b) \right| \leq C \|a - b\|^s$$

for all $a, b \in \mathbb{R}^d$.

We are now ready to define the hierarchical composition function class.

Definition 2 (Space of hierarchical composition models (Kohler & Langer, 2021))**.** *For $l = 1$ and smoothness constraint $\mathcal{P} \subseteq (0, \infty) \times \mathbb{N}$ the space of hierarchical composition models is defined as*

$$\begin{aligned} \mathcal{H}(1, \mathcal{P}) := \big\{ & h : \mathbb{R}^d \to \mathbb{R} : h(a) = \mu\left(a_{\pi(1)}, \ldots, a_{\pi(K)}\right), \\ & \text{where } \mu : \mathbb{R}^K \to \mathbb{R} \text{ is } (p, C)\text{-smooth for some} \\ & (p, K) \in \mathcal{P} \text{ and } \pi : \{1, \ldots, K\} \to \{1, \ldots, d\} \big\}. \end{aligned}$$

For $l > 1$, we recursively define

$$\begin{aligned} \mathcal{H}(l, \mathcal{P}) := \big\{ & h : \mathbb{R}^d \to \mathbb{R} : h(a) = \mu\left(f_1(a), \ldots, f_K(a)\right), \\ & \text{where } \mu : \mathbb{R}^K \to \mathbb{R} \text{ is } (p, C)\text{-smooth for some} \\ & (p, K) \in \mathcal{P} \text{ and } \ f_i \in \mathcal{H}(l-1, \mathcal{P}) \big\}. \end{aligned}$$

B Proofs for the General Theory

B.1 Proof of Theorem 1

Proof. We proceed as in the proof of Theorem 3 in Padilla et al. (2025). Notice that for any $A \in \Lambda$ we have that

$$\|F_A - H_A\|_n^2 \;\le\; 4U^2$$

for all $F_A, H_A \in \mathcal{F}_A$. It follows that that $\sup_{A\in\Lambda} \|\widehat{F}_A^{\text{pre}} - G_A\|_n^2 \le 4U^2$. Use $J = \lceil \log_2(4nU^2/\eta^2) \rceil$ and radius $r_j = 2^{j/2}\eta/\sqrt{n}$ as in the theorem. All probabilities and expectations below are conditional on the design, with the measurability conventions of Section 1.3.

Next, to control the probability of large deviations, we focus on

$$\left\{ \sup_{A\in\Lambda} \|\widehat{F}_A^{\text{pre}} - G_A\|_n^2 > \eta^2/n \right\} \cap \left\{ \sup_{A\in\Lambda} \|\widehat{F}_A^{\text{pre}} - G_A\|_n^2 \le 4U^2 \right\},$$

which confines the deviations to the range $(\eta^2/n, 4U^2]$. Proceeding with a peeling argument, we obtain that

$$\begin{aligned}
&\mathbb{P}\left(\sup_{A\in\mathcal{A}} \|\widehat{F}_A^{\text{pre}} - G_A\|_n \;>\; \tfrac{\eta}{\sqrt{n}} \,\middle|\, \{x_i\}_{i=1}^n \right) \\
&= \mathbb{P}\left(\sup_{A\in\Lambda} \|\widehat{F}_A^{\text{pre}} - G_A\|_n^2 \;>\; \frac{\eta^2}{n},\; \sup_{A\in\Lambda} \|\widehat{F}_A^{\text{pre}} - G_A\|_n^2 \le 4U^2 \,\middle|\, \{x_i\}_{i=1}^n \right) \\
&\le \sum_{j=1}^{J} \mathbb{P}\left(\sup_{A\in\Lambda} \|\widehat{F}_A^{\text{pre}} - G_A\|_n^2 \;>\; \frac{2^{j-1}\eta^2}{n},\; \sup_{A\in\Lambda} \|\widehat{F}_A^{\text{pre}} - G_A\|_n^2 \le \frac{2^j\eta^2}{n} \,\middle|\, \{x_i\}_{i=1}^n \right) \\
&\le \sum_{j=1}^{J} \mathbb{P}\Bigg(\sup_{A\in\Lambda} 2\sum_{i=1}^{n} (G_A(x_i) - \mathbb{1}_{\{y_i\in A\}})(G_A(x_i) - \widehat{F}_A^{\text{pre}}(x_i)) \;>\; 2^{j-1}\eta^2, \\
&\qquad\qquad \sup_{A\in\Lambda} \|\widehat{F}_A - G_A\|_n^2 \le 2^j\eta^2/n \,\Bigg|\, \{x_i\}_{i=1}^n \Bigg) \\
&\le \sum_{j=1}^{J} \mathbb{P}\Bigg(\sup_{A\in\Lambda} \sup_{F_A\in\mathcal{F}_A\,:\,\|F_A - G_A\|_n^2 \le 2^j\eta^2/n} \sum_{i=1}^{n} (G_A(x_i) - \mathbb{1}_{\{y_i\in A\}})(G_A(x_i) - F_A(x_i)) \\
&\qquad > \; 2^{j-2}\eta^2 \,\Bigg|\, \{x_i\}_{i=1}^n \Bigg),
\end{aligned}$$

where the first inequality follows from the union bound and the second from the basic inequality. Indeed, expanding the least-squares inequality against G_A gives

$$n\|\widehat{F}_A^{\text{pre}} - G_A\|_n^2 \le 2\sum_{i=1}^{n} \big(G_A(x_i) - \mathbf{1}_{\{y_i\in A\}}\big)\big(G_A(x_i) - \widehat{F}_A^{\text{pre}}(x_i)\big). \tag{R4}$$

Furthermore,

$$\begin{aligned}
&\mathbb{P}\Bigg(\sup_{A\in\Lambda} \sup_{F_A\in\mathcal{F}_A\,:\,\|F_A - G_A\|_n^2 \le 2^j\eta^2/n} \sum_{i=1}^{n} (G_A(x_i) - \mathbb{1}_{\{y_i\in A\}})(G_A(x_i) - F_A(x_i)) \;>\; 2^{j-2}\eta^2 \,\Bigg|\, \{x_i\}_{i=1}^n \Bigg) \\
&\le \frac{1}{2^{j-2}\eta^2}\mathbb{E}\left(\sup_{A\in\Lambda} \sup_{F_A\in\mathcal{F}_A\,:\,\|F_A - G_A\|^2/n \le 2^j\eta^2/n} \sum_{i=1}^{n} (G_A(x_i) - \mathbb{1}_{\{y_i\in A\}})(G_A(x_i) - F_A(x_i)) \,\middle|\, \{x_i\}_{i=1}^n \right) \\
&\le \frac{1}{2^{j-2}\eta^2}\mathbb{E}\left(\sup_{A\in\Lambda} \sup_{F_A\in\mathcal{F}_A\,:\,\|F_A - G_A\|_n^2 \le 2^j\eta^2/n} \sum_{i=1}^{n} (F_A^*(x_i) - \mathbb{1}_{\{y_i\in A\}})(G_A(x_i) - F_A(x_i)) \,\middle|\, \{x_i\}_{i=1}^n \right) \\
&\quad + \frac{1}{2^{j-2}\eta^2} \sup_{A\in\Lambda} \sup_{F_A\in\mathcal{F}_A\,:\,\|F_A - G_A\|_n^2 \le 2^j\eta^2/n} \sum_{i=1}^{n} (G_A(x_i) - F_A^*(x_i))(G_A(x_i) - F_A(x_i) \,\middle|\, \{x_i\}_{i=1}^n) \\
&\le \frac{1}{2^{j-2}\eta^2}\mathbb{E}\left(\sup_{A\in\Lambda} \sup_{F_A\in\mathcal{F}_A\,:\,\|F_A - G_A\|_n^2 \le 2^j\eta^2/n} \sum_{i=1}^{n} (F_A^*(x_i) - \mathbb{1}_{\{y_i\in A\}})(G_A(x_i) - F_A(x_i)) \,\middle|\, \{x_i\}_{i=1}^n \right) \\
&\quad + \frac{4\sqrt{n}}{2^{j/2}\eta} \sup_{A\in\Lambda} \|G_A - F_A^*\|_\infty
\end{aligned} \tag{7}$$

where the first inequality follows from Markov's inequality, the last inequality holds by Cauchy–Schwarz inequality, and $C > 0$ is a constant. Moreover, for $\xi_1, \dots, \xi_n$ independent Rademacher random variables independent $\{(x_i, y_i)\}_{i=1}^n$, we have, by Symmetrization, that

$$\begin{aligned}
&\mathbb{E}\left(\sup_{A\in\Lambda}\sup_{F_A\in\mathcal{F}_A\,:\,\|F_A-G_A\|_n^2\le 2^j\eta^2/n}\sum_{i=1}^n(F_A^*(x_i)-\mathbb{1}_{\{y_i\in A\}})(G_A(x_i)-F_A(x_i))\,\middle|\,\{x_i\}_{i=1}^n\right)\\
&\le 2\mathbb{E}\left(\sup_{A\in\Lambda}\sup_{F_A\in\mathcal{F}_A\,:\,\|F_A-G_A\|_n^2\le 2^j\eta^2/n}\sum_{i=1}^n\xi_i\mathbb{1}_{\{y_i\in A\}}(G_A(x_i)-F_A(x_i))\,\middle|\,\{x_i\}_{i=1}^n\right)\\
&= 2\mathbb{E}\left(\mathbb{E}\left(\sup_{A\in\Lambda}\sup_{F_A\in\mathcal{F}_A\,:\,\|F_A-G_A\|_n^2\le 2^j\eta^2/n}\sum_{i=1}^n\xi_i\mathbb{1}_{\{y_i\in A\}}(G_A(x_i)-F_A(x_i))\,\middle|\,\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n\right)\,\middle|\,\{x_i\}_{i=1}^n\right).
\end{aligned}\tag{8}$$

Next notice that for fixed $\{(x_i, y_i)\}_{i=1}^n$, the vectors

$$\mathcal{V}_n(\Lambda) = \{(\mathbf{1}_{\{y_i\in A\}})_{i=1}^n : A \in \Lambda\}$$

form a chain in $\{0,1\}^n$ under coordinatewise order, by Assumption 1. Distinct vectors in such a chain have distinct sums of coordinates, each in $\{0,\dots,n\}$. Choose a finite collection $\Lambda' \subseteq \Lambda$ with one representative for each distinct sample vector. Then

$$|\Lambda'| = |\mathcal{V}_n(\Lambda)| \le n+1. \tag{9}$$

To use this finite collection, first enlarge the admissible differences to $\mathcal{D}(r_j) = (\mathcal{F}-\mathcal{F})\cap B_{r_j}(0)$. Indeed, $G_A - F_A \in \mathcal{D}(r_j)$ whenever $\|F_A - G_A\|_n^2 \le 2^j\eta^2/n$, so

$$\begin{aligned}
&\sup_{A\in\Lambda}\sup_{F_A\in\mathcal{F}_A:\,\|F_A-G_A\|_n^2\le 2^j\eta^2/n}\sum_{i=1}^n\xi_i\mathbf{1}_{\{y_i\in A\}}\big(G_A(x_i)-F_A(x_i)\big)\\
&\le \sup_{A\in\Lambda}\sup_{h\in\mathcal{D}(r_j)}\sum_{i=1}^n\xi_i\mathbf{1}_{\{y_i\in A\}}h(x_i) = \max_{A\in\Lambda'}\sup_{h\in\mathcal{D}(r_j)}\sum_{i=1}^n\xi_i\mathbf{1}_{\{y_i\in A\}}h(x_i).
\end{aligned}\tag{145}$$

The first step is an inequality, not an identity: G_A and $\mathcal{F}_A$ may differ for sets with the same sample vector.

Therefore, by Padilla et al. (2025, Lemma 3), it follows that

$$\begin{aligned}
&\mathbb{E}\left[\sup_{A\in\Lambda}\sup_{F_A\in\mathcal{F}_A:\,\|F_A-G_A\|_n^2\le 2^j\eta^2/n}\sum_{i=1}^n\xi_i\mathbf{1}_{\{y_i\in A\}}\big(G_A(x_i)-F_A(x_i)\big)\,\middle|\,\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n\right]\\
&\le\mathbb{E}\left[\max_{A\in\Lambda'}\sup_{h\in\mathcal{D}(r_j)}\sum_{i=1}^n\xi_i\mathbf{1}_{\{y_i\in A\}}h(x_i)\,\middle|\,\{x_i\}_{i=1}^n,\{y_i\}_{i=1}^n\right]\\
&\le C\int_0^{2^{j/2}\eta}\sqrt{\log N\big(\varepsilon/\sqrt{n},\mathcal{D}(r_j),\|\cdot\|_n\big)}\,d\varepsilon + C2^{j/2}\eta\sqrt{\log(2|\Lambda'|)}\\
&\le C\sqrt{n}\,\mathcal{J}(r_j) + C2^{j/2}\eta\sqrt{\log(2n+2)}.
\end{aligned}\tag{146}$$

for a positive constant C. Here $(h(x_i))_{i=1}^n$ has Euclidean norm at most $2^{j/2}\eta$. For each $A \in \Lambda'$, the multipliers $(\xi_i\mathbf{1}_{\{y_i\in A\}})_{i=1}^n$ have independent, centered, 1-sub-Gaussian coordinates; independence between different A's is not required. The covering-number form of the lemma follows by the packing–covering comparison at a fixed numerical scale.

Combining equation equation 146 with equation 7, equation 8, and equation 145, and using $\sum_{j=1}^J 2^{-j/2} \le \sum_{j\ge1} 2^{-j/2} < \infty$, gives

$$\begin{aligned}
&\mathbb{P}\left(\sup_{A\in\Lambda}\|\widehat{F}_A^{\mathrm{pre}} - G_A\|_n > \frac{\eta}{\sqrt{n}}\,\middle|\,\{x_i\}_{i=1}^n\right)\\
&\qquad\le C\left\{\frac{\sqrt{n}}{\eta^2}\sum_{j=1}^J 2^{-j}\mathcal{J}(r_j) + \frac{\sqrt{\log(2n+2)}}{\eta} + \frac{\sqrt{n}}{\eta}\sup_{A\in\Lambda}\|G_A - F_A^*\|_\infty\right\}.
\end{aligned}$$

Finally, $\sup_{A\in\Lambda}\|\widehat{F}_A^{\mathrm{pre}} - F_A^*\|_n \le \sup_{A\in\Lambda}\|\widehat{F}_A^{\mathrm{pre}} - G_A\|_n + b$ by the triangle inequality, which proves the desired claim

□

B.2 Proof of Corollary 1

Proof. All steps of Appendix B.1 remain unchanged except for the cardinality bound on $\mathcal{V}_n(\Lambda)$. For each coordinate k, the threshold class

$$\left\{\left(\mathbb{1}_{\{Y_{1k}\le t\}},\dots,\mathbb{1}_{\{Y_{nk}\le t\}}\right) :\ t\in\mathbb{R}\right\}$$

induces at most $n+1$ distinct binary vectors on the sample, corresponding to the $n+1$ possible threshold positions relative to the ordered values $Y_{1k},\dots,Y_{nk}$, including the all-zero vector obtained when $t<\min_i Y_{ik}$. For $t=(t_1,\dots,t_u)\in\mathbb{R}^u$, the sample indicator vector associated with the lower orthant $A(t)$ satisfies

$$\left(\mathbb{1}_{\{Y_i\in A(t)\}}\right)_{i=1}^n = \bigodot_{k=1}^u \left(\mathbb{1}_{\{Y_{ik}\le t_k\}}\right)_{i=1}^n,$$

where $\bigodot$ denotes coordinatewise multiplication. Consequently,

$$|\mathcal{V}_n(\Lambda)| \le (n+1)^u. \tag{R9}$$

Ties only decrease this count. Use equation equation R9 in the penultimate line of equation equation 146 to replace $\log(2n+2)$ by $\log 2 + u\log(n+1)$. This proves the desired claim.

□

B.3 Proof of Theorem 2

Proof. Condition on the design, let

$$\varepsilon_{i,j} = z_i^j - F_{A_j}^*(x_i), \qquad R(\boldsymbol{g}) = \sum_{j=1}^m\sum_{i=1}^n \left\{F_{A_j}^*(x_i) - H_j(\boldsymbol{g})(x_i)\right\}^2,$$

and write $\widehat{\boldsymbol{g}} = (\widehat{g}_1,\dots,\widehat{g}_m)$. The least-squares basic inequality gives

$$\begin{aligned} R(\widehat{\boldsymbol{g}}) &\le R(\widetilde{\boldsymbol{g}}) + 2\sum_{j=1}^m\sum_{i=1}^n \varepsilon_{i,j}\left\{H_j(\widehat{\boldsymbol{g}})(x_i) - H_j(\widetilde{\boldsymbol{g}})(x_i)\right\} \\ &\le nm\{\Delta_{n,m}(\widetilde{\boldsymbol{g}})\}^2 + 2\sum_{j=1}^m\sum_{i=1}^n \varepsilon_{i,j}\left\{H_j(\widehat{\boldsymbol{g}})(x_i) - H_j(\widetilde{\boldsymbol{g}})(x_i)\right\}. \end{aligned} \tag{10}$$

Since $R(\widehat{\boldsymbol{g}}) \le nm$, define

$$\mathcal{S}_k = \left\{2^{k-1}\eta^2 < R(\widehat{\boldsymbol{g}}) \le 2^k\eta^2\right\}, \qquad k=1,\dots,K.$$

Then

$$\{R(\widehat{\boldsymbol{g}}) > \eta^2\} \subseteq \bigcup_{k=1}^K \mathcal{S}_k.$$

Choose $C_1 \ge 2$. On $\mathcal{S}_k$, the condition $nm\{\Delta_{n,m}(\widetilde{\boldsymbol{g}})\}^2 < C_1^{-1}\eta^2$ and equation 10 imply

$$\begin{aligned} &2\sum_{j=1}^m\sum_{i=1}^n \varepsilon_{i,j}\left\{H_j(\widehat{\boldsymbol{g}})(x_i) - H_j(\widetilde{\boldsymbol{g}})(x_i)\right\} \\ &\quad\ge R(\widehat{\boldsymbol{g}}) - nm\{\Delta_{n,m}(\widetilde{\boldsymbol{g}})\}^2 \\ &\quad> 2^{k-1}\eta^2 - C_1^{-1}\eta^2 \ge 2^{k-2}\eta^2. \end{aligned}$$

Moreover, for every j,

$$\begin{aligned}\|H_j(\widehat{\boldsymbol{g}}) - H_j(\widetilde{\boldsymbol{g}})\|_n &\le \|H_j(\widehat{\boldsymbol{g}}) - F^*_{A_j}\|_n + \|H_j(\widetilde{\boldsymbol{g}}) - F^*_{A_j}\|_n \\ &\le \frac{2^{k/2}\eta}{\sqrt{n}} + \Delta_{n,m}(\widetilde{\boldsymbol{g}}) \\ &\le \frac{2^{k/2+1}\eta}{\sqrt{n}}.\end{aligned}$$

Therefore, by Markov's inequality,

$$\mathbb{P}(\mathcal{S}_k) \le \frac{C}{2^k\eta^2}\sum_{j=1}^{m}\mathbb{E}\left\{\sup_{\substack{f\in\mathcal{I}_{j,n}(\widetilde{\boldsymbol{g}})\\ \|f\|_n\le 2^{k/2+1}\eta/\sqrt{n}}}\left|\sum_{i=1}^{n}\varepsilon_{i,j}f(x_i)\right|\right\}.$$

Let $\xi_1,\dots,\xi_n$ be independent Rademacher variables. Conditional symmetrization and contraction give

$$\begin{aligned}&\mathbb{E}\left\{\sup_{\substack{f\in\mathcal{I}_{j,n}(\widetilde{\boldsymbol{g}})\\ \|f\|_n\le 2^{k/2+1}\eta/\sqrt{n}}}\left|\sum_{i=1}^{n}\varepsilon_{i,j}f(x_i)\right|\right\} \\ &\quad\le C\mathbb{E}_\xi\left\{\sup_{\substack{f\in\mathcal{I}_{j,n}(\widetilde{\boldsymbol{g}})\\ \|f\|_n\le 2^{k/2+1}\eta/\sqrt{n}}}\left|\sum_{i=1}^{n}\xi_i f(x_i)\right|\right\} \\ &\quad\le C\left[1+\sqrt{n}\int_{1/n}^{C_3 2^{k/2}\eta/\sqrt{n}}\sqrt{\log N\big(\epsilon,\mathcal{I}_{j,n}(\widetilde{\boldsymbol{g}}),\|\cdot\|_n\big)}\,d\epsilon\right],\end{aligned}$$

where the last inequality follows from Dudley's entropy bound. Hence,

$$\begin{aligned}\mathbb{P}\{R(\widehat{\boldsymbol{g}}) > \eta^2\} &\le \sum_{k=1}^{K}\mathbb{P}(\mathcal{S}_k) \\ &\le C_2\sum_{j=1}^{m}\sum_{k=1}^{K}\frac{1}{2^{k-2}\eta^2}\left[1+\sqrt{n}\int_{1/n}^{C_3 2^{k/2}\eta/\sqrt{n}}\sqrt{\log N\big(\epsilon,\mathcal{I}_{j,n}(\widetilde{\boldsymbol{g}}),\|\cdot\|_n\big)}\,d\epsilon\right],\end{aligned}$$

which proves equation 5. □

B.4 Auxiliary results for Theorem 2

The first lemma records how perturbations of the base functions propagate through the DAG. It will also be used to relate the covering entropy of the base class to that of each constrained output.

Lemma 1 (Pathwise representation and stability). *For any $\boldsymbol{g} = (g_1,\dots,g_m)$,*

$$\widetilde{H}_j(\boldsymbol{g})(x) = \max_{p=(i_1,\dots,i_\ell)\in\mathcal{PR}(j)}\left\{g_{i_1}(x) + \sum_{r=2}^{\ell}\log\big(1+e^{g_{i_r}(x)}\big)\right\}. \tag{11}$$

Consequently, for any $\boldsymbol{g},\boldsymbol{h}\in\mathcal{F}^m$,

$$|H_j(\boldsymbol{g})(x) - H_j(\boldsymbol{h})(x)| \le \frac{1}{4}\max_{p\in\mathcal{PR}(j)}\sum_{k\in p}|g_k(x)-h_k(x)|, \tag{12}$$

$$\|H_j(\boldsymbol{g}) - H_j(\boldsymbol{h})\|_n^2 \le \frac{\kappa_j}{16}\sum_{k\in\overline{\mathcal{A}}(j)}\|g_k - h_k\|_n^2. \tag{13}$$

Proof. Equation equation 11 follows directly by recursively expanding the definition of $\widetilde{H}_j$ along all root-to-j paths. Since the maximum map and the softplus function are one-Lipschitz,

$$\begin{aligned}
&\left|\widetilde{H}_j(\boldsymbol{g})(x) - \widetilde{H}_j(\boldsymbol{h})(x)\right| \\
&\quad\le \max_{p=(i_1,\dots,i_\ell)\in\mathcal{PR}(j)} \left\{ |g_{i_1}(x) - h_{i_1}(x)| + \sum_{r=2}^{\ell} \left|\log\bigl(1+e^{g_{i_r}(x)}\bigr) - \log\bigl(1+e^{h_{i_r}(x)}\bigr)\right| \right\} \\
&\quad\le \max_{p\in\mathcal{PR}(j)} \sum_{k\in p} |g_k(x) - h_k(x)|.
\end{aligned}$$

The $1/4$-Lipschitz property of the sigmoid therefore gives

$$\begin{aligned}
|H_j(\boldsymbol{g})(x) - H_j(\boldsymbol{h})(x)| &\le \frac{1}{4}\left|\widetilde{H}_j(\boldsymbol{g})(x) - \widetilde{H}_j(\boldsymbol{h})(x)\right| \\
&\le \frac{1}{4} \max_{p\in\mathcal{PR}(j)} \sum_{k\in p} |g_k(x) - h_k(x)|,
\end{aligned}$$

which proves equation 12. Moreover, by Cauchy–Schwarz, $|p| \le \kappa_j$, and $p \subseteq \overline{\mathcal{A}}(j)$,

$$\begin{aligned}
\|H_j(\boldsymbol{g}) - H_j(\boldsymbol{h})\|_n^2 &\le \frac{1}{16n}\sum_{i=1}^{n} \max_{p\in\mathcal{PR}(j)} \left\{ \sum_{k\in p} |g_k(x_i) - h_k(x_i)| \right\}^2 \\
&\le \frac{1}{16n}\sum_{i=1}^{n} \max_{p\in\mathcal{PR}(j)} |p| \sum_{k\in p} |g_k(x_i) - h_k(x_i)|^2 \\
&\le \frac{\kappa_j}{16n}\sum_{i=1}^{n} \sum_{k\in\overline{\mathcal{A}}(j)} |g_k(x_i) - h_k(x_i)|^2 \\
&= \frac{\kappa_j}{16} \sum_{k\in\overline{\mathcal{A}}(j)} \|g_k - h_k\|_n^2,
\end{aligned}$$

which proves equation 13. □

Lemma 2 (Covering entropy of a constrained output)**.** *For every $\widetilde{\boldsymbol{g}} \in \mathcal{F}^m$,*

$$N\bigl(\epsilon, \mathcal{I}_{j,n}(\widetilde{\boldsymbol{g}}), \|\cdot\|_n\bigr) \le \left[N\left(\frac{4\epsilon}{\sqrt{\kappa_j a_j}}, \mathcal{F}, \|\cdot\|_n\right)\right]^{a_j}. \tag{14}$$

Under the covering condition following Theorem 2, this implies

$$\log N\bigl(\epsilon, \mathcal{I}_{j,n}(\widetilde{\boldsymbol{g}}), \|\cdot\|_n\bigr) \le a_j v_n \log\left(\frac{C A_n \sqrt{\kappa_j a_j}}{\epsilon}\right) \tag{15}$$

for a universal constant C.

Proof. Let $\{f_1, \dots, f_M\}$ be a $4\epsilon/\sqrt{\kappa_j a_j}$-cover of $\mathcal{F}$ under $\|\cdot\|_n$, where

$$M = N\left(\frac{4\epsilon}{\sqrt{\kappa_j a_j}}, \mathcal{F}, \|\cdot\|_n\right).$$

For each $k \in \overline{\mathcal{A}}(j)$, choose h_k from this cover such that

$$\|g_k - h_k\|_n \le \frac{4\epsilon}{\sqrt{\kappa_j a_j}}.$$

By equation 13,

$$\begin{aligned}
\|H_j(\boldsymbol{g}) - H_j(\boldsymbol{h})\|_n^2 &\le \frac{\kappa_j}{16} \sum_{k\in\overline{\mathcal{A}}(j)} \|g_k - h_k\|_n^2 \\
&\le \frac{\kappa_j a_j}{16}\left(\frac{4\epsilon}{\sqrt{\kappa_j a_j}}\right)^2 = \epsilon^2.
\end{aligned}$$

There are at most M^{a_j} such coordinatewise choices. Since translation by $H_j(\widetilde{\boldsymbol{g}})$ does not change covering numbers,

$$N\big(\epsilon, \mathcal{I}_{j,n}(\widetilde{\boldsymbol{g}}), \|\cdot\|_n\big) \le M^{a_j},$$

which proves equation 14. Under the stated covering condition,

$$\begin{aligned}\log N\big(\epsilon, \mathcal{I}_{j,n}(\widetilde{\boldsymbol{g}}), \|\cdot\|_n\big) &\le a_j \log N\left(\frac{4\epsilon}{\sqrt{\kappa_j a_j}}, \mathcal{F}, \|\cdot\|_n\right) \\ &\le a_j v_n \log\left(\frac{C A_n \sqrt{\kappa_j a_j}}{\epsilon}\right),\end{aligned}$$

proving equation 15. □

The following localized bound is used to derive the explicit rate following Theorem 2.

Lemma 3 (Localized multiplier bound)**.** *Suppose that*

$$\log N(u, \mathcal{F}, \|\cdot\|_n) \le v_n \log(A_n/u), \qquad 0 < u \le B_n,$$

where $v_n \ge 1$ and $A_n \ge eB_n$. Fix $\widetilde{\boldsymbol{g}} \in \mathcal{F}^m$ and let

$$\varepsilon_{i,j} = z_i^j - F^*_{A_j}(x_i).$$

For every $\delta \in (0, 1/2)$, conditionally on the design, with probability at least $1-\delta$, simultaneously for all $j \le m$ and $f \in \mathcal{I}_{j,n}(\widetilde{\boldsymbol{g}})$,

$$\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_{i,j} f(x_i)\right| \le \frac{1}{16}\|f\|_n^2 + C\frac{a_j v_n}{n}\log\left(\frac{eA_n n m\sqrt{\kappa_m a_{\max,m}}}{\delta}\right). \tag{16}$$

Proof. Condition on $x_1, \dots, x_n$ and, for $r \in [1/n, 1]$, define

$$Z_j(r) = \sup_{\substack{f \in \mathcal{I}_{j,n}(\widetilde{\boldsymbol{g}}) \\ \|f\|_n \le r}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_{i,j} f(x_i)\right|.$$

Since $\mathbb{E}(\varepsilon_{i,j} \mid x_i) = 0$ and $|\varepsilon_{i,j}| \le 1$, symmetrization, contraction, Dudley's inequality, and equation 15 give

$$\begin{aligned}\mathbb{E}\{Z_j(r)\} &\le \frac{C}{n} + \frac{C}{\sqrt{n}}\int_{1/n}^r \sqrt{a_j v_n \log\left(\frac{C A_n\sqrt{\kappa_j a_j}}{u}\right)}\,du \\ &\le \frac{C}{n} + Cr\sqrt{\frac{a_j v_n}{n}\log\big(eA_n n\sqrt{\kappa_j a_j}\big)}.\end{aligned} \tag{17}$$

Because $\mathcal{I}_{j,n}(\widetilde{\boldsymbol{g}})$ has envelope one, Bousquet's inequality gives, for every $t \ge 1$,

$$\mathbb{P}\left[Z_j(r) > C\left\{\mathbb{E}Z_j(r) + r\sqrt{\frac{t}{n}} + \frac{t}{n}\right\}\right] \le e^{-t}. \tag{18}$$

In the usual form, the square-root term also contains $\mathbb{E}Z_j(r)$; Young's inequality absorbs that term into $C\mathbb{E}Z_j(r) + Ct/n$. No independence across event indices j is assumed.

Let

$$S = \lceil \log_2 n \rceil, \qquad r_s = 2^{-s}, \qquad t_s = \log\left\{\frac{Cm(s+2)^2}{\delta}\right\}, \qquad s = 0, \dots, S,$$

and define

$$q_j = a_j v_n \log\left(\frac{eA_n n m\sqrt{\kappa_m a_{\max,m}}}{\delta}\right).$$

Then, uniformly over $j \le m$ and $s \le S$,

$$a_j v_n \log\big(eA_n n\sqrt{\kappa_j a_j}\big) + t_s \le C q_j,$$

and

$$\sum_{j=1}^{m}\sum_{s=0}^{S} e^{-t_s} \le \delta.$$

Therefore, by equation 17, equation 18, and a union bound, with probability at least $1-\delta$,

$$Z_j(r_s) \le C\left\{r_s\sqrt{\frac{q_j}{n}} + \frac{q_j}{n}\right\}, \qquad j \le m, \quad s \le S. \tag{19}$$

For any $f \in \mathcal{I}_{j,n}(\widetilde{\boldsymbol{g}})$ with $\|f\|_n > 1/n$, choose s such that $r_s/2 < \|f\|_n \le r_s$. Then equation 19 gives

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{i,j}f(x_i)\right| &\le Z_j(r_s)\\ &\le C\left\{\|f\|_n\sqrt{\frac{q_j}{n}} + \frac{q_j}{n}\right\}.\end{aligned}$$

If $\|f\|_n \le 1/n$, then

$$\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{i,j}f(x_i)\right| \le \|f\|_n \le \frac{1}{n} \le \frac{q_j}{n}.$$

Thus, uniformly over j and f,

$$\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{i,j}f(x_i)\right| \le C\left\{\|f\|_n\sqrt{\frac{q_j}{n}} + \frac{q_j}{n}\right\}.$$

Finally, Young's inequality,

$$C\|f\|_n\sqrt{\frac{q_j}{n}} \le \frac{1}{16}\|f\|_n^2 + C\frac{q_j}{n},$$

proves equation 16. □

B.5 Proof of the explicit bound equation 6

Proof. This proof starts from the basic inequality (10) and Lemma 3. Fix $\widetilde{\boldsymbol{g}} \in \mathcal{F}^m$ and write

$$\mathcal{E} = \sum_{j=1}^{m}\|H_j(\widehat{\boldsymbol{g}}) - F^*_{A_j}\|_n^2, \qquad \mathcal{A} = \sum_{j=1}^{m}\|H_j(\widetilde{\boldsymbol{g}}) - F^*_{A_j}\|_n^2.$$

By the definition of $\Delta_{n,m}(\widetilde{\boldsymbol{g}})$,

$$\mathcal{A} \le m\{\Delta_{n,m}(\widetilde{\boldsymbol{g}})\}^2. \tag{20}$$

Let

$$f_j = H_j(\widehat{\boldsymbol{g}}) - H_j(\widetilde{\boldsymbol{g}}) \in \mathcal{I}_{j,n}(\widetilde{\boldsymbol{g}}).$$

Dividing equation 10 by n and applying Lemma 3 give, with probability at least $1-\delta$,

$$\begin{aligned}\mathcal{E} &\le \mathcal{A} + 2\sum_{j=1}^{m}\left|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{i,j}f_j(x_i)\right|\\ &\le \mathcal{A} + \frac{1}{8}\sum_{j=1}^{m}\|f_j\|_n^2 + C\frac{v_n}{n}\sum_{j=1}^{m}a_j\log\left(\frac{eA_n n m\sqrt{\kappa_m a_{\max,m}}}{\delta}\right).\end{aligned} \tag{21}$$

Moreover,

$$\begin{aligned}\sum_{j=1}^{m}\|f_j\|_n^2 &\le 2\sum_{j=1}^{m}\|H_j(\widehat{\boldsymbol{g}}) - F^*_{A_j}\|_n^2 + 2\sum_{j=1}^{m}\|H_j(\widetilde{\boldsymbol{g}}) - F^*_{A_j}\|_n^2\\ &= 2\mathcal{E} + 2\mathcal{A}.\end{aligned} \tag{22}$$

Substituting equation 22 into equation 21 yields

$$\mathcal{E} \leq \mathcal{A} + \frac{1}{4}\mathcal{E} + \frac{1}{4}\mathcal{A} + C\frac{v_n}{n}\sum_{j=1}^{m} a_j \log\left(\frac{eA_n nm\sqrt{\kappa_m a_{\max,m}}}{\delta}\right),$$

and hence, by equation 20,

$$\mathcal{E} \leq C\left\{m\{\Delta_{n,m}(\widetilde{\boldsymbol{g}})\}^2 + \frac{v_n}{n}\sum_{j=1}^{m} a_j \log\left(\frac{eA_n nm\sqrt{\kappa_m a_{\max,m}}}{\delta}\right)\right\}.$$

Dividing by m and using $m^{-1}\sum_{j=1}^{m} a_j = \bar{a}_m$ gives

$$\frac{1}{m}\sum_{j=1}^{m} \|\widehat{F}_{A_j} - F^*_{A_j}\|_n^2 \leq C\left\{\{\Delta_{n,m}(\widetilde{\boldsymbol{g}})\}^2 + \frac{\bar{a}_m v_n}{n}\log\left(\frac{eA_n nm\sqrt{\kappa_m a_{\max,m}}}{\delta}\right)\right\}.$$

For every $\gamma > 0$, choose $\widetilde{\boldsymbol{g}}_\gamma \in \mathcal{F}^m$ satisfying

$$\Delta_{n,m}(\widetilde{\boldsymbol{g}}_\gamma) \leq \Delta_{n,m} + \gamma.$$

Letting $\gamma \downarrow 0$ proves equation 6. □

C DNN THEORY FOR THE PRELIMINARY ESTIMATOR

Corollary 5. ***[Detailed statement of Corollary 3].*** *Consider the notation and assumptions from Theorem 1, and that for all $A \in \Lambda$ it holds that $F^*_A \in \mathcal{H}(l_A, \mathcal{P}_A)$ for some $l_A \in \mathbb{N}$ and $\mathcal{P}_A \subset [1,\infty) \times \mathbb{N}$. In addition, suppose that each function μ in the definition of F^*_A can have different smoothness $p_\mu = q_\mu + u_\mu$, for $q_\mu \in \mathbb{N}$, $u_\mu \in (0,1]$, of perhaps different input dimension d_μ, so that $(p_\mu, d_\mu) \in \mathcal{P}_A$. Let $d^A_{\max}$ be the largest input dimension and $p^A_{\max}$ the largest smoothness of any function μ associated with F^*_A. We require that all the partial derivatives of order less than or equal to q_μ are uniformly bounded by constant that does not depend on A, and each function μ is Lipschitz continuous with Lipschitz constant $C_{\text{Lip}} \geq 1$. Also, assume that $\sup_{A\in\Lambda} \max\{p^A_{\max}, d^A_{\max}\} = O(1)$. Let*

$$\phi_n = \sup_{A\in\Lambda} \max_{(p,d)\in\mathcal{P}_A} n^{\frac{-2p}{(2p+d)}}.$$

Then there exists choices of L and w such that for sufficiently large positive constants c_3 and c_4 if

$$L = \lceil c_3 \log n \rceil \quad \textit{and} \quad w = \left\lceil c_4 \sup_{A\in\Lambda} \max_{(p,d)\in\mathcal{P}_A} n^{\frac{d}{2(2p+d)}} \right\rceil \tag{23}$$

or

$$L = \left\lceil c_3 \sup_{A\in\Lambda} \max_{(p,d)\in\mathcal{P}_A} n^{\frac{d}{2(2p+d)}} \log n \right\rceil \quad \textit{and} \quad w = \lceil c_4 \rceil, \tag{24}$$

then, taking $\mathcal{F} = \{f \in \mathcal{F}(L,w) : \|f\|_\infty \leq B\}$ for a sufficiently large constant $B = O(1)$, we obtain that

$$\sup_{A\in\Lambda} \frac{1}{n}\sum_{i=1}^{n} (F^*_A(x_i) - \widehat{F}^{pre}_A(x_i))^2 = O_{\mathbb{P}}\left(\phi_n \log^4 n\right).$$

C.1 PROOF OF COROLLARY 3 AND 5

Proof. We give all approximation, entropy, integration, and probability-balancing steps.

Uniform approximation. The uniform hierarchy assumptions allow one common pair of sufficiently large architecture constants in equations equation 23 or equation 24. Kohler & Langer's hierarchical approximation construction then gives, for every A, an admissible network G_A with

$$b_n := \sup_{A\in\Lambda} \|G_A - F^*_A\|_\infty \leq C\sqrt{\phi_n}.$$

The constants depend on the common hierarchy bounds and fixed input dimension, not on A.

Pseudodimension and covering entropy. Fix $B \geq 2$ as a constant independent of n and $A \in \Lambda$, and let

$$\mathcal{F} = \{f \in \mathcal{F}(L, w) : \|f\|_\infty \leq B\}.$$

Since $0 \leq F_A^* \leq 1$ for every $A \in \Lambda$ and the uniform approximation error tends to zero, B can be chosen sufficiently large so that the approximating networks considered below belong to $\mathcal{F}$ for all sufficiently large n.

A fully connected ReLU network in $\mathcal{F}(L, w)$ with input dimension d, L hidden layers of width w, and one scalar output has

$$W = (d+1)w + (L-1)(w+1)w + (w+1)$$

trainable parameters and

$$U_{\text{net}} = Lw + 1$$

computational units. Let $C_{\text{pd}} > 0$ denote the universal constant in the pseudo-dimension bound for fully connected ReLU networks, and define

$$v_n := \max\left\{1,\, C_{\text{pd}} LW \log\big(eU_{\text{net}}\big)\right\}.$$

By the same pseudo-dimension and bounded VC-subgraph argument as in Lemma 6, there exists a universal constant $C_{\text{ent}} \geq 1$, independent of n, L, w, B, u, such that

$$\log N(u, \mathcal{F}, \|\cdot\|_n) \leq C_{\text{ent}} v_n \log\left(\frac{C_{\text{ent}} B}{u}\right), \qquad 0 < u \leq B.$$

Moreover, the product of two $(u/2)$-covers of $\mathcal{F}$ induces a u-cover of $\mathcal{F} - \mathcal{F}$. Since $\mathcal{D}(r) \subseteq \mathcal{F} - \mathcal{F}$,

$$\begin{aligned}\log N(u, \mathcal{D}(r), \|\cdot\|_n) &\leq 2 \log N\Big(\frac{u}{2}, \mathcal{F}, \|\cdot\|_n\Big) \\ &\leq 2C_{\text{ent}} v_n \log\left(\frac{2C_{\text{ent}} B}{u}\right), \qquad 0 < u \leq r \leq 2\sqrt{2}\, B.\end{aligned}$$

The integral. For $A_0 = CB$ with $A_0/r \geq e$, substitute $u = rt$:

$$\begin{aligned}\int_0^r \sqrt{\log(A_0/u)}\, du &= r \int_0^1 \sqrt{\log(A_0/r) + \log(1/t)}\, dt \\ &\leq r\left\{\sqrt{\log(A_0/r)} + \int_0^1 \sqrt{\log(1/t)}\, dt\right\} \\ &= r\{\sqrt{\log(A_0/r)} + \Gamma(3/2)\} \leq Cr\sqrt{\log(eA_0/r)}.\end{aligned}$$

For Theorem 1, $\eta > 1$ and every shell has $r_j \geq n^{-1/2}$, while $r_J \leq 2\sqrt{2}B$. Thus

$$\mathcal{J}(r_j) \leq Cr_j\sqrt{v_n \log(en)}, \qquad \frac{\sqrt{n}}{\eta^2} \sum_{j=1}^{J} 2^{-j} \mathcal{J}(r_j) \leq \frac{C\sqrt{v_n \log(en)}}{\eta}. \tag{R10}$$

Architecture size. Let $s_n = \sqrt{n\phi_n} = \sup_A \max_{(p,d_0)\in\mathcal{P}_A} n^{d_0/[2(2p+d_0)]}$. For the wide architecture, $L \asymp \log(en)$, $w \asymp s_n$, so $W \lesssim n\phi_n \log(en)$ and $v_n \lesssim n\phi_n \log^3(en)$. For the narrow architecture, $L \asymp s_n \log(en)$, $w = O(1)$, so $W \lesssim L$ and again $v_n \lesssim L^2 \log(eL) \lesssim n\phi_n \log^3(en)$. Therefore

$$\sqrt{v_n \log(en)} + \sqrt{\log(en)} + \sqrt{n}\, b_n \leq C\sqrt{n\phi_n} \log^2(en). \tag{R11}$$

Probability and rate. Take $\eta = M\sqrt{n\phi_n} \log^2(en)$ for a fixed constant $M \geq 1$. Equations equation 4, equation R10, and equation R11 give an upper-tail probability at most C/M, and the threshold is at most $(M + C)\sqrt{\phi_n} \log^2(en)$. For every prescribed probability tolerance, choose M sufficiently large. This is precisely

$$\sup_A \|\widehat{F}_A^{\text{pre}} - F_A^*\|_n^2 = O_{\mathbb{P}}\{\phi_n \log^4(en)\}.$$

□

C.2 COROLLARY 4

Proof. This follows from Corollary 1, using the same network-entropy calculation as in the proof of Corollary 3 and taking

$$\eta \asymp \sqrt{u \log n} + \sqrt{LW \log(U) \log n}.$$

□

D DNN THEORY FOR THE CONSTRAINED ESTIMATOR

This section develops the DNN approximation theory underlying Corollary 2. We proceed from the basic definitions and the required ReLU approximation results to the DAG-constrained network construction and its approximation properties. With these ingredients in place, we present the detailed statement of Corollary 2 in Corollary 6 and prove the resulting DNN rate.

We use the DAG notation from Theorem 2 and the dense ReLU class $\mathcal{F}(L, w)$ from Appendix A; L counts hidden layers and w is their common width. Let

$$\ell(t) = \log \frac{t}{1-t}, \quad 0 < t < 1, \qquad \psi(t) = \log(e^t - 1), \quad t > 0,$$

be the inverse functions of the sigmoid τ and the softplus $\mathrm{sp}(t) = \log(1 + e^t)$, respectively. For $0 < s \le 1$, write

$$\Lambda_s = \log(e/s),$$

and, for an interval I, write

$$\|f\|_{\infty,I} = \sup_{t \in I} |f(t)|.$$

Lemma 4 (Deep–narrow ReLU approximation of logit and inverse softplus)**.** *There is a universal integer $w_{\mathrm{cal}} \ge 1$ with the following properties.*

1. *For every $0 < \epsilon < 1/4$ and $0 < \eta < 1$, there exists a scalar ReLU network $\mathsf{L}_{\epsilon,\eta}$ whose weights depend only on (ϵ, η) and such that*

$$\left\|\mathsf{L}_{\epsilon,\eta} - \ell\right\|_{\infty,[\epsilon,1-\epsilon]} \le \eta. \tag{25}$$

   Its width is at most w_{cal}, and its number of hidden layers, denoted by $D_\ell(\epsilon, \eta)$, satisfies

$$D_\ell(\epsilon, \eta) \le C\Lambda_\epsilon \Lambda_\eta (\Lambda_\epsilon + \Lambda_\eta). \tag{26}$$

2. *For every $0 < \delta \le 1$, $U > \delta$, and $0 < \zeta < 1$, there exists a scalar ReLU network $\mathsf{S}_{\delta,U,\zeta}$ whose weights depend only on (δ, U, ζ) and such that*

$$\left\|\mathsf{S}_{\delta,U,\zeta} - \psi\right\|_{\infty,[\delta,U]} \le \zeta. \tag{27}$$

   Its width is at most w_{cal}, and its number of hidden layers, denoted by $D_\psi(\delta, U, \zeta)$, satisfies

$$D_\psi(\delta, U, \zeta) \le C(1 + U + \Lambda_\delta)\Lambda_\zeta(1 + U + \Lambda_\delta + \Lambda_\zeta). \tag{28}$$

   Moreover,

$$-\psi(\delta) \le \log(1/\delta), \qquad \psi(U) \le U. \tag{29}$$

All constants in this lemma are universal and independent of $(\epsilon, \eta, \delta, U, \zeta)$.

Proof outline. We first show that a polynomial can be approximated by a ReLU network of universal width by combining a fixed-width multiplication network with Horner's recursion. We then extend this construction to continuous piecewise polynomials by evaluating the pieces successively, so that the number of pieces affects the depth but not the width. Finally, we partition the domains of $\log$ and ψ into intervals on which, after rescaling to $[0, 1]$, their nonlinear parts form uniformly holomorphic families. Polynomial approximation on these intervals then gives the stated networks.

Polynomial evaluation. We begin with a fixed-width network for approximate multiplication and use it as the basic module for evaluating a polynomial recursively. The construction keeps the width universal, while the polynomial degree and approximation accuracy enter through the depth. By Kohler & Langer (2021, Lemma 4), for every $B \geq 1$ and $0 < \omega < 1$, there exists a two-input scalar ReLU network of width 18 and depth at most $C \log(eB^2/\omega)$,

$$\mathsf{M}_{B,\omega} \in \mathcal{F}\left(C \log \frac{eB^2}{\omega}, 18\right),$$

such that

$$\sup_{|u|,|v| \leq B} |\mathsf{M}_{B,\omega}(u,v) - uv| \leq \omega. \tag{30}$$

Then, let $q \in \mathbb{N}_0$ and let $b_0, \ldots, b_q \in \mathbb{R}$ be arbitrary. Consider the polynomial

$$P : [0,1] \to \mathbb{R}, \qquad P(t) = \sum_{r=0}^{q} b_r t^r.$$

Define

$$A := 1 + \sum_{r=0}^{q} |b_r|.$$

Fix a target approximation accuracy $\xi \in (0,1)$ and set

$$\omega := \frac{\xi}{q+1}.$$

Then $0 < \omega < 1$ and $2A \geq 1$, so equation 30, applied with $B = 2A$, yields a two-input ReLU network $\mathsf{M}_{2A,\omega}$ satisfying

$$\sup_{|u|,|v| \leq 2A} |\mathsf{M}_{2A,\omega}(u,v) - uv| \leq \omega.$$

For each $r \in \{0, \ldots, q\}$, define the exact Horner tail

$$y_r : [0,1] \to \mathbb{R}, \qquad y_r(t) := \sum_{k=r}^{q} b_k t^{\,k-r}.$$

Thus,

$$y_q(t) = b_q, \qquad y_r(t) = b_r + t\, y_{r+1}(t), \quad r = q-1, \ldots, 0,$$

and, in particular,

$$y_0(t) = P(t).$$

We define the corresponding approximate Horner iterates recursively by

$$\widehat{y}_q(t) := b_q$$

and, for $r = q-1, \ldots, 0$,

$$\widehat{y}_r(t) := b_r + \mathsf{M}_{2A,\omega}\big(t, \widehat{y}_{r+1}(t)\big), \qquad 0 \leq t \leq 1.$$

Hence $\widehat{y}_r$ is obtained from the exact Horner recursion by replacing each multiplication $t\, y_{r+1}(t)$ with the approximate multiplication network $\mathsf{M}_{2A,\omega}$. We next verify simultaneously that the multiplication network is always evaluated on its prescribed domain and that the approximation errors do not accumulate excessively. First, for every $r \in \{0, \ldots, q\}$ and $t \in [0,1]$,

$$|y_r(t)| \leq \sum_{k=r}^{q} |b_k|\, t^{k-r} \leq \sum_{k=r}^{q} |b_k| \leq A - 1.$$

We claim that, for every $r \in \{0, \ldots, q\}$,

$$\sup_{t \in [0,1]} |\widehat{y}_r(t) - y_r(t)| \leq (q-r)\omega \tag{31}$$

and

$$\sup_{t \in [0,1]} |\widehat{y}_r(t)| \leq 2A. \tag{32}$$

We prove equation 31 and equation 32 simultaneously by backward induction on r. For $r = q$,

$$\widehat{y}_q(t) = y_q(t) = b_q,$$

so the error is zero and

$$|\widehat{y}_q(t)| = |b_q| \le A - 1 < 2A.$$

Now let $r \in \{0, \dots, q-1\}$ and suppose that both bounds hold at level $r+1$. In particular,

$$|\widehat{y}_{r+1}(t)| \le 2A, \qquad t \in [0,1].$$

Since $A \ge 1$, we also have $|t| \le 1 \le 2A$. Hence

$$\big(t, \widehat{y}_{r+1}(t)\big) \in [-2A, 2A]^2,$$

so the approximation guarantee equation 30 applies and gives

$$\big|\mathsf{M}_{2A,\omega}\big(t, \widehat{y}_{r+1}(t)\big) - t\,\widehat{y}_{r+1}(t)\big| \le \omega.$$

Using the exact and approximate Horner recursions,

$$\begin{aligned} |\widehat{y}_r(t) - y_r(t)| &\le \big|\mathsf{M}_{2A,\omega}\big(t, \widehat{y}_{r+1}(t)\big) - t\,\widehat{y}_{r+1}(t)\big| \\ &\quad + |t|\,|\widehat{y}_{r+1}(t) - y_{r+1}(t)| \\ &\le \omega + (q - r - 1)\omega \\ &= (q-r)\omega. \end{aligned}$$

Thus equation 31 holds at level r. Moreover, since $q - r \le q$ and

$$q\omega = \frac{q}{q+1}\xi < \xi < 1,$$

the bound $|y_r(t)| \le A - 1$ yields

$$|\widehat{y}_r(t)| \le |y_r(t)| + |\widehat{y}_r(t) - y_r(t)| \le (A-1) + \xi < A \le 2A.$$

Hence equation 32 also holds at level r, completing the induction. In particular, since $y_0 = P$,

$$\|\widehat{y}_0 - P\|_{\infty,[0,1]} \le q\omega < \xi.$$

It remains to identify the approximate Horner recursion with a ReLU network. Define

$$\widetilde{P}(t) := \widehat{y}_0(t), \qquad t \in [0,1].$$

By construction, $\widehat{y}_0$ is obtained by successively composing the q multiplication networks $\mathsf{M}_{2A,\omega}$ with the affine updates

$$v \mapsto b_r + v, \qquad r = q-1, \dots, 0.$$

Moreover, since

$$z = \rho(z) - \rho(-z),$$

the input t and the current approximate Horner iterate can be carried through the successive multiplication blocks using only a universal number of additional neurons per layer. Consequently, $\widetilde{P} = \widehat{y}_0$ admits a ReLU representation satisfying

$$\widetilde{P} \in \mathcal{F}\left(C(q+1)\log\frac{e(q+1)A}{\xi}, C\right), \tag{33}$$

where $C > 0$ is a universal constant. Since $y_0 = P$ and $\widetilde{P} = \widehat{y}_0$, the preceding Horner error bound gives

$$\|\widetilde{P} - P\|_{\infty,[0,1]} = \|\widehat{y}_0 - y_0\|_{\infty,[0,1]} \le q\omega < \xi. \tag{34}$$

Piecewise-polynomial approximation. We next extend the polynomial construction in equation 33 to continuous piecewise-polynomial functions. The main idea is to decompose such a function into a global piecewise-linear interpolant and polynomial residuals supported on the individual subintervals. This will later allow the residual corrections to be evaluated sequentially while keeping the network width uniformly bounded.

Let $N \in \mathbb{N}$ and let

$$a = x_0 < x_1 < \cdots < x_N = b$$

be a partition of the interval $[a, b]$. Suppose that $P : [a, b] \to \mathbb{R}$ is continuous and that, on each subinterval $[x_{r-1}, x_r]$, $r \in \{1, \ldots, N\}$, the function P is a polynomial of degree at most q. More precisely, introduce the local coordinate

$$s = \frac{t - x_{r-1}}{x_r - x_{r-1}} \in [0, 1], \qquad t = x_{r-1} + (x_r - x_{r-1})s,$$

and define the corresponding rescaled polynomial

$$P_r(s) := P\big(x_{r-1} + (x_r - x_{r-1})s\big), \qquad s \in [0, 1].$$

Assume that

$$P_r(s) = \sum_{k=0}^{q} b_{r,k} s^k, \qquad \sum_{k=0}^{q} |b_{r,k}| \le A, \qquad r = 1, \ldots, N, \tag{35}$$

for some constant $A \ge 1$ that is uniform over the subintervals.

We first construct the continuous piecewise-linear function that interpolates P at the partition points $x_0, \ldots, x_N$. For $r = 1, \ldots, N$, define the slope of the secant line over $[x_{r-1}, x_r]$ by

$$d_r := \frac{P(x_r) - P(x_{r-1})}{x_r - x_{r-1}}.$$

Recall that $\rho(u) = \max\{u, 0\}$ denotes the ReLU activation. Define

$$I_P(t) := P(x_0) + d_1(t - x_0) + \sum_{r=1}^{N-1} (d_{r+1} - d_r)\rho(t - x_r), \qquad t \in [a, b]. \tag{36}$$

Then I_P is precisely the piecewise-linear interpolant of the data

$$\big\{(x_r, P(x_r)) : r = 0, \ldots, N\big\}.$$

Indeed, define recursively

$$v_0(t) := P(x_0) + d_1(t - x_0)$$

and, for $r = 1, \ldots, N - 1$,

$$v_r(t) := v_{r-1}(t) + (d_{r+1} - d_r)\rho(t - x_r).$$

By construction,

$$v_r(t) = P(x_0) + d_1(t - x_0) + \sum_{j=1}^{r} (d_{j+1} - d_j)\rho(t - x_j),$$

and hence

$$v_{N-1}(t) = I_P(t).$$

It remains to record the ReLU-network complexity of this representation. At step r, the recursion only requires the original input t and the current value $v_{r-1}(t)$. Since every scalar quantity z can be propagated through a ReLU layer using the identity

$$z = \rho(z) - \rho(-z),$$

the pair $(t, v_{r-1}(t))$ can be carried through the successive updates using only a universal number of neurons per layer. Each update

$$v_{r-1}(t) \longmapsto v_{r-1}(t) + (d_{r+1} - d_r)\rho(t - x_r)$$

requires only a constant number of additional ReLU units. Consequently, the function I_P admits a scalar ReLU representation of universal width and depth at most CN; equivalently,

$$I_P \in \mathcal{F}(CN, C), \tag{37}$$

where $C > 0$ is a universal constant independent of N, q, A, the partition $\{x_r\}_{r=0}^{N}$, and the coefficients $\{b_{r,k}\}$.

Localization of the residuals. We separate the piecewise polynomial into a global linear interpolant and residual terms that vanish at the endpoints of their respective intervals. This makes each residual local after composition with the clipped coordinate below.

For each interval, define

$$R_r(s) = P_r(s) - (1-s)P(x_{r-1}) - sP(x_r).$$

Then

$$R_r(0) = R_r(1) = 0,$$

and, by equation 35,

$$\sum_{k=0}^{\max\{q,1\}} |[s^k]R_r(s)| \le 4A. \tag{38}$$

Also define

$$s_r(t) = \frac{\rho(t-x_{r-1}) - \rho(t-x_r)}{x_r - x_{r-1}}. \tag{39}$$

Then

$$s_r(t) = \begin{cases} 0, & t \le x_{r-1}, \\ \dfrac{t-x_{r-1}}{x_r - x_{r-1}}, & x_{r-1} < t < x_r, \\ 1, & t \ge x_r. \end{cases}$$

By equation 33–equation 34, for each r there exists a universal-width ReLU network R_r' such that

$$\|R_r' - R_r\|_{\infty,[0,1]} \le \frac{\xi}{2}$$

and

$$R_r' \in \mathcal{F}\left(C(q+1)\log\frac{e(q+1)A}{\xi}, C\right).$$

Set

$$\widetilde{R}_r(s) = R_r'(s) - (1-s)R_r'(0) - sR_r'(1).$$

Then

$$\widetilde{R}_r(0) = \widetilde{R}_r(1) = 0,$$

and

$$\begin{aligned} |\widetilde{R}_r(s) - R_r(s)| &\le |R_r'(s) - R_r(s)| \\ &\quad + (1-s)|R_r'(0) - R_r(0)| + s|R_r'(1) - R_r(1)| \\ &\le \xi. \end{aligned}$$

Hence

$$\|\widetilde{R}_r - R_r\|_{\infty,[0,1]} \le \xi. \tag{40}$$

Moreover, since both R_r and $\widetilde{R}_r$ vanish at 0 and 1,

$$R_r\{s_r(t)\} = \widetilde{R}_r\{s_r(t)\} = 0, \qquad t \notin (x_{r-1}, x_r). \tag{41}$$

For $t \in [x_{r-1}, x_r]$, by definition,

$$P(t) = I_P(t) + R_r\{s_r(t)\}.$$

Thus equation 41 gives

$$P(t) = I_P(t) + \sum_{r=1}^{N} R_r\{s_r(t)\}. \tag{42}$$

Define

$$\widetilde{P}(t) = I_P(t) + \sum_{r=1}^{N} \widetilde{R}_r\{s_r(t)\}.$$

For every fixed t, at most one summand in equation 42 is nonzero. Therefore, by equation 40,

$$\|\widetilde{P} - P\|_{\infty,[a,b]} \le \xi. \tag{43}$$

Serial evaluation. Rather than evaluating all localized residual networks in parallel, we add them successively. This makes the number of pieces contribute to the depth while the width remains universal.

To evaluate $\widetilde{P}$ with universal width, let

$$z_0(t) = I_P(t), \qquad z_r(t) = z_{r-1}(t) + \widetilde{R}_r\{s_r(t)\}, \quad r = 1, \dots, N.$$

At each step only t and z_{r-1} need to be carried. Combining equation 37 with equation 33 therefore gives

$$\widetilde{P} \in \mathcal{F}\left(CN(q+1)\log\frac{e(q+1)A}{\xi}, C\right). \tag{44}$$

Analytic polynomial approximation. We next obtain polynomial approximants whose degree and coefficient size can both be controlled. These are the two quantities needed in the ReLU construction above.

We next recall a standard polynomial-approximation bound for analytic functions. Map $[0,1]$ to $[-1,1]$ by $x = 2s - 1$ and define

$$\widetilde{f}(x) := f\left(\frac{x+1}{2}\right), \qquad x \in [-1,1].$$

Suppose that, for some $\rho > 1$, $\widetilde{f}$ extends holomorphically to the Bernstein ellipse E_ρ with foci $\{-1,1\}$ and satisfies

$$\sup_{z \in E_\rho} |\widetilde{f}(z)| \le M,$$

enlarging the upper bound so that $M \ge 1$. Then $\widetilde{f}$ admits a Chebyshev expansion

$$\widetilde{f}(x) = \sum_{k=0}^{\infty} c_k T_k(x),$$

where T_k is the kth Chebyshev polynomial of the first kind. Equivalently, returning to $s \in [0,1]$,

$$f(s) = \sum_{k=0}^{\infty} c_k T_k(2s-1).$$

Then Trefethen (2019, Theorems 8.1–8.2) gives

$$|c_k| \le 2M\rho^{-k}, \qquad \left\| f - \sum_{k=0}^{q} c_k T_k(2\,\cdot -1) \right\|_{\infty,[0,1]} \le \frac{2M}{\rho - 1}\rho^{-q}. \tag{45}$$

Recall that the Chebyshev polynomials of the first kind are characterized by

$$T_k(\cos\theta) = \cos(k\theta), \qquad k \in \mathbb{N}_0.$$

The trigonometric identity

$$\cos\{(k+1)\theta\} = 2\cos(\theta)\cos(k\theta) - \cos\{(k-1)\theta\}$$

therefore yields the standard three-term recurrence

$$T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x), \qquad k \ge 1,$$

with

$$T_0(x) = 1, \qquad T_1(x) = x.$$

We next control the coefficients after the change of variables $x = 2s - 1$. For a polynomial

$$Q(s) = \sum_{r \ge 0} a_r s^r,$$

write

$$[s^r]Q(s) := a_r$$

for the coefficient of s^r in Q, and define its coefficient ℓ_1-norm by

$$\|Q\|_{\mathrm{coef},1} := \sum_{r\geq 0} |a_r|.$$

Equivalently,

$$\|Q\|_{\mathrm{coef},1} = \sum_{r\geq 0} |[s^r]Q(s)|.$$

Let

$$A_k := \|T_k(2\,\cdot -1)\|_{\mathrm{coef},1} = \sum_{r=0}^{k} |[s^r]T_k(2s-1)|.$$

Since

$$T_0(2s-1) = 1, \qquad T_1(2s-1) = 2s-1,$$

we have

$$A_0 = 1, \qquad A_1 = 3.$$

Substituting $x = 2s-1$ into the Chebyshev recurrence gives

$$T_{k+1}(2s-1) = (4s-2)T_k(2s-1) - T_{k-1}(2s-1).$$

For every polynomial Q, multiplication by s only shifts its coefficients, and therefore

$$\|sQ\|_{\mathrm{coef},1} = \|Q\|_{\mathrm{coef},1}.$$

Consequently, by the triangle inequality,

$$\begin{aligned} A_{k+1} &\leq 4\|sT_k(2s-1)\|_{\mathrm{coef},1} + 2\|T_k(2s-1)\|_{\mathrm{coef},1} + \|T_{k-1}(2s-1)\|_{\mathrm{coef},1} \\ &= 6A_k + A_{k-1}. \end{aligned}$$

We now prove by induction that

$$A_k \leq 7^k, \qquad k \geq 0.$$

The claim holds for $k = 0, 1$, since $A_0 = 1$ and $A_1 = 3 \leq 7$. If it holds at levels k and $k-1$, then

$$A_{k+1} \leq 6 \cdot 7^k + 7^{k-1} = 43 \cdot 7^{k-1} \leq 49 \cdot 7^{k-1} = 7^{k+1}.$$

Hence, for every $k \geq 0$,

$$\sum_{r=0}^{k} |[s^r]T_k(2s-1)| \leq 7^k.$$

Next, define the degree-q truncated Chebyshev approximation by

$$p_q(s) := \sum_{k=0}^{q} c_k T_k(2s-1), \qquad s \in [0,1].$$

Since

$$[s^r]p_q(s) = \sum_{k=r}^{q} c_k [s^r]T_k(2s-1),$$

the triangle inequality and a change in the order of summation give

$$\begin{aligned} 1 + \sum_{r=0}^{q} |[s^r]p_q(s)| &\leq 1 + \sum_{r=0}^{q}\sum_{k=r}^{q} |c_k|\,|[s^r]T_k(2s-1)| \\ &= 1 + \sum_{k=0}^{q} |c_k| \sum_{r=0}^{k} |[s^r]T_k(2s-1)|. \end{aligned}$$

Using the bounds

$$|c_k| \le 2M\rho^{-k} \qquad \text{and} \qquad \sum_{r=0}^{k} |[s^r]T_k(2s-1)| \le 7^k,$$

we obtain

$$1 + \sum_{r=0}^{q} |[s^r]p_q(s)| \le 1 + 2M \sum_{k=0}^{q} \left(\frac{7}{\rho}\right)^k.$$

Since $\rho > 1$ is fixed, there exist constants $C > 0$ and $C_0 > 1$, depending only on ρ (equivalently, only on the fixed analytic neighborhood), such that

$$1 + 2M \sum_{k=0}^{q} \left(\frac{7}{\rho}\right)^k \le CMC_0^q.$$

Define

$$\vartheta := \rho^{-1} \in (0,1).$$

then

$$\|f - p_q\|_{\infty,[0,1]} \le \frac{2M}{\rho - 1}\rho^{-q} \le CM\vartheta^q,$$

where $C > 0$ depends only on the fixed value of ρ. Combining this with the coefficient bound above, we obtain

$$\|f - p_q\|_{\infty,[0,1]} \le CM\vartheta^q, \qquad 1 + \sum_{r=0}^{q} |[s^r]p_q(s)| \le CMC_0^q, \tag{46}$$

where $\vartheta \in (0,1)$, $C_0 > 1$, and $C > 0$ depend only on ρ, equivalently only on the fixed analytic neighborhood.

Replacing p_q by

$$p_q^0(s) = p_q(s) + (1-s)\{f(0) - p_q(0)\} + s\{f(1) - p_q(1)\}$$

gives

$$p_q^0(0) = f(0), \qquad p_q^0(1) = f(1),$$

and

$$\begin{aligned} |p_q^0(s) - f(s)| &\le |p_q(s) - f(s)| + (1-s)|p_q(0) - f(0)| \\ &\quad + s|p_q(1) - f(1)| \\ &\le 2\|p_q - f\|_{\infty,[0,1]}. \end{aligned}$$

The coefficient sum of p_q^0 is bounded by CMC_0^q after changing C.

Approximation of the logarithm. We use geometrically growing intervals so that, after rescaling each interval to $[0,1]$, the nonlinear part of $\log$ belongs to a uniformly holomorphic family.

We first apply these bounds to $\log$. Let

$$a_0 = \epsilon, \qquad a_{r+1} = \min\{2a_r, 1\},$$

and stop at $a_N = 1$. Then

$$N \le 1 + \log_2(1/\epsilon) \le C\Lambda_\epsilon. \tag{47}$$

For $t \in [a_r, a_{r+1}]$, write

$$t = a_r(1 + \lambda_r s), \qquad \lambda_r = \frac{a_{r+1} - a_r}{a_r}, \qquad 0 \le s \le 1.$$

Then

$$0 < \lambda_r \le 1, \qquad \log t = \log a_r + \log(1 + \lambda_r s).$$

Choose a universal constant $c \in (0, 1/2]$ and define

$$\mathcal{N} = \{z \in \mathbb{C} : \operatorname{dist}(z, [0,1]) < c\}.$$

If $z \in \mathcal{N}$, then $\mathrm{Re}(z) \geq -c$. Hence, for $0 \leq \lambda \leq 1$,

$$\mathrm{Re}(1 + \lambda z) = 1 + \lambda \,\mathrm{Re}(z) \geq 1 - \lambda c \geq 1 - c \geq \frac{1}{2}.$$

Thus $1 + \lambda z$ remains uniformly inside the right half-plane, where the principal logarithm is holomorphic. Moreover, $|z| \leq 1 + c$, so

$$\frac{1}{2} \leq |1 + \lambda z| \leq 2 + c.$$

Therefore the family $z \mapsto \log(1 + \lambda z)$ is uniformly bounded on $\mathcal{N}$:

$$\sup_{\substack{z \in \mathcal{N} \\ 0 \leq \lambda \leq 1}} |\log(1 + \lambda z)| \leq C, \tag{48}$$

for a universal constant $C > 0$.

For each $r \in \{0, \ldots, N-1\}$ and $t \in [a_r, a_{r+1}]$, introduce the rescaled coordinate

$$s = \frac{t - a_r}{a_{r+1} - a_r} \in [0, 1]$$

and set

$$\lambda_r := \frac{a_{r+1} - a_r}{a_r} \in (0, 1].$$

Then

$$t = a_r(1 + \lambda_r s)$$

and therefore

$$\log t = \log a_r + h_r(s), \qquad h_r(s) := \log(1 + \lambda_r s). \tag{49}$$

By equation 48, the family $\{h_r\}_{r=0}^{N-1}$ is holomorphic and uniformly bounded on a common complex neighborhood of $[0, 1]$. Hence equation 46 applies uniformly in r: there exist constants $\vartheta \in (0, 1)$ and $C_0 > 1$, independent of r, such that for every $q \in \mathbb{N}_0$ there is a polynomial $p_{r,q}$ of degree at most q satisfying

$$\|h_r - p_{r,q}\|_{\infty,[0,1]} \leq C\vartheta^q, \qquad 1 + \sum_{k=0}^{q} |[s^k]p_{r,q}(s)| \leq CC_0^q. \tag{50}$$

Choose

$$q = \lceil C\Lambda_\xi \rceil$$

with the universal constant $C > 0$ sufficiently large that $C\vartheta^q \leq \xi/4$. Thus

$$\|h_r - p_{r,q}\|_{\infty,[0,1]} \leq \frac{\xi}{4}$$

uniformly in r. Define the endpoint-corrected polynomial

$$\widetilde{p}_{r,q}(s) := p_{r,q}(s) + (1 - s)\{h_r(0) - p_{r,q}(0)\} + s\{h_r(1) - p_{r,q}(1)\}.$$

Then

$$\widetilde{p}_{r,q}(0) = h_r(0), \qquad \widetilde{p}_{r,q}(1) = h_r(1),$$

and, by the same endpoint-correction argument used above,

$$\|\widetilde{p}_{r,q} - h_r\|_{\infty,[0,1]} \leq \frac{\xi}{2}.$$

Consequently, the local polynomials

$$Q_r(s) := \log a_r + \widetilde{p}_{r,q}(s)$$

match $\log t$ at the endpoints of $[a_r, a_{r+1}]$. Hence the functions

$$P_{\epsilon,\xi}^{\log}(t) := Q_r\left(\frac{t - a_r}{a_{r+1} - a_r}\right), \qquad t \in [a_r, a_{r+1}],$$

fit together to form a continuous piecewise polynomial on $[\epsilon, 1]$ satisfying

$$\|P^{\log}_{\epsilon,\xi} - \log\|_{\infty,[\epsilon,1]} \le \frac{\xi}{2}. \tag{51}$$

We next bound the coefficients of the local polynomials Q_r. Define

$$e_{r,0} := h_r(0) - p_{r,q}(0), \qquad e_{r,1} := h_r(1) - p_{r,q}(1).$$

Since

$$\|h_r - p_{r,q}\|_{\infty,[0,1]} \le \frac{\xi}{4},$$

we have

$$|e_{r,0}| \le \frac{\xi}{4}, \qquad |e_{r,1}| \le \frac{\xi}{4}.$$

The endpoint-corrected polynomial can be written as

$$\widetilde{p}_{r,q}(s) = p_{r,q}(s) + e_{r,0} + (e_{r,1} - e_{r,0})s.$$

Thus the correction is affine, and the sum of the absolute values of its coefficients is bounded by

$$|e_{r,0}| + |e_{r,1} - e_{r,0}| \le 2|e_{r,0}| + |e_{r,1}| \le \frac{3\xi}{4}.$$

Consequently, by the triangle inequality for polynomial coefficients and equation 50,

$$\begin{aligned} 1 + \sum_{k=0}^{q} |[s^k]\widetilde{p}_{r,q}(s)| &\le 1 + \sum_{k=0}^{q} |[s^k]p_{r,q}(s)| + \frac{3\xi}{4} \\ &\le CC_0^q, \end{aligned}$$

after enlarging the universal constant C, since $\xi < 1$ and $C_0^q \ge 1$.

Moreover, since $\epsilon \le a_r \le 1$,

$$|\log a_r| = \log(1/a_r) \le \log(1/\epsilon) \le \Lambda_\epsilon.$$

Thus the coefficient ℓ_1-norm of every Q_r is bounded uniformly by

$$A \le C(1 + \Lambda_\epsilon)C_0^q. \tag{52}$$

Applying equation 43–equation 44 to $P^{\log}_{\epsilon,\xi}$ with network-approximation error $\xi/2$ therefore yields a scalar ReLU network

$$G_{\epsilon,\xi} \in \mathcal{F}\left(CN(q+1)\log\frac{e(q+1)A}{\xi}, C\right)$$

such that, by equation 51,

$$\|G_{\epsilon,\xi} - \log\|_{\infty,[\epsilon,1]} \le \xi. \tag{53}$$

Finally, equation 47, $q + 1 \lesssim \Lambda_\xi$, and equation 52 imply

$$\begin{aligned} N(q+1)\log\frac{e(q+1)A}{\xi} &\le C\Lambda_\epsilon\Lambda_\xi\left\{\Lambda_\xi + \log(1+\Lambda_\epsilon)\right\} \\ &\le C\Lambda_\epsilon\Lambda_\xi(\Lambda_\epsilon + \Lambda_\xi). \end{aligned}$$

Hence

$$G_{\epsilon,\xi} \in \mathcal{F}\left(C\Lambda_\epsilon\Lambda_\xi(\Lambda_\epsilon + \Lambda_\xi), C\right).$$

Construction of the logit network. Since $\ell(t) = \log t - \log(1-t)$, the logit approximator is obtained from two copies of the logarithm network evaluated in parallel.

Let $\eta \in (0,1)$ and set

$$\mathsf{L}_{\epsilon,\eta}(t) = G_{\epsilon,\eta/2}(t) - G_{\epsilon,\eta/2}(1-t).$$

For $t \in [\epsilon, 1-\epsilon]$,

$$\begin{aligned} |\mathsf{L}_{\epsilon,\eta}(t) - \ell(t)| &\le |G_{\epsilon,\eta/2}(t) - \log t| \\ &\quad + |G_{\epsilon,\eta/2}(1-t) - \log(1-t)| \\ &\le \eta. \end{aligned}$$

Evaluating the two universal-width networks in parallel changes the width only by a universal constant. Since $\Lambda_{\eta/2} \le C\Lambda_\eta$,

$$\mathsf{L}_{\epsilon,\eta} \in \mathcal{F}\left(C\Lambda_\epsilon\Lambda_\eta(\Lambda_\epsilon + \Lambda_\eta), C\right),$$

which proves equation 25–equation 26.

Approximation of inverse softplus below one. Near zero, ψ has the same logarithmic singularity as $\log t$. We therefore separate this term from a remainder that is uniformly analytic after rescaling.

It remains to treat ψ. For $0 < t \le 1$,

$$\psi(t) = \log t + r(t), \qquad r(t) = \log\left(\frac{e^t - 1}{t}\right), \qquad r(0) = 0. \tag{54}$$

With the removable singularity at zero filled in,

$$\frac{e^z - 1}{z} = 0 \quad \Longleftrightarrow \quad z = 2\pi i k, \qquad k \in \mathbb{Z} \setminus \{0\}.$$

Thus, after decreasing the universal constant c defining $\mathcal{N}$ if necessary,

$$\frac{e^z - 1}{z} \neq 0$$

on a simply connected neighborhood containing

$$\{b(1+\lambda z) : z \in \mathcal{N},\ 0 < b \le 1,\ 0 \le \lambda \le 1,\ b(1+\lambda) \le 1\}.$$

Let

$$h(w) := \begin{cases} \dfrac{e^w - 1}{w}, & w \neq 0, \\ 1, & w = 0. \end{cases}$$

By the preceding zero-free argument, after decreasing c if necessary, there exists a fixed simply connected open set D containing

$$K := \left\{b(1+\lambda z) : z \in \overline{\mathcal{N}},\ 0 \le b \le 1,\ 0 \le \lambda \le 1,\ b(1+\lambda) \le 1\right\},$$

such that h has no zeros on D. Hence h admits a holomorphic logarithm r on D, chosen to agree with $r(t) = \log\{(e^t - 1)/t\}$ for $t > 0$. Since K is compact and r is continuous on K, there is a universal constant $C > 0$ such that

$$\sup_{\substack{z \in \mathcal{N} \\ 0 < b \le 1,\ 0 \le \lambda \le 1 \\ b(1+\lambda) \le 1}} |r\{b(1+\lambda z)\}| \le C. \tag{55}$$

Partition $[\delta, \min\{U, 1\}]$ geometrically as follows. Set

$$b_0 = \delta, \qquad b_{j+1} = \min\{2b_j, \min\{U, 1\}\},$$

and stop at the first index N_0 for which $b_{N_0} = \min\{U, 1\}$. Then every interval $[b_j, b_{j+1}]$ satisfies

$$1 < \frac{b_{j+1}}{b_j} \le 2,$$

except possibly for equality at the final endpoint. Since $\min\{U, 1\} \le 1$,

$$N_0 \le 1 + \log_2(1/\delta) \le C\Lambda_\delta, \qquad \Lambda_\delta = \log(e/\delta).$$

On each interval $[b_j, b_{j+1}]$ of the geometric partition, write

$$t = b(1+\lambda s), \qquad b = b_j, \qquad \lambda = \frac{b_{j+1} - b_j}{b_j} \in (0, 1], \qquad s \in [0, 1].$$

Using the decomposition

$$\psi(t) = \log t + r(t),$$

we obtain

$$\psi\{b(1+\lambda s)\} = \log b + \log(1+\lambda s) + r\{b(1+\lambda s)\}. \tag{56}$$

The term $\log b$ is constant in s. By equation 48, the functions $s \mapsto \log(1+\lambda s)$ are holomorphic and uniformly bounded on the same complex neighborhood, uniformly over $\lambda \in (0, 1]$. Likewise, by equation 55, the functions $s \mapsto r\{b(1+\lambda s)\}$ are holomorphic and uniformly bounded there, uniformly over all admissible b and λ. Hence the nonconstant part of $\psi\{b(1+\lambda s)\}$ satisfies the assumptions of equation 46 with constants independent of the interval index.

Approximation of inverse softplus above one. For $t \geq 1$, we separate the linear growth of ψ from a bounded remainder. Since

$$e^t - 1 = e^t(1 - e^{-t}),$$

we may write

$$\psi(t) = \log(e^t - 1) = t + r_+(t), \qquad r_+(t) := \log(1 - e^{-t}). \tag{57}$$

The function r_+ is bounded on $[1, \infty)$ and converges to zero as $t \to \infty$.

When $U > 1$, partition $[1, U]$ into consecutive intervals $[b_j, b_j + h_j]$, $j = 1, \ldots, N_1$, with

$$b_j \geq 1, \qquad 0 < h_j \leq 1.$$

For example, one may use intervals of length one, except possibly for the final interval. Their number satisfies

$$N_1 \leq \lceil U - 1 \rceil \leq 1 + U. \tag{58}$$

On the jth interval, introduce the local coordinate $u \in [0, 1]$ by

$$t = b_j + h_j u.$$

Then

$$\psi(b_j + h_j u) = b_j + h_j u + r_+(b_j + h_j u),$$

where the first two terms are affine in u and only the bounded remainder $r_+(b_j + h_j u)$ requires polynomial approximation.

For $z \in \mathcal{N}$, we have $\mathrm{Re}(z) \geq -c$. Hence, for $b \geq 1$ and $0 < h \leq 1$,

$$\mathrm{Re}(b + hz) = b + h\,\mathrm{Re}(z) \geq 1 - c.$$

Choosing $c \leq 1/4$ gives

$$\mathrm{Re}(b + hz) \geq \frac{3}{4},$$

and therefore

$$|e^{-(b+hz)}| = e^{-\mathrm{Re}(b+hz)} \leq e^{-3/4} < 1.$$

Consequently,

$$1 - e^{-(b+hz)} \in \{w \in \mathbb{C} : |w - 1| \leq e^{-3/4}\},$$

a fixed disk contained in the right half-plane and bounded away from zero. Hence the principal logarithm is holomorphic on this disk, so

$$z \mapsto r_+(b + hz) = \log\{1 - e^{-(b+hz)}\}$$

is holomorphic on $\mathcal{N}$. Since the disk is compactly contained in the domain of the logarithm, there exists a universal constant $C > 0$ such that

$$\sup_{\substack{z \in \mathcal{N} \\ b \geq 1,\ 0 < h \leq 1}} |r_+(b + hz)| \leq C. \tag{59}$$

Assembly of the inverse-softplus network. Let N_0 denote the number of intervals in the geometric partition of $[\delta, \min\{U, 1\}]$ constructed above, with $N_0 = 0$ if this interval is degenerate. By construction,

$$N_0 \leq C\Lambda_\delta.$$

Similarly, when $U > 1$, let N_1 denote the number of intervals $[b_j, b_j + h_j]$ in the partition of $[1, U]$ constructed above, where $b_j \geq 1$ and $0 < h_j \leq 1$; set $N_1 = 0$ when $U \leq 1$. By equation 58,

$$N_1 \leq 1 + U.$$

The two collections of intervals together form a partition of $[\delta, U]$ (with the common endpoint 1 shared when $U > 1$). Therefore, if N denotes the total number of intervals,

$$N = N_0 + N_1 \leq C\Lambda_\delta + (1 + U) \leq C(1 + U + \Lambda_\delta). \tag{60}$$

On every interval below 1, the rescaled function has the form (56). The preceding argument shows that $s \mapsto \log(1 + \lambda s)$ is holomorphic and uniformly bounded on a fixed complex neighborhood of

$[0, 1]$, uniformly over $0 < \lambda \le 1$. Moreover, by equation 55, $s \mapsto r\{b(1 + \lambda s)\}$ is holomorphic and uniformly bounded on the same neighborhood, uniformly over all admissible b and λ. On every interval above 1, writing $t = b + hs$ gives

$$\psi(b + hs) = b + hs + r_+(b + hs),$$

where r_+ is holomorphic and uniformly bounded by equation 59. Hence equation 46 applies on every interval with constants independent of the interval.

Choose

$$q = \lceil C\Lambda_\zeta \rceil$$

with $C > 0$ sufficiently large that the degree-q polynomial approximation error on each interval is at most $\zeta/4$. After the endpoint correction, the error on each interval is at most $\zeta/2$, and adjacent polynomial pieces agree at their common endpoints. Thus they define a continuous piecewise polynomial P_ψ on $[\delta, U]$ satisfying

$$\|P_\psi - \psi\|_{\infty,[\delta,U]} \le \frac{\zeta}{2}. \tag{61}$$

We next bound the coefficients of its local polynomial pieces. For intervals below 1, the constant term satisfies

$$|\log b| \le \log(1/\delta) \le \Lambda_\delta,$$

while the remaining analytic terms have coefficient ℓ_1-norm at most CC_0^q. For intervals above 1, the affine term $b + hs$ has coefficient ℓ_1-norm $b + h \le U + 1$, while the polynomial approximation to $r_+(b + hs)$ again has coefficient ℓ_1-norm at most CC_0^q. The endpoint corrections contribute only $O(\zeta)$ to these coefficient sums. Consequently, uniformly over all local pieces,

$$A \le C(1 + U + \Lambda_\delta)C_0^q. \tag{62}$$

Applying equation 43–equation 44 to P_ψ with network-approximation error $\zeta/2$ yields a universal-width ReLU network $\mathsf{S}_{\delta,U,\zeta}$ satisfying

$$\|\mathsf{S}_{\delta,U,\zeta} - \psi\|_{\infty,[\delta,U]} \le \zeta. \tag{63}$$

Its depth satisfies

$$\begin{aligned} D_\psi(\delta, U, \zeta) &\le CN(q + 1) \log \frac{e(q + 1)A}{\zeta} \\ &\le C(1 + U + \Lambda_\delta)\Lambda_\zeta \log \left(\frac{e\Lambda_\zeta(1 + U + \Lambda_\delta)C_0^{C\Lambda_\zeta}}{\zeta} \right) \\ &\le C(1 + U + \Lambda_\delta)\Lambda_\zeta(1 + U + \Lambda_\delta + \Lambda_\zeta). \end{aligned}$$

Here the first inequality follows from the piecewise-polynomial network bound equation 44, applied with network-approximation error $\zeta/2$ (with the resulting numerical factor absorbed into C). The second inequality uses equation 60, the choice $q = \lceil C\Lambda_\zeta \rceil$, which implies $q + 1 \le C\Lambda_\zeta$, and the coefficient bound equation 62; in particular, $C_0^q \le C_0^{C\Lambda_\zeta}$. Finally, the last inequality follows by expanding the logarithm and using

$$\log(1/\zeta) \le \Lambda_\zeta, \qquad \log \Lambda_\zeta \le \Lambda_\zeta, \qquad \log(1 + U + \Lambda_\delta) \le 1 + U + \Lambda_\delta,$$

together with the fact that $C_0 > 1$ is a fixed constant. This proves equation 27–equation 28.

Endpoint bounds. Finally, since $e^\delta - 1 \ge \delta$ and $e^U - 1 \le e^U$,

$$-\psi(\delta) = \log \frac{1}{e^\delta - 1} \le \log(1/\delta), \qquad \psi(U) = \log(e^U - 1) \le U.$$

This proves equation 29. Enlarging the universal width constant appearing above, if necessary, defines w_{cal} and completes the proof. □

Lemma 5 (Parallelization, composition, and maximum of ReLU networks)**.** *Let $f_s : \mathbb{R}^d \to \mathbb{R}$, $s = 1, \dots, r$, be ReLU networks with depths L_s and widths w_s. Put*

$$L_{\max} = \max_{s \le r} L_s, \qquad w_\Sigma = \sum_{s=1}^{r} (w_s + 1).$$

Then for a constant $C > 0$ that can change from line to line:

1. *$(f_1, \dots, f_r)$ has a ReLU representation of depth at most $L_{\max} + C$ and width at most Cw_Σ;*
2. *if $q : \mathbb{R}^r \to \mathbb{R}$ has depth L_q and width w_q, then $q \circ (f_1, \dots, f_r)$ has depth at most $L_{\max} + L_q + C$ and width at most $C \max\{w_\Sigma, w_q\}$;*
3. *$\max_{s \le r} f_s$ is represented exactly with depth at most $L_{\max} + \lceil \log_2 r \rceil + C$ and width at most Cw_Σ.*

Proof. The first two statements follow from the parallelization and composition constructions in the proof of Kohler & Langer (2021, Theorem 3). To equalize the depths, use

$$z = \rho(z) - \rho(-z). \tag{64}$$

Thus, each f_s can be padded to depth $L_{\max}$ with width at most $C(w_s + 1)$. Placing the padded networks in parallel gives

$$L_{\boldsymbol{f}} \le L_{\max} + C, \qquad w_{\boldsymbol{f}} \le C \sum_{s=1}^{r} (w_s + 1) = Cw_\Sigma,$$

where $\boldsymbol{f} = (f_1, \dots, f_r)$. Composing this representation with q gives

$$L_{q \circ \boldsymbol{f}} \le L_{\max} + L_q + C, \qquad w_{q \circ \boldsymbol{f}} \le C \max\{w_\Sigma, w_q\},$$

which proves the first two statements.

For the third statement, let

$$R = 2^{\lceil \log_2 r \rceil}$$

and, if $R > r$, set $f_{r+1} = \cdots = f_R = f_r$. Define

$$v_s^{(0)} = f_s, \qquad s = 1, \dots, R,$$

and recursively, for $t = 1, \dots, \log_2 R$,

$$v_s^{(t)} = \rho\left(v_{2s-1}^{(t-1)} - v_{2s}^{(t-1)}\right) + v_{2s}^{(t-1)}, \qquad s = 1, \dots, \frac{R}{2^t}. \tag{65}$$

Since

$$\rho(a - b) + b = \max\{a, b\},$$

induction gives

$$v_s^{(t)} = \max_{(s-1)2^t < k \le s2^t} f_k,$$

and hence

$$v_1^{(\log_2 R)} = \max_{1 \le s \le r} f_s.$$

Using

$$b = \rho(b) - \rho(-b),$$

each update in equation 65 is represented by one ReLU layer with at most a constant number of units per active value. At level t there are $R/2^t$ active values, and therefore

$$w_t \le C \frac{R}{2^t} \le CR \le 2Cr \le Cw_\Sigma.$$

Since

$$\log_2 R = \lceil \log_2 r \rceil,$$

the resulting network has depth at most

$$L_{\max} + \lceil \log_2 r \rceil + C$$

and width at most Cw_Σ. □

Definition 3 (Construction of the latent ReLU networks)**.** *Let $F^*_{A_1}, \dots, F^*_{A_m}$ be the target probability functions and let $h_1, \dots, h_m$ be $[0,1]$-valued ReLU networks. Fix*

$$0 < \epsilon < \frac{1}{4}, \qquad 0 < \eta < 1, \qquad 0 < \delta \le 1, \qquad 0 < \zeta < 1.$$

Define

$$F^*_{A_j,\epsilon} = \epsilon + (1-2\epsilon)F^*_{A_j}, \qquad \widetilde{h}_j = \epsilon + (1-2\epsilon)h_j, \qquad \lambda_\epsilon = \log\frac{1-\epsilon}{\epsilon}. \tag{66}$$

Let $\mathsf{L}_{\epsilon,\eta}$ be the logit approximation from Lemma 4, and set

$$q_j = \mathsf{L}_{\epsilon,\eta} \circ \widetilde{h}_j. \tag{67}$$

Define the DAG level recursively by

$$r_j = 0, \quad j \in \mathcal{R}, \qquad r_j = 1 + \max_{k\in\mathcal{P}(j)} r_k, \quad j \notin \mathcal{R}. \tag{68}$$

Then, with the notation of Theorem 2,

$$r_j \le \kappa_j - 1 \le \kappa_m - 1.$$

Define

$$\bar{q}_j = q_j, \quad j \in \mathcal{R}, \qquad \bar{q}_j = \max\left\{q_j, \max_{k\in\mathcal{P}(j)}(\bar{q}_k + \delta)\right\}, \quad j \notin \mathcal{R}. \tag{69}$$

For $j \notin \mathcal{R}$, set

$$u_j = \bar{q}_j - \max_{k\in\mathcal{P}(j)} \bar{q}_k, \qquad U_{\epsilon,\delta,\eta} = 2\lambda_\epsilon + 2\eta + \delta\kappa_m. \tag{70}$$

Let $\mathsf{S}_{\delta,U,\zeta}$ be the inverse-softplus approximation from Lemma 4, and define

$$g_j = \bar{q}_j, \quad j \in \mathcal{R}, \qquad g_j = \mathsf{S}_{\delta,U_{\epsilon,\delta,\eta},\zeta} \circ u_j, \quad j \notin \mathcal{R}. \tag{71}$$

Theorem 3 (Probability-scale ReLU approximation)**.** *Suppose that*

$$F^*_{A_k}(x) \le F^*_{A_j}(x), \qquad (k,j) \in E, \quad x \in [0,1]^d,$$

and that, for some $a \ge 0$, with the notation in Definition 3, we have

$$\max_{j\le m} \|h_j - F^*_{A_j}\|_\infty \le a. \tag{72}$$

Let $\boldsymbol{g} = (g_1, \dots, g_m)$ be the networks constructed in Definition 3. Then

$$\|H_j(\boldsymbol{g}) - F^*_{A_j}\|_\infty \le \epsilon + a + \frac{1}{4}\left(\eta + r_j\delta + r_j\zeta\right), \qquad j \in [m], \tag{73}$$

and therefore

$$\max_{j\le m} \|H_j(\boldsymbol{g}) - F^*_{A_j}\|_\infty \le \epsilon + a + \frac{1}{4}\left(\eta + \kappa_m\delta + \kappa_m\zeta\right). \tag{74}$$

Moreover,

$$\max_{j\le m} \|g_j\|_\infty \le C\left\{\lambda_\epsilon + \eta + \delta\kappa_m + \log(1/\delta) + \zeta + 1\right\}. \tag{75}$$

If the h_j have depth at most L and width at most w, and

$$\deg^-_{\max,m} = \max_{j\le m} |\mathcal{P}(j)|, \tag{76}$$

where $\deg^-_{\max,m}$ is the maximum in-degree of the DAG. Then there exist integers $L_g, w_g \ge 1$ such that, after padding with identity layers and zero weights if necessary, every latent network g_j, $j \in [m]$, admits a representation in the common dense ReLU class

$$g_j \in \mathcal{F}(L_g, w_g), \qquad j \in [m].$$

Moreover, the common depth and width may be chosen so that

$$L_g \le L + D_\ell(\epsilon,\eta) + D_\psi(\delta, U_{\epsilon,\delta,\eta}, \zeta) + C\kappa_m\left\lceil \log_2(\deg^-_{\max,m} + 2)\right\rceil + C, \tag{77}$$

$$w_g \le C a_{\max,m}(\deg^-_{\max,m} + 2)\max\{w, w_{\mathrm{cal}}, 1\}. \tag{78}$$

Here $C > 0$ is a universal constant independent of the node index j.

Proof. By equation 66 and equation 72,

$$\|\widetilde{h}_j - F^*_{A_j,\epsilon}\|_\infty = (1-2\epsilon)\|h_j - F^*_{A_j}\|_\infty \le a, \tag{79}$$

$$\widetilde{h}_j(x) \in [\epsilon, 1-\epsilon]. \tag{80}$$

Lemma 4 therefore gives

$$\|q_j - \ell \circ \widetilde{h}_j\|_\infty \le \eta. \tag{81}$$

Since 80 holds, because the logit function

$$\ell(t) = \log \frac{t}{1-t}$$

is increasing on $(0,1)$ and

$$\ell(\epsilon) = -\lambda_\epsilon, \qquad \ell(1-\epsilon) = \lambda_\epsilon,$$

we obtain

$$-\lambda_\epsilon \le \ell(\widetilde{h}_j) \le \lambda_\epsilon.$$

equation equation 81 implies

$$-\lambda_\epsilon - \eta \le q_j \le \lambda_\epsilon + \eta. \tag{82}$$

We next show that

$$-\lambda_\epsilon - \eta \le \bar{q}_j \le \lambda_\epsilon + \eta + r_j\delta. \tag{83}$$

For $j \in \mathcal{R}$, the claim follows directly from equation 82. Now let $j \notin \mathcal{R}$ and assume the bound holds for all parents $k \in \mathcal{P}(j)$. Then by equation 83,

$$\bar{q}_k \le \lambda_\epsilon + \eta + r_k\delta.$$

Since

$$r_j = 1 + \max_{k\in\mathcal{P}(j)} r_k,$$

we have $r_k + 1 \le r_j$ for every parent k, and therefore

$$\bar{q}_k + \delta \le \lambda_\epsilon + \eta + r_j\delta. \tag{84}$$

Since

$$q_j \le \lambda_\epsilon + \eta \le \lambda_\epsilon + \eta + r_j\delta,$$

and, for every $k \in \mathcal{P}(j)$ we have that (84) holds, then every term in the maximum defining $\bar{q}_j$ is bounded above by $\lambda_\epsilon + \eta + r_j\delta$. Hence

$$\bar{q}_j \le \lambda_\epsilon + \eta + r_j\delta.$$

Moreover, since $\bar{q}_j \ge q_j$, equation 82 gives

$$\bar{q}_j \ge -\lambda_\epsilon - \eta.$$

This completes the induction and proves equation 83.

For $j \notin \mathcal{R}$, equation 69 gives

$$\bar{q}_j \ge \max_{k\in\mathcal{P}(j)} \bar{q}_k + \delta,$$

and hence $u_j \ge \delta$. Moreover, by equation 83,

$$\begin{aligned} u_j &= \bar{q}_j - \max_{k\in\mathcal{P}(j)} \bar{q}_k \\ &\le \lambda_\epsilon + \eta + r_j\delta - (-\lambda_\epsilon - \eta) \\ &= 2\lambda_\epsilon + 2\eta + r_j\delta \le U_{\epsilon,\delta,\eta}, \end{aligned} \tag{85}$$

where the last inequality follows from (70). Thus

$$u_j \in [\delta, U_{\epsilon,\delta,\eta}], \qquad j \notin \mathcal{R}. \tag{86}$$

Set

$$\bar{F}_j = \tau(\bar{q}_j).$$

Since $\tau \circ \ell$ is the identity on $(0,1)$ and $\|\tau'\|_\infty \le 1/4$,

$$\begin{aligned}\|\tau(q_j) - F^*_{A_j,\epsilon}\|_\infty &\le \|\tau(q_j) - \tau(\ell \circ \widetilde{h}_j)\|_\infty + \|\widetilde{h}_j - F^*_{A_j,\epsilon}\|_\infty \\ &\le \frac{1}{4}\|q_j - \ell \circ \widetilde{h}_j\|_\infty + a \\ &\le \frac{\eta}{4} + a, \end{aligned} \tag{87}$$

where the second inequality follows from (79), and the last from (81).

For every edge $(k,j) \in E$,

$$F^*_{A_j,\epsilon} - F^*_{A_k,\epsilon} = (1-2\epsilon)\left(F^*_{A_j} - F^*_{A_k}\right) \ge 0. \tag{88}$$

We now claim that

$$F^*_{A_j,\epsilon} - a - \frac{\eta}{4} \le \bar{F}_j \le F^*_{A_j,\epsilon} + a + \frac{\eta}{4} + \frac{r_j\delta}{4}. \tag{89}$$

Since $\bar{q}_j \ge q_j$ and τ is increasing, from (87),

$$\bar{F}_j \ge \tau(q_j) \ge F^*_{A_j,\epsilon} - a - \frac{\eta}{4}.$$

We prove the upper bound in (89) by induction. For a root, the upper bound in (73) follows from equation 87. For $j \notin \mathcal{R}$, monotonicity of τ gives

$$\bar{F}_j = \max\left\{\tau(q_j),\ \max_{k\in\mathcal{P}(j)} \tau(\bar{q}_k + \delta)\right\}. \tag{90}$$

For every $k \in \mathcal{P}(j)$, the induction hypothesis for the upper bound in equation 89 gives

$$\tau(\bar{q}_k) = \bar{F}_k \le F^*_{A_k,\epsilon} + a + \frac{\eta}{4} + \frac{r_k\delta}{4}.$$

Since τ is $1/4$-Lipschitz, equation 88 holds, and $r_k + 1 \le r_j$, we obtain

$$\begin{aligned}\tau(\bar{q}_k + \delta) &\le \tau(\bar{q}_k) + \frac{\delta}{4} \\ &\le F^*_{A_k,\epsilon} + a + \frac{\eta}{4} + \frac{(r_k+1)\delta}{4} \\ &\le F^*_{A_j,\epsilon} + a + \frac{\eta}{4} + \frac{r_j\delta}{4}. \end{aligned}$$

Together with equation 87, this proves equation 89. Hence

$$\|\bar{F}_j - F^*_{A_j,\epsilon}\|_\infty \le a + \frac{1}{4}(\eta + r_j\delta). \tag{91}$$

Let

$$e_j = \|\widetilde{H}_j(\boldsymbol{g}) - \bar{q}_j\|_\infty.$$

For $j \in \mathcal{R}$,

$$e_j = 0.$$

For $j \notin \mathcal{R}$,

$$\widetilde{H}_j(\boldsymbol{g}) = \max_{k\in\mathcal{P}(j)} \widetilde{H}_k(\boldsymbol{g}) + \mathrm{sp}(g_j),$$

whereas, by equation 70,

$$\bar{q}_j = \max_{k\in\mathcal{P}(j)} \bar{q}_k + u_j.$$

Therefore, For $j \notin \mathcal{R}$, using the recursions for $\widetilde{H}_j(\boldsymbol{g})$ and $\bar{q}_j$, we have pointwise

$$\widetilde{H}_j(\boldsymbol{g}) - \bar{q}_j = \left\{\max_{k\in\mathcal{P}(j)} \widetilde{H}_k(\boldsymbol{g}) - \max_{k\in\mathcal{P}(j)} \bar{q}_k\right\} + \{\mathrm{sp}(g_j) - u_j\}.$$

Since, for any finite collections (a_k) and (b_k),

$$\left|\max_k a_k - \max_k b_k\right| \le \max_k |a_k - b_k|,$$

the triangle inequality and the definition $e_k = \|\widetilde{H}_k(\boldsymbol{g}) - \bar{q}_k\|_\infty$ give

$$e_j \le \max_{k\in\mathcal{P}(j)} e_k + \|\operatorname{sp}(g_j) - u_j\|_\infty.$$

Therefore,

$$\begin{aligned} e_j &\le \max_{k\in\mathcal{P}(j)} e_k + \|\operatorname{sp}(g_j) - u_j\|_\infty \\ &= \max_{k\in\mathcal{P}(j)} e_k + \left\|\operatorname{sp}\left\{\mathsf{S}_{\delta,U_{\epsilon,\delta,\eta},\zeta}(u_j)\right\} - \operatorname{sp}\{\psi(u_j)\}\right\|_\infty \\ &\le \max_{k\in\mathcal{P}(j)} e_k + \left\|\mathsf{S}_{\delta,U_{\epsilon,\delta,\eta},\zeta}(u_j) - \psi(u_j)\right\|_\infty \\ &\le \max_{k\in\mathcal{P}(j)} e_k + \zeta. \end{aligned} \tag{92}$$

Here, the equality uses equation 71 and $\operatorname{sp}\{\psi(u_j)\} = u_j$, since ψ is the inverse of sp. The next inequality follows from $\|\operatorname{sp}'\|_\infty \le 1$, so sp is 1-Lipschitz. Finally, equation 86 ensures that $u_j \in [\delta, U_{\epsilon,\delta,\eta}]$, and hence equation 27 gives

$$\left\|\mathsf{S}_{\delta,U_{\epsilon,\delta,\eta},\zeta}(u_j) - \psi(u_j)\right\|_\infty \le \zeta.$$

Induction gives

$$e_j \le r_j\zeta. \tag{93}$$

Since

$$H_j(\boldsymbol{g}) = \tau\{\widetilde{H}_j(\boldsymbol{g})\}, \qquad \bar{F}_j = \tau(\bar{q}_j),$$

from (93), we obtain

$$\|H_j(\boldsymbol{g}) - \bar{F}_j\|_\infty \le \frac{1}{4}e_j \le \frac{r_j\zeta}{4}. \tag{94}$$

Finally,

$$\|F^*_{A_j,\epsilon} - F^*_{A_j}\|_\infty = \epsilon\|1 - 2F^*_{A_j}\|_\infty \le \epsilon. \tag{95}$$

Combining equation 91, equation 94, and equation 95 gives

$$\begin{aligned} \|H_j(\boldsymbol{g}) - F^*_{A_j}\|_\infty &\le \|H_j(\boldsymbol{g}) - \bar{F}_j\|_\infty + \|\bar{F}_j - F^*_{A_j,\epsilon}\|_\infty \\ &\quad + \|F^*_{A_j,\epsilon} - F^*_{A_j}\|_\infty \\ &\le \epsilon + a + \frac{1}{4}\left(\eta + r_j\delta + r_j\zeta\right), \end{aligned}$$

which proves equation 73. Since $r_j \le \kappa_m$, equation 74 follows.

For $j \in \mathcal{R}$, equation 71 and equation 82 give

$$\|g_j\|_\infty \le \lambda_\epsilon + \eta.$$

For $j \notin \mathcal{R}$, Lemma 4 and equation 86 give

$$\begin{aligned} \|g_j\|_\infty &= \left\|S_{\delta,U_{\epsilon,\delta,\eta},\zeta}(u_j)\right\|_\infty \\ &\le \|\psi(u_j)\|_\infty + \left\|S_{\delta,U_{\epsilon,\delta,\eta},\zeta}(u_j) - \psi(u_j)\right\|_\infty \\ &\le \sup_{u\in[\delta,U_{\epsilon,\delta,\eta}]} |\psi(u)| + \zeta \\ &= \max\{-\psi(\delta), \psi(U_{\epsilon,\delta,\eta})\} + \zeta \\ &\le \log(1/\delta) + U_{\epsilon,\delta,\eta} + \zeta, \end{aligned}$$

where the second inequality uses $u_j \in [\delta, U_{\epsilon,\delta,\eta}]$ and the uniform approximation bound

$$\left\|S_{\delta,U_{\epsilon,\delta,\eta},\zeta} - \psi\right\|_{\infty,[\delta,U_{\epsilon,\delta,\eta}]} \le \zeta.$$

Since $\psi(u) = \log(e^u - 1)$ is increasing, the maximum of $|\psi(u)|$ over the interval is attained at one of its endpoints. Moreover,

$$-\psi(\delta) = \log \frac{1}{e^\delta - 1} \le \log(1/\delta), \qquad \psi(U_{\epsilon,\delta,\eta}) = \log\big(e^{U_{\epsilon,\delta,\eta}} - 1\big) \le U_{\epsilon,\delta,\eta},$$

where we used $e^\delta - 1 \ge \delta$ and $e^{U_{\epsilon,\delta,\eta}} - 1 \le e^{U_{\epsilon,\delta,\eta}}$. Setting

$$U_{\epsilon,\delta,\eta} = 2\lambda_\epsilon + 2\eta + \delta\kappa_m$$

proves equation 75.

It remains to establish the architecture bounds. Indeed,

$$q_j = \mathsf{L}_{\epsilon,\eta} \circ \widetilde{h}_j, \qquad \widetilde{h}_j = \epsilon + (1 - 2\epsilon)h_j.$$

The affine transformation defining $\widetilde{h}_j$ does not change the architecture order of h_j. Since h_j has depth at most L and width at most w, while $\mathsf{L}_{\epsilon,\eta}$ has depth $D_\ell(\epsilon,\eta)$ and width at most w_{cal}, Lemma 5 yields

$$q_j \in \mathcal{F}\left(L + D_\ell(\epsilon,\eta) + C,\, C\max\{w, w_{\mathrm{cal}}, 1\}\right).$$

Fix $j \in [m]$. The networks

$$q_s, \qquad s \in \overline{\mathcal{A}}(j),$$

can be evaluated in parallel with width at most

$$C a_j \max\{w, w_{\mathrm{cal}}, 1\}. \tag{96}$$

At each DAG level, every maximum in equation 69 contains at most $\deg^-_{\max,m} + 1$ inputs. Lemma 5 therefore adds at most

$$C\left\lceil \log_2(\deg^-_{\max,m} + 2)\right\rceil$$

hidden layers per graph level and increases the common width to at most

$$C a_j (\deg^-_{\max,m} + 2)\max\{w, w_{\mathrm{cal}}, 1\}.$$

Iterating this construction over the r_j graph levels yields

$$L_{\bar{q}_j} \le L + D_\ell(\epsilon,\eta) + C r_j \left\lceil \log_2(\deg^-_{\max,m} + 2)\right\rceil + C, \tag{97}$$

$$w_{\bar{q}_j} \le C a_j (\deg^-_{\max,m} + 2)\max\{w, w_{\mathrm{cal}}, 1\}. \tag{98}$$

For $j \notin \mathcal{R}$, Lemma 5 shows that computing $\max_{k\in\mathcal{P}(j)} \bar{q}_k$ adds at most

$$C\left\lceil \log_2(\deg^-_{\max,m} + 2)\right\rceil$$

hidden layers. Since $r_j \ge 1$, this additional depth is absorbed into the term in equation 97. The subtraction defining u_j is then affine. For $j \notin \mathcal{R}$, composition with $\mathsf{S}_{\delta,U_{\epsilon,\delta,\eta},\zeta}$ gives

$$\begin{aligned} L_{g_j} &\le L + D_\ell(\epsilon,\eta) + D_\psi(\delta, U_{\epsilon,\delta,\eta}, \zeta) \\ &\quad + C r_j \left\lceil \log_2(\deg^-_{\max,m} + 2)\right\rceil + C, \\ w_{g_j} &\le C a_j (\deg^-_{\max,m} + 2)\max\{w, w_{\mathrm{cal}}, 1\}. \end{aligned}$$

The same bounds hold for roots after padding. Taking the maximum over j, using

$$r_j \le \kappa_m, \qquad a_j \le a_{\max,m},$$

gives equation 77–equation 78. □

Definition 4 (Constructed functions and clipped dense ReLU class)**.** *Let $\mathcal{B}$ be a class of $[0,1]$-valued ReLU networks. For fixed*

$$0 < \epsilon < \frac{1}{4}, \qquad 0 < \eta < 1, \qquad 0 < \delta \le 1, \qquad 0 < \zeta < 1,$$

let $\mathcal{T}_j(\mathcal{B}; \epsilon, \eta, \delta, \zeta)$ be the set of all functions g_j obtained from Definition 3 with

$$h_s \in \mathcal{B}, \qquad s \in \overline{\mathcal{A}}(j),$$

and define

$$\mathcal{T}_m(\mathcal{B};\epsilon,\eta,\delta,\zeta)=\bigcup_{j=1}^{m}\mathcal{T}_j(\mathcal{B};\epsilon,\eta,\delta,\zeta). \tag{99}$$

For $R\geq 1$, define

$$\pi_R(t)=-R+\rho(t+R)-\rho(t-R)=\max\{-R,\min(t,R)\}, \tag{100}$$

and, for integers $L,w\geq 1$, let

$$\mathcal{F}^{\mathrm{cl}}(L,w,R)=\{\pi_R\circ f : f\in\mathcal{F}(L,w)\}. \tag{101}$$

All weights of f are trainable, whereas the final clipping map π_R is fixed.

Lemma 6 (Embedding and entropy of the ambient class)**.** *Suppose that every $h\in\mathcal{B}$ has depth at most $L_{\mathcal{B}}$ and width at most $w_{\mathcal{B}}$. With the notation of Theorem 2, equation 76, equation 70, and Lemma 4, define*

$$L_{\mathrm{amb}}=\left\lceil L_{\mathcal{B}}+D_\ell(\epsilon,\eta)+D_\psi(\delta,U_{\epsilon,\delta,\eta},\zeta)+C\kappa_m\left\lceil\log_2(\deg^-_{\max,m}+2)\right\rceil+C\right\rceil, \tag{102}$$

$$w_{\mathrm{amb}}=\left\lceil Ca_{\max,m}(\deg^-_{\max,m}+2)\max\{w_{\mathcal{B}},w_{\mathrm{cal}},1\}\right\rceil, \tag{103}$$

and

$$R_{\mathrm{amb}}=\max\left[1,\,C\left\{\lambda_\epsilon+\eta+\delta\kappa_m+\log(1/\delta)+\zeta+1\right\}\right]. \tag{104}$$

Then

$$\mathcal{T}_m(\mathcal{B};\epsilon,\eta,\delta,\zeta)\subseteq\mathcal{F}^{\mathrm{cl}}(L_{\mathrm{amb}},w_{\mathrm{amb}},R_{\mathrm{amb}}). \tag{105}$$

Moreover, uniformly over probability measures Q on $[0,1]^d$ and $0<u\leq R_{\mathrm{amb}}$,

$$\log N\left(u,\,\mathcal{F}^{\mathrm{cl}}(L_{\mathrm{amb}},w_{\mathrm{amb}},R_{\mathrm{amb}}),\|\cdot\|_{L^2(Q)}\right)\leq v_{\mathrm{amb}}\log\left(\frac{CR_{\mathrm{amb}}}{u}\right), \tag{106}$$

where

$$\overline{L}=L_{\mathrm{amb}}+1,\qquad \overline{w}=\max\{w_{\mathrm{amb}},2\},$$

and

$$W_{\mathrm{amb}}=(d+1)\overline{w}+(\overline{L}-1)(\overline{w}+1)\overline{w}+(\overline{w}+1), \tag{107}$$

$$U_{\mathrm{amb}}=\overline{L}\,\overline{w}+1,\qquad v_{\mathrm{amb}}=C\overline{L}W_{\mathrm{amb}}\log(eU_{\mathrm{amb}}). \tag{108}$$

Proof. By Theorem 3, every

$$g\in\mathcal{T}_m(\mathcal{B};\epsilon,\eta,\delta,\zeta)$$

admits a representation in

$$\mathcal{F}(L_{\mathrm{amb}},w_{\mathrm{amb}})$$

and satisfies

$$\|g\|_\infty\leq R_{\mathrm{amb}}.$$

Hence, by equation 100,

$$\pi_{R_{\mathrm{amb}}}\{g(x)\}=g(x),\qquad x\in[0,1]^d,$$

and therefore

$$g=\pi_{R_{\mathrm{amb}}}\circ g\in\mathcal{F}^{\mathrm{cl}}(L_{\mathrm{amb}},w_{\mathrm{amb}},R_{\mathrm{amb}}).$$

This proves equation 105.

The representation

$$\pi_R(t)=-R+\rho(t+R)-\rho(t-R)$$

adds one hidden layer with two ReLU units. After padding the hidden layers to common width

$$\overline{w}=\max\{w_{\mathrm{amb}},2\},$$

the clipped class is therefore contained in a dense ReLU architecture with

$$\overline{L}=L_{\mathrm{amb}}+1$$

hidden layers and width $\overline{w}$. Its number of parameters is at most

$$(d+1)\overline{w} + \sum_{\ell=2}^{\overline{L}} (\overline{w}+1)\overline{w} + (\overline{w}+1)$$
$$= (d+1)\overline{w} + (\overline{L}-1)(\overline{w}+1)\overline{w} + (\overline{w}+1) = W_{\mathrm{amb}},$$

and its number of computational units is at most

$$\overline{L}\,\overline{w} + 1 = U_{\mathrm{amb}}. \tag{109}$$

Since the clipping weights are fixed, $\mathcal{F}^{\mathrm{cl}}$ is a subclass of this dense ReLU architecture. Hence Bartlett et al. (2019, Theorem 7) gives

$$\begin{aligned}\mathrm{Pdim}\left\{\mathcal{F}^{\mathrm{cl}}(L_{\mathrm{amb}}, w_{\mathrm{amb}}, R_{\mathrm{amb}})\right\} &\le C\overline{L}W_{\mathrm{amb}}\log(eU_{\mathrm{amb}}) \\ &= v_{\mathrm{amb}}.\end{aligned} \tag{110}$$

Set

$$\mathcal{G} = \left\{ \frac{f}{R_{\mathrm{amb}}} : f \in \mathcal{F}^{\mathrm{cl}}(L_{\mathrm{amb}}, w_{\mathrm{amb}}, R_{\mathrm{amb}}) \right\}.$$

By construction,

$$\sup_{g\in\mathcal{G}} \|g\|_\infty \le 1, \qquad \mathrm{Pdim}(\mathcal{G}) = \mathrm{Pdim}\left\{\mathcal{F}^{\mathrm{cl}}(L_{\mathrm{amb}}, w_{\mathrm{amb}}, R_{\mathrm{amb}})\right\} \le v_{\mathrm{amb}}. \tag{111}$$

The bounded VC-subgraph covering inequality van der Vaart & Wellner (1996, Theorem 2.6.7) therefore gives, for $0 < \varepsilon \le 1$,

$$\log N\left(\varepsilon, \mathcal{G}, \|\cdot\|_{L^2(Q)}\right) \le v_{\mathrm{amb}} \log\left(\frac{C}{\varepsilon}\right). \tag{112}$$

Taking

$$\varepsilon = \frac{u}{R_{\mathrm{amb}}}$$

and using

$$N\left(u,\, \mathcal{F}^{\mathrm{cl}}(L_{\mathrm{amb}}, w_{\mathrm{amb}}, R_{\mathrm{amb}}), \|\cdot\|_{L^2(Q)}\right) = N\left(\frac{u}{R_{\mathrm{amb}}}, \mathcal{G}, \|\cdot\|_{L^2(Q)}\right)$$

proves equation 106. □

Corollary 6 (Detailed statement of Corollary 2)**.** *Suppose that*

$$F^*_{A_j} \in \mathcal{H}(l_j, \mathcal{P}_j), \qquad j = 1, \dots, m,$$

where $l_j \in \mathbb{N}$ and $\mathcal{P}_j \subset [1,\infty)\times\mathbb{N}$. Each constituent function μ may have smoothness $p_\mu = q_\mu + u_\mu$, where $q_\mu \in \mathbb{N}_0$ and $u_\mu \in (0,1]$, and input dimension d_μ, with $(p_\mu, d_\mu) \in \mathcal{P}_j$. Assume uniformly over $j \le m$ that the hierarchy depths, constituent counts and ranges, smoothness constants, the relevant derivative bounds, and the Lipschitz constants are bounded, and that

$$\max_{j\le m} \max_{(p,d_0)\in\mathcal{P}_j} \max\{p, d_0\} = O(1). \tag{113}$$

Define

$$\phi_n = \max_{j\le m} \max_{(p,d_0)\in\mathcal{P}_j} n^{-2p/(2p+d_0)}. \tag{114}$$

For sufficiently large constants c_3, c_4, set

$$L_{0,n} = \left\lceil c_3 \max_{j\le m} \max_{(p,d_0)\in\mathcal{P}_j} n^{d_0/[2(2p+d_0)]} \log(en) \right\rceil, \qquad w_{0,n} = \lceil c_4 \rceil. \tag{115}$$

For a sufficiently large positive constant C_0, set

$$\alpha_n = C_0\sqrt{\phi_n}. \tag{116}$$

For all sufficiently large n such that $\alpha_n < 1/4$, let

$$\epsilon_n = \eta_n = \alpha_n, \qquad \delta_n = \zeta_n = \frac{\alpha_n}{\kappa_m}, \tag{117}$$

and define

$$U_n = 2\log\frac{1-\alpha_n}{\alpha_n} + 3\alpha_n. \tag{118}$$

With the notation of Theorem 2, equation 76, equation 115, equation 116, and equation 118, and with D_ℓ and D_ψ defined in equation 26 and equation 28, respectively, set

$$L_n = \Bigg\lceil L_{0,n} + 1 + D_\ell(\alpha_n, \alpha_n) + D_\psi\left(\frac{\alpha_n}{\kappa_m}, U_n, \frac{\alpha_n}{\kappa_m}\right) + C\kappa_m\left\lceil \log_2(\deg^-_{\max,m} + 2)\right\rceil + C\Bigg\rceil, \tag{119}$$

$$w_n = \left\lceil C a_{\max,m}(\deg^-_{\max,m} + 2)\max\{w_{0,n}, w_{\mathrm{cal}}, 2\}\right\rceil, \tag{120}$$

$$R_n = \max\left\{1,\, C\left[\log\frac{1-\alpha_n}{\alpha_n} + \log\frac{\kappa_m}{\alpha_n} + 1\right]\right\}, \tag{121}$$

for $C > 0$ large enough. In the constrained least-squares estimator in equations (3)–(4), take the network class to be

$$\mathcal{F}^{\mathrm{cl}}(L_n, w_n, R_n).$$

Let v_n denote the quantity in equation 108 evaluated at

$$(L_{\mathrm{amb}}, w_{\mathrm{amb}}) = (L_n, w_n).$$

Then, conditionally on the design, for every $\gamma \in (0, 1/2)$, with probability at least $1-\gamma$,

$$\frac{1}{m}\sum_{j=1}^{m} \|\widehat{F}_{A_j} - F^*_{A_j}\|_n^2 \le C\left[\phi_n + \frac{\bar{a}_m v_n}{n}\log\left\{\frac{C R_n n m \sqrt{\kappa_m a_{\max,m}}}{\gamma}\right\}\right]. \tag{122}$$

If m grows no faster than polynomially in n, then

$$\frac{1}{m}\sum_{j=1}^{m} \|\widehat{F}_{A_j} - F^*_{A_j}\|_n^2 = O_{\mathbb{P}}\left(\phi_n + \frac{\bar{a}_m v_n \log(en)}{n}\right). \tag{123}$$

For a fixed DAG,

$$L_n \lesssim \sqrt{n\phi_n}\log(en), \qquad w_n = O(1), \qquad R_n \lesssim \log(en), \qquad v_n \lesssim n\phi_n \log^3(en), \tag{124}$$

and hence

$$\frac{1}{m}\sum_{j=1}^{m} \|\widehat{F}_{A_j} - F^*_{A_j}\|_n^2 = O_{\mathbb{P}}\{\phi_n \log^4(en)\}. \tag{125}$$

The constant in equation 122 depends only on the fixed input dimension and the uniform hierarchical-composition bounds, and not on n or the DAG.

Proof. For brevity, write

$$D_{\ell,n} = D_\ell(\alpha_n, \alpha_n), \qquad D_{\psi,n} = D_\psi\left(\frac{\alpha_n}{\kappa_m}, U_n, \frac{\alpha_n}{\kappa_m}\right),$$

and, within this proof, let

$$\mathcal{F} = \mathcal{F}^{\mathrm{cl}}(L_n, w_n, R_n), \qquad A_n = CeR_n. \tag{126}$$

Approximation. By Kohler & Langer (2021, Theorem 3(b)), for each $j \in [m]$ there exists

$$f_j \in \mathcal{F}(L_{0,n}, w_{0,n})$$

here $L_{0,n}$ and $w_{0,n}$ are defined in equation 115, and such that

$$\|f_j - F^*_{A_j}\|^2_\infty \le C \max_{(p,d_0)\in\mathcal{P}_j} n^{-2p/(2p+d_0)} \le C\phi_n. \tag{127}$$

Define

$$c(t) = \rho(t) - \rho(t-1) = \max\{0, \min(t,1)\},$$

and, only for this proof, set

$$\mathcal{B}_n = \{c \circ f : f \in \mathcal{F}(L_{0,n}, w_{0,n})\}. \tag{128}$$

Since

$$|c(t) - y| \le |t - y|, \qquad t \in \mathbb{R}, \quad y \in [0,1],$$

the networks

$$h_j = c \circ f_j \in \mathcal{B}_n$$

satisfy

$$\max_{j\le m} \|h_j - F^*_{A_j}\|_\infty \le \max_{j\le m} \|f_j - F^*_{A_j}\|_\infty \le C\sqrt{\phi_n} \le \alpha_n, \tag{129}$$

where C_0 is chosen sufficiently large.

Applying Theorem 3 with $a = \alpha_n$ and

$$\epsilon_n = \eta_n = \alpha_n, \qquad \delta_n = \zeta_n = \frac{\alpha_n}{\kappa_m},$$

gives

$$\begin{aligned}\max_{j\le m} \|H_j(\boldsymbol{g}) - F^*_{A_j}\|_\infty &\le \alpha_n + \alpha_n + \frac{1}{4}\left(\alpha_n + \kappa_m \frac{\alpha_n}{\kappa_m} + \kappa_m \frac{\alpha_n}{\kappa_m}\right) \\ &= \frac{11}{4}\alpha_n. \end{aligned} \tag{130}$$

Every $h \in \mathcal{B}_n$ has depth at most $L_{0,n} + 1$ and width at most

$$\max\{w_{0,n}, 2\}.$$

Hence Lemma 6, applied with

$$L_{\mathcal{B}} = L_{0,n} + 1, \qquad w_{\mathcal{B}} = \max\{w_{0,n}, 2\},$$

gives

$$L_{\mathrm{amb}} \le L_n, \qquad w_{\mathrm{amb}n}, \tag{131}$$

because L_n and w_n in equation 119–equation 120 are defined as the corresponding ambient architecture bounds after these substitutions.

Moreover, by the tuning parameters,

$$\begin{aligned} L_{\epsilon_n} + \eta_n + \kappa_m\delta_n + \log(1/\delta_n) + \zeta_n + 1 &= \log\frac{1-\alpha_n}{\alpha_n} + 2\alpha_n + \log\frac{\kappa_m}{\alpha_n} + \frac{\alpha_n}{\kappa_m} + 1 \\ &\le C\left\{\log\frac{1-\alpha_n}{\alpha_n} + \log\frac{\kappa_m}{\alpha_n} + 1\right\}, \end{aligned} \tag{132}$$

for a postive constant C. The first equality follows by substituting the tuning choices $\epsilon_n = \eta_n = \alpha_n$ and $\delta_n = \zeta_n = \alpha_n/\kappa_m$, together with the definition $L_\epsilon = \log\{(1-\epsilon)/\epsilon\}$ in equation 66. The second inequality follows from $\alpha_n < 1/4$ and $\kappa_m \ge 1$.

By equation 128, the functions $h_s = c \circ f_s$ satisfy $h_s \in \mathcal{B}_n$. Hence, by Definition 4, the corresponding latent functions satisfy

$$g_j \in \mathcal{T}_m(\mathcal{B}_n; \epsilon_n, \eta_n, \delta_n, \zeta_n), \qquad j \in [m].$$

Moreover, from equation 104, equation 121, (132) we obtain that $R_{\mathrm{amb}} \le R_n$. Hence, from equation 105 and definition of $\mathcal{F}$ in equation 126, yields

$$g_j \in \mathcal{T}_m(\mathcal{B}_n; \epsilon_n, \eta_n, \delta_n, \zeta_n) \subseteq \mathcal{F}, \qquad j \in [m]. \tag{133}$$

Hence $\boldsymbol{g} \in \mathcal{F}^m$. By the definition of $\Delta_{n,m}$ following Theorem 2, its infimum over $\mathcal{F}^m$ is no larger than the approximation error of this particular $\boldsymbol{g}$. Therefore, using equation 130 and $\alpha_n = C_0\sqrt{\phi_n}$ from equation 116,

$$\Delta_{n,m}^2 \le \max_{j\le m} \|H_j(\boldsymbol{g}) - F_{A_j}^*\|_\infty^2 \le \frac{121}{16}\alpha_n^2 \lesssim \phi_n. \tag{134}$$

Estimation.

By equation 126, $\mathcal{F} = \mathcal{F}^{\mathrm{cl}}(L_n, w_n, R_n)$. Hence, the argument in Lemma 6 yields, uniformly over all probability measures Q on $[0,1]^d$,

$$\log N\left(u, \mathcal{F}, \|\cdot\|_{L^2(Q)}\right) \le v_n \log\left(\frac{A_n}{u}\right), \qquad 0 < u \le R_n, \tag{135}$$

where v_n is defined by equation 108 with $(L_{\mathrm{amb}}, w_{\mathrm{amb}})$ replaced by (L_n, w_n) and $A_n = CeR_n \ge eR_n$.

Thus the approximation condition and the entropy condition in the explicit bound following Theorem 2 are satisfied by equation 134 and equation 135, respectively. Applying that bound with $B_n = R_n$ therefore gives equation 122, after enlarging the universal constant C if necessary.

Suppose now that

$$m = O(n^{C_m})$$

for some fixed $C_m > 0$. Then

$$1 \le \bar{a}_m \le a_{\max,m} \le m, \qquad 1 \le \kappa_m \le m, \qquad 0 \le \deg^-_{\max,m} \le m-1. \tag{136}$$

By equation 114, equation 116, and equation 113,

$$\log(1/\alpha_n) = O(\log(en)).$$

Hence, using equation 121, $m = O(n^{C_m})$, (121), and equation 136,

$$R_n \lesssim \log(en), \qquad \log\left\{\frac{eA_n n m\sqrt{\kappa_m a_{\max,m}}}{\gamma}\right\} = O_\gamma(\log(en)), \tag{137}$$

where we also used $A_n = CeR_n$ from equation 126. Substituting equation 137 into equation 122 proves equation 123.

Fixed DAG. By the definition of ϕ_n,

$$\begin{aligned}\max_{j\le m}\max_{(p,d_0)\in\mathcal{P}_j} n^{d_0/[2(2p+d_0)]} &= \max_{j\le m}\max_{(p,d_0)\in\mathcal{P}_j}\left\{n\, n^{-2p/(2p+d_0)}\right\}^{1/2} \\ &= \sqrt{n\phi_n}.\end{aligned} \tag{138}$$

Thus

$$L_{0,n} \asymp \sqrt{n\phi_n}\log(en). \tag{139}$$

Since p is uniformly bounded and $d_0 \ge 1$, there exists $c_* > 0$ such that, for all sufficiently large n,

$$n\phi_n = \max_{j\le m}\max_{(p,d_0)\in\mathcal{P}_j} n^{d_0/(2p+d_0)} \ge n^{c_*}. \tag{140}$$

For a fixed DAG,

$$\kappa_m = O(1), \qquad \deg^-_{\max,m} = O(1), \qquad a_{\max,m} = O(1),$$

and

$$U_n + \Lambda_{\epsilon_n} + \Lambda_{\eta_n} + \Lambda_{\delta_n} + \Lambda_{\zeta_n} \lesssim \log(en). \tag{141}$$

Here U_n is defined in equation 118, $(\epsilon_n, \eta_n, \delta_n, \zeta_n)$ are defined in equation 117, and $\Lambda_s = \log(e/s)$ is defined at the beginning of Section D. Lemma 4 therefore gives

$$D_{\ell,n} + D_{\psi,n} \lesssim \log^3(en). \tag{142}$$

By equation 139 and equation 140,

$$\log^3(en) = o(L_{0,n}),$$

and hence

$$D_{\ell,n} + D_{\psi,n} = o(L_{0,n}). \tag{143}$$

It follows from the definitions of L_n, w_n, and R_n in equation 119–equation 121 that

$$L_n \lesssim \sqrt{n\phi_n}\log(en), \qquad w_n = O(1), \qquad R_n \lesssim \log(en). \tag{144}$$

Since $w_n = O(1)$, the parameter and unit counts in equation 107–equation 108, replacing $(L_{\mathrm{amb}}, w_{\mathrm{amb}})$ with (L_n, w_n), satisfy

$$W_{\mathrm{amb}} \lesssim L_n, \qquad U_{\mathrm{amb}} \lesssim L_n.$$

Recall from the statement of Corollary 6 that v_n is defined by equation 108 replacing $(L_{\mathrm{amb}}, w_{\mathrm{amb}})$ with (L_n, w_n). Hence, since $\overline{L} = L_n + 1$, $W_{\mathrm{amb}} \lesssim L_n$, and $U_{\mathrm{amb}} \lesssim L_n$,

$$\begin{aligned} v_n &\lesssim L_n^2 \log(eL_n) \\ &\lesssim n\phi_n \log^3(en). \end{aligned} \tag{145}$$

where the second inequality follows from equation 144. This proves equation 124. Finally, since $\bar{a}_m = O(1)$ for a fixed DAG,

$$\begin{aligned} \phi_n + \frac{\bar{a}_m v_n \log(en)}{n} &\lesssim \phi_n + \frac{n\phi_n \log^3(en)\log(en)}{n} \\ &\lesssim \phi_n \log^4(en). \end{aligned}$$

Equation equation 125 follows from equation 123. □

E TREND FILTERING AND ORTHOGONAL SERIES REGRESSION

Theorems 1 and 2 apply once the approximation error and the covering entropy of the function class are controlled. In this section, we derive results for trend filtering and orthogonal series regression, including concrete rates for both the preliminary and constrained estimators.

E.1 TREND FILTERING

Trend filtering estimates a univariate regression function by controlling the total variation of one of its derivatives, allowing the fitted function to adapt to spatially varying smoothness. We consider the corresponding bounded function class below.

Let $d = 1$ and fix an integer $r \geq 1$. We first define the ambient smoothness class

$$\mathcal{S}_r = \left\{ f : [0,1] \to \mathbb{R} : \mathrm{TV}\big(f^{(r-1)}\big) < \infty \right\}, \tag{146}$$

consisting of functions whose $(r-1)$st derivative has finite total variation. For $r = 1$, $\mathcal{S}_r$ is understood in the usual pointwise bounded-variation sense; for $r \geq 2$, we take $f \in W^{r-1,1}(0,1)$ with weak derivative $f^{(r-1)} \in BV(0,1)$ and use the continuous representative of f on $[0,1]$.

For $V \geq 0$ and $B \geq 1$, define the bounded trend-filtering class

$$\mathcal{G}_r(V,B) = \left\{ f \in \mathcal{S}_r : \mathrm{TV}\big(f^{(r-1)}\big) \leq V,\ \|f\|_\infty \leq B \right\}. \tag{147}$$

PRELIMINARY ESTIMATOR

For $A \in \Lambda$, define

$$\widehat{F}_A^{\mathrm{pre,TF}} \in \arg\min_{f \in \mathcal{G}_r(V_n,B)} \sum_{i=1}^n \{z_i^A - f(x_i)\}^2, \tag{148}$$

and let

$$\phi_{n,\mathrm{TF}}^{\mathrm{pre}} = \sup_{A \in \Lambda} \inf_{g \in \mathcal{G}_r(V_n,B)} \|F_A^* - g\|_\infty^2. \tag{149}$$

Corollary 7 (Preliminary trend-filtering estimator)**.** *Suppose Assumption 1 holds, $B = O(1)$, and*

$$(1 \vee V_n)^{\frac{2}{2r+1}} n^{-\frac{2r}{2r+1}} + \frac{\log n}{n} + \phi_{n,\mathrm{TF}}^{\mathrm{pre}} = o(1). \tag{150}$$

Then

$$\sup_{A \in \Lambda} \|\widehat{F}_A^{\mathrm{pre,TF}} - F_A^*\|_n^2 = O_{\mathbb{P}}\left((1 \vee V_n)^{\frac{2}{2r+1}} n^{-\frac{2r}{2r+1}} + \frac{\log n}{n} + \phi_{n,\mathrm{TF}}^{\mathrm{pre}} \right). \tag{151}$$

If, for some constant $V < \infty$,

$$\sup_{A \in \Lambda} \mathrm{TV}\big((F_A^*)^{(r-1)}\big) \le V, \tag{152}$$

then, for $V \le V_n \le \bar{V} < \infty$, with $\bar{V}$ fixed, and $B \ge 1$,

$$\sup_{A \in \Lambda} \|\widehat{F}_A^{\mathrm{pre,TF}} - F_A^*\|_n^2 = O_{\mathbb{P}}\left(n^{-\frac{2r}{2r+1}} \right). \tag{153}$$

Proof. We use the standard entropy bound

$$\sup_{q \ge 1} \sup_{x_1,\dots,x_q \in [0,1]} \log N(u, \mathcal{G}_r(1,1), \|\cdot\|_q) \le C_r u^{-1/r}, \qquad 0 < u \le 1, \tag{154}$$

where C_r depends only on r; see the proof of Sadhanala & Tibshirani (2019a;b, Corollary 1).

Let

$$\mathcal{D}_{r,n} = \mathcal{G}_r(V_n, B) - \mathcal{G}_r(V_n, B).$$

For $h \in \mathcal{D}_{r,n}$,

$$\mathrm{TV}\big(h^{(r-1)}\big) \le 2V_n, \qquad \|h\|_\infty \le 2B.$$

Hence, after rescaling equation 154,

$$\log N(u, \mathcal{D}_{r,n}, \|\cdot\|_n) \le C_r (1 + B + V_n)^{1/r} u^{-1/r}. \tag{155}$$

Therefore, for $t > 0$,

$$\int_0^t \sqrt{\log N(u, \mathcal{D}_{r,n}, \|\cdot\|_n)}\, du \le C_r (1 + B + V_n)^{1/(2r)} t^{1-1/(2r)}. \tag{156}$$

Substituting equation 156 into Theorem 1, with $\rho_j = 2^{j/2}\eta/\sqrt{n}$, gives

$$\frac{\sqrt{n}}{\eta^2} \sum_{j=1}^{J} \frac{\mathcal{J}(\rho_j)}{2^j} \lesssim (1 + B + V_n)^{1/(2r)} n^{1/(4r)} \eta^{-(2r+1)/(2r)},$$

where we used $\sum_{j \ge 1} 2^{-j(2r+1)/(4r)} < \infty$. Consequently,

$$\sup_{A \in \Lambda} \|\widehat{F}_A^{\mathrm{pre,TF}} - F_A^*\|_n = O_{\mathbb{P}}\left((1 + B + V_n)^{\frac{1}{2r+1}} n^{-\frac{r}{2r+1}} + \sqrt{\frac{\log n}{n}} + \sqrt{\phi_{n,\mathrm{TF}}^{\mathrm{pre}}} \right). \tag{157}$$

Since $B = O(1)$, squaring yields equation 151. Under equation 152, $F_A^* \in \mathcal{G}_r(V_n, B)$ because $0 \le F_A^* \le 1$. Hence

$$\phi_{n,\mathrm{TF}}^{\mathrm{pre}} = 0.$$

Since

$$\frac{\log n}{n} = o\left(n^{-2r/(2r+1)} \right),$$

equation 153 follows. □

CONSTRAINED ESTIMATOR

Take $\mathcal{F} = \mathcal{G}_r(V_n, B)$ and define

$$\widehat{\boldsymbol{g}}^{\mathrm{TF}} \in \arg\min_{\boldsymbol{g}\in\mathcal{G}_r(V_n,B)^m} \sum_{j=1}^{m}\sum_{i=1}^{n}\{z_i^j - H_j(\boldsymbol{g})(x_i)\}^2, \tag{158}$$

with

$$\widehat{F}_{A_j}^{\mathrm{con,TF}} = H_j(\widehat{\boldsymbol{g}}^{\mathrm{TF}}), \qquad j = 1, \ldots, m. \tag{159}$$

Define

$$\Delta_{n,m}^{\mathrm{TF}} = \inf_{\boldsymbol{g}\in\mathcal{G}_r(V_n,B)^m} \max_{j\le m} \|F_{A_j}^* - H_j(\boldsymbol{g})\|_\infty. \tag{160}$$

Corollary 8 (Constrained trend-filtering estimator)**.** *Suppose*

$$(\Delta_{n,m}^{\mathrm{TF}})^2 + (1 \vee V_n)^{\frac{2}{2r+1}} n^{-\frac{2r}{2r+1}} \frac{1}{m} \left\{ \sum_{j=1}^{m} a_j^{\frac{2r+1}{4r}} \kappa_j^{\frac{1}{4r}} \right\}^{\frac{4r}{2r+1}} = o(1). \tag{161}$$

Then

$$\frac{1}{m}\sum_{j=1}^{m} \|\widehat{F}_{A_j}^{\mathrm{con,TF}} - F_{A_j}^*\|_n^2 = O_{\mathbb{P}}\Bigg\{ (\Delta_{n,m}^{\mathrm{TF}})^2 + (1 \vee V_n)^{\frac{2}{2r+1}} n^{-\frac{2r}{2r+1}} \frac{1}{m} \left(\sum_{j=1}^{m} a_j^{\frac{2r+1}{4r}} \kappa_j^{\frac{1}{4r}} \right)^{\frac{4r}{2r+1}} \Bigg\}. \tag{162}$$

Proof. By equation 154 and scaling,

$$\log N\big(u, \mathcal{G}_r(V_n, B), \|\cdot\|_n\big) \le C(1 \vee V_n)^{1/r} u^{-1/r}. \tag{163}$$

Hence, by Lemma 2, for every $\widetilde{\boldsymbol{g}} \in \mathcal{G}_r(V_n, B)^m$,

$$\log N\big(u, \mathcal{I}_{j,n}(\widetilde{\boldsymbol{g}}), \|\cdot\|_n\big) \le C a_j^{1+\frac{1}{2r}} \kappa_j^{\frac{1}{2r}} (1 \vee V_n)^{1/r} u^{-1/r}. \tag{164}$$

Therefore,

$$\sqrt{n}\int_{1/n}^{C_3 2^{k/2}\eta/\sqrt{n}} \sqrt{\log N\big(u, \mathcal{I}_{j,n}(\widetilde{\boldsymbol{g}}), \|\cdot\|_n\big)}\, du \le C a_j^{\frac{2r+1}{4r}} \kappa_j^{\frac{1}{4r}} (1 \vee V_n)^{\frac{1}{2r}} n^{\frac{1}{4r}} 2^{\frac{k(2r-1)}{4r}} \eta^{1-\frac{1}{2r}}. \tag{165}$$

Substitution into Theorem 2 and

$$\sum_{k\ge 1} 2^{-k(2r+1)/(4r)} < \infty$$

give

$$\mathbb{P}\left\{ \sum_{j=1}^{m}\sum_{i=1}^{n} \big(F_{A_j}^*(x_i) - \widehat{F}_{A_j}^{\mathrm{con,TF}}(x_i)\big)^2 > \eta^2 \right\} \le C\Bigg\{ \frac{m}{\eta^2} + (1 \vee V_n)^{\frac{1}{2r}} n^{\frac{1}{4r}} \eta^{-\frac{2r+1}{2r}} \sum_{j=1}^{m} a_j^{\frac{2r+1}{4r}} \kappa_j^{\frac{1}{4r}} \Bigg\}. \tag{166}$$

Let $T_{n,m}$ denote the stochastic term in equation 162, and set

$$R_{n,m} = (\Delta_{n,m}^{\mathrm{TF}})^2 + T_{n,m}.$$

By the definition of the infimum, choose a comparator $\widetilde{\boldsymbol{g}} \in \mathcal{G}_r(V_n, B)^m$ with squared approximation error at most $2R_{n,m}$. Taking

$$\eta^2 = MnmR_{n,m}$$

for a sufficiently large fixed constant M, condition equation 161 ensures that the requirements on η in Theorem 2 hold for all sufficiently large n.

Since

$$\sum_{j=1}^{m} a_j^{\frac{2r+1}{4r}} \kappa_j^{\frac{1}{4r}} \geq m,$$

we have

$$nT_{n,m} \geq n^{1/(2r+1)}.$$

Hence the first term in equation 166 satisfies

$$\frac{m}{\eta^2} \leq M^{-1} n^{-1/(2r+1)},$$

while direct substitution into the second term gives an upper bound of order $M^{-(2r+1)/(4r)}$. Taking M sufficiently large proves equation 162. □

To simplify the derivation of the concrete constrained rates, we use the following bounded latent representation condition.

Assumption 2 (Bounded latent representation)**.** *For every $j \leq m$ and $x \in [0,1]^d$,*

$$0 < F^*_{A_j}(x) < 1, \tag{167}$$

and, for every edge $(k, j) \in E$,

$$F^*_{A_k}(x) < F^*_{A_j}(x). \tag{168}$$

For $j \in \mathcal{R}$, define

$$g^*_j = \tau^{-1} \circ F^*_{A_j}, \tag{169}$$

and, for $j \notin \mathcal{R}$,

$$g^*_j(x) = \log\left[\exp\left\{\tau^{-1}\big(F^*_{A_j}(x)\big) - \max_{k \in \mathcal{P}(j)} \tau^{-1}\big(F^*_{A_k}(x)\big)\right\} - 1\right]. \tag{170}$$

Suppose

$$\max_{j \leq m} \|g^*_j\|_\infty \leq C_0 \tag{171}$$

for a constant $C < \infty$.

For $j \notin \mathcal{R}$,

$$\log(1 + e^{g^*_j}) = \tau^{-1}(F^*_{A_j}) - \max_{k \in \mathcal{P}(j)} \tau^{-1}(F^*_{A_k}). \tag{172}$$

Together with equation 169, induction over a topological ordering of the DAG gives

$$F^*_{A_j} = H_j(\boldsymbol{g}^*), \qquad j = 1, \ldots, m. \tag{173}$$

CONSTRAINED ESTIMATOR: CONCRETE TREND-FILTERING RATE

Corollary 9 (Constrained trend filtering: bounded latent functions)**.** *Suppose Assumption 2 holds and, for a constant $V < \infty$,*

$$\max_{j \leq m} \mathrm{TV}\big((g^*_j)^{(r-1)}\big) \leq V. \tag{174}$$

Set $V_n = V$ and take a fixed $B \geq \max\{1, C\}$. If

$$n^{-\frac{2r}{2r+1}} \frac{1}{m} \left\{\sum_{j=1}^{m} a_j^{\frac{2r+1}{4r}} \kappa_j^{\frac{1}{4r}}\right\}^{\frac{4r}{2r+1}} = o(1), \tag{175}$$

then

$$\frac{1}{m} \sum_{j=1}^{m} \|\widehat{F}^{\mathrm{con,TF}}_{A_j} - F^*_{A_j}\|_n^2 = O_{\mathbb{P}}\left[n^{-\frac{2r}{2r+1}} \frac{1}{m} \left\{\sum_{j=1}^{m} a_j^{\frac{2r+1}{4r}} \kappa_j^{\frac{1}{4r}}\right\}^{\frac{4r}{2r+1}}\right]. \tag{176}$$

For a fixed DAG, equation 176 reduces to $O_{\mathbb{P}}(n^{-2r/(2r+1)})$.

Proof. By Assumption 2, equation 173, and equation 174,

$$\boldsymbol{g}^* \in \mathcal{G}_r(V, B)^m, \qquad \Delta^{\mathrm{TF}}_{n,m} = 0. \tag{177}$$

The result follows from Corollary 8. For a fixed DAG, m, a_j, and κ_j are constants. □

E.2 Orthogonal series regression

Let $\{\varphi_k\}_{k\geq 1}$ be a fixed orthonormal basis of $L^2([0,1]^d)$, with the basis functions defined pointwise on $[0,1]^d$. For an integer $J \geq 1$, define

$$\mathcal{L}_J = \operatorname{span}\{\varphi_1, \ldots, \varphi_J\}, \qquad \mathcal{O}_J(B) = \{f \in \mathcal{L}_J : \|f\|_\infty \leq B\}, \tag{178}$$

where $B \geq 1$ is fixed. The empirical quotient of $\mathcal{L}_J$ has dimension at most J. Hence, for $0 < u \leq B$, the standard volumetric bound gives

$$N(u, \mathcal{O}_J(B), \|\cdot\|_n) \leq (1 + 2B/u)^J. \tag{179}$$

Preliminary estimator

For $A \in \Lambda$, define

$$\widehat{F}^{\mathrm{pre,OS}}_{A,J} \in \arg\min_{f \in \mathcal{O}_J(B)} \sum_{i=1}^n \{z_i^A - f(x_i)\}^2 \tag{180}$$

and let

$$\phi^{\mathrm{pre,OS}}_{n,J} = \sup_{A\in\Lambda} \inf_{g \in \mathcal{O}_J(B)} \|F_A^* - g\|_\infty^2. \tag{181}$$

Corollary 10 (Preliminary orthogonal-series estimator)**.** *Suppose Assumption 1 holds, $J = J_n$, and*

$$\frac{J + \log n}{n} + \phi^{\mathrm{pre,OS}}_{n,J} = o(1). \tag{182}$$

Then

$$\sup_{A\in\Lambda} \|\widehat{F}^{\mathrm{pre,OS}}_{A,J} - F_A^*\|_n^2 = O_{\mathbb{P}}\left(\frac{J + \log n}{n} + \phi^{\mathrm{pre,OS}}_{n,J}\right). \tag{183}$$

If, for some $s > 0$,

$$F_A^* = \sum_{k=1}^\infty \theta_{A,k}\varphi_k, \qquad \sup_{A\in\Lambda} \sum_{k=1}^\infty k^{s/d}|\theta_{A,k}|\,\|\varphi_k\|_\infty \leq C_s, \tag{184}$$

then, for $B \geq 1 + C_s$ and $J_n \asymp n^{d/(2s+d)}$,

$$\sup_{A\in\Lambda} \|\widehat{F}^{\mathrm{pre,OS}}_{A,J_n} - F_A^*\|_n^2 = O_{\mathbb{P}}\left(n^{-\frac{2s}{2s+d}}\right). \tag{185}$$

Proof. For $t > 0$, let

$$\mathcal{D}_J(t) = \{f - g : f, g \in \mathcal{O}_J(B),\ \|f - g\|_n \leq t\}.$$

The volumetric bound in the empirical quotient space gives

$$N(u, \mathcal{D}_J(t), \|\cdot\|_n) \leq (1 + 2t/u)^J,$$

and hence

$$\int_0^t \sqrt{\log N(u, \mathcal{D}_J(t), \|\cdot\|_n)}\, du \leq Ct\sqrt{J}. \tag{186}$$

For the shell radii $\rho_\ell = 2^{\ell/2}\eta/\sqrt{n}$ in Theorem 1, equation 186 gives

$$\frac{\sqrt{n}}{\eta^2} \sum_\ell \frac{\mathcal{J}(\rho_\ell)}{2^\ell} \lesssim \frac{\sqrt{J}}{\eta} \sum_{\ell\geq 1} 2^{-\ell/2} \lesssim \frac{\sqrt{J}}{\eta}.$$

Taking η to be a sufficiently large constant multiple of

$$\sqrt{J + \log n} + \sqrt{n\phi^{\mathrm{pre,OS}}_{n,J}}$$

in Theorem 1 yields

$$\sup_{A\in\Lambda} \|\widehat{F}^{\mathrm{pre,OS}}_{A,J} - F_A^*\|_n = O_{\mathbb{P}}\left(\sqrt{\frac{J + \log n}{n}} + \sqrt{\phi^{\mathrm{pre,OS}}_{n,J}}\right),$$

which proves equation 183.

Under equation 184, let

$$G_{A,J} = \sum_{k=1}^{J} \theta_{A,k}\varphi_k.$$

Then

$$\|F_A^* - G_{A,J}\|_\infty \le J^{-s/d} \sum_{k>J} k^{s/d}|\theta_{A,k}|\,\|\varphi_k\|_\infty \le C_s J^{-s/d}. \tag{187}$$

Since $\|F_A^*\|_\infty \le 1$, $\|G_{A,J}\|_\infty \le 1 + C_s \le B$, and therefore

$$\phi_{n,J}^{\mathrm{pre,OS}} \le C_s^2 J^{-2s/d}.$$

Balancing J/n and $J^{-2s/d}$ proves equation 185. □

CONSTRAINED ESTIMATOR

Take $\mathcal{F} = \mathcal{O}_J(B)$ and define

$$\widehat{\boldsymbol{g}}^{\mathrm{OS}} \in \arg\min_{\boldsymbol{g}\in\mathcal{O}_J(B)^m} \sum_{j=1}^{m}\sum_{i=1}^{n}\{z_i^j - H_j(\boldsymbol{g})(x_i)\}^2, \tag{188}$$

with

$$\widehat{F}_{A_j}^{\mathrm{con,OS}} = H_j(\widehat{\boldsymbol{g}}^{\mathrm{OS}}), \qquad j = 1,\dots,m. \tag{189}$$

Define

$$\Delta_{n,m}^{\mathrm{OS}} = \inf_{\boldsymbol{g}\in\mathcal{O}_J(B)^m} \max_{j\le m} \|F_{A_j}^* - H_j(\boldsymbol{g})\|_\infty. \tag{190}$$

Corollary 11 (Constrained orthogonal-series estimator)**.** *For every $\delta \in (0, 1/2)$, with probability at least $1-\delta$,*

$$\frac{1}{m}\sum_{j=1}^{m} \|\widehat{F}_{A_j}^{\mathrm{con,OS}} - F_{A_j}^*\|_n^2 \le C\left\{(\Delta_{n,m}^{\mathrm{OS}})^2 + \frac{\bar{a}_m J}{n}\log\!\left(\frac{eBnm\sqrt{\kappa_m a_{\max,m}}}{\delta}\right)\right\}. \tag{191}$$

Proof. By equation 179,

$$\log N(u, \mathcal{O}_J(B), \|\cdot\|_n) \le J\log(3B/u), \qquad 0 < u \le B. \tag{192}$$

Thus the covering condition following Theorem 2 holds with $v_n = J$ and $A_n = 3B$. Substitution into equation 6 proves equation 191. □

CONSTRAINED ESTIMATOR: CONCRETE ORTHOGONAL-SERIES RATE

Corollary 12 (Constrained orthogonal series: bounded latent functions)**.** *Suppose Assumption 2 holds and, for some $s > 0$,*

$$g_j^* = \sum_{k=1}^{\infty}\gamma_{j,k}\varphi_k, \qquad \max_{j\le m}\sum_{k=1}^{\infty} k^{s/d}|\gamma_{j,k}|\,\|\varphi_k\|_\infty \le C_s. \tag{193}$$

If $B \ge \max\{1, C_0 + C_s\}$ is fixed, then

$$\frac{1}{m}\sum_{j=1}^{m} \|\widehat{F}_{A_j}^{\mathrm{con,OS}} - F_{A_j}^*\|_n^2 = O_{\mathbb{P}}\left\{\kappa_m^2 J^{-2s/d} + \frac{\bar{a}_m J}{n}\log\big(eBnm\sqrt{\kappa_m a_{\max,m}}\big)\right\}. \tag{194}$$

Choosing the integer truncation level

$$J_n = \max\left\{1, \left\lceil\left[\frac{\kappa_m^2 n}{\bar{a}_m \log\big(eBnm\sqrt{\kappa_m a_{\max,m}}\big)}\right]^{\frac{d}{2s+d}}\right\rceil\right\},$$

we further obtain

$$\frac{1}{m}\sum_{j=1}^{m}\|\widehat{F}_{A_j}^{\mathrm{con,OS}}-F_{A_j}^{*}\|_n^2 = O_{\mathbb{P}}\left\{\kappa_m^{\frac{2d}{2s+d}}\left[\frac{\bar{a}_m}{n}\log\big(eBnm\sqrt{\kappa_m a_{\max,m}}\big)\right]^{\frac{2s}{2s+d}} + \frac{\bar{a}_m}{n}\log\big(eBnm\sqrt{\kappa_m a_{\max,m}}\big)\right\}. \tag{195}$$

In particular, if $\kappa_m, \bar{a}_m, a_{\max,m} = O(1)$ and m grows at most polynomially in n, then, ignoring logarithmic factors,

$$\frac{1}{m}\sum_{j=1}^{m}\|\widehat{F}_{A_j}^{\mathrm{con,OS}} - F_{A_j}^{*}\|_n^2 = O_{\mathbb{P}}\Big(n^{-\frac{2s}{2s+d}}\Big).$$

Proof. Let

$$g_{j,J} = \sum_{k=1}^{J}\gamma_{j,k}\varphi_k.$$

By equation 193,

$$\max_{j\le m}\|g_j^{*} - g_{j,J}\|_\infty \le C_s J^{-s/d}. \tag{196}$$

Moreover,

$$\|g_{j,J}\|_\infty \le C_0 + C_s \le B,$$

so $\boldsymbol{g}_J \in \mathcal{O}_J(B)^m$. By equation 173 and equation 12,

$$\max_{j\le m}\|H_j(\boldsymbol{g}_J) - F_{A_j}^{*}\|_\infty \le C\kappa_m J^{-s/d}. \tag{197}$$

Hence

$$(\Delta_{n,m}^{\mathrm{OS}})^2 \le C\kappa_m^2 J^{-2s/d}. \tag{198}$$

For the stated choice of J_n, put

$$t = \left\{\frac{\kappa_m^2 n}{\bar{a}_m\log\big(eBnm\sqrt{\kappa_m a_{\max,m}}\big)}\right\}^{\frac{d}{2s+d}}.$$

Since $t \le J_n \le t+1$,

$$\begin{aligned}&\kappa_m^2 J_n^{-2s/d} + \frac{\bar{a}_m J_n}{n}\log\big(eBnm\sqrt{\kappa_m a_{\max,m}}\big)\\ &\quad\lesssim \kappa_m^{\frac{2d}{2s+d}}\left[\frac{\bar{a}_m}{n}\log\big(eBnm\sqrt{\kappa_m a_{\max,m}}\big)\right]^{\frac{2s}{2s+d}} + \frac{\bar{a}_m}{n}\log\big(eBnm\sqrt{\kappa_m a_{\max,m}}\big).\end{aligned}$$

This proves equation 195. If $\kappa_m, \bar{a}_m, a_{\max,m} = O(1)$ and m grows at most polynomially in n, ignoring logarithmic factors gives $O_{\mathbb{P}}(n^{-2s/(2s+d)})$. □

F REPRESENTATIVE DAG STRUCTURES

This section illustrates how the topology-dependent quantities κ_m and $\bar{a}_m$ vary across several representative DAG structures. Recall that κ_m is the maximum number of vertices on a root-to-node path, and $\bar{a}_m = \frac{1}{m}\sum_{j=1}^{m} a_j$, where a_j denotes the number of ancestors of node j, including the node itself. Ignoring logarithmic factors, equation 6 yields an estimation term of order $\bar{a}_m v_n/n$.

We consider the following DAGs.

- **Isolated roots.** The graph contains m nodes and no edges. Hence every node is a root and has only itself as an ancestor.
- **Chain.** The nodes are ordered as $1 \to 2 \to \cdots \to m$. Thus node j has ancestor set $\{1, \ldots, j\}$.
- **Rooted tree of depth D.** The DAG is a rooted directed tree with all edges oriented away from the root, and the longest root-to-node path contains $D+1$ vertices. A node at depth r has $r+1$ ancestors.
- **Star with $m-1$ roots and one sink.** There are $m-1$ root nodes, each with an edge to a single sink node. Thus the sink has all m nodes as ancestors, whereas every root has only itself as an ancestor.
- **Complete two-layer DAG.** The nodes are partitioned into two layers, with every node in the first layer being a root and every possible edge directed from the first layer to the second layer. We take both layers to have order m, up to constant factors. Hence each node in the second layer has order m ancestors, while root nodes have only themselves as ancestors.

The resulting quantities and rates are summarized below.

DAG structure	κ_m	$\bar{a}_m$	Average MSE bound
m isolated roots	1	1	$\Delta_{n,m}^2 + v_n/n$
Chain of length m	m	$(m+1)/2$	$\Delta_{n,m}^2 + mv_n/n$
Rooted tree of depth D	$D+1$	$\leq D+1$	$\Delta_{n,m}^2 + Dv_n/n$
Star with $m-1$ roots and one sink	2	$2-1/m$	$\Delta_{n,m}^2 + v_n/n$
Complete two-layer DAG	2	$\asymp m$	$\Delta_{n,m}^2 + mv_n/n$

These examples show that the number of nodes m alone does not determine the estimation rate. For isolated roots and the star, $\bar{a}_m$ remains bounded, so the estimation term is of order v_n/n up to logarithmic factors. For a rooted tree, the dependence is controlled by its depth; in particular, a balanced tree has depth of order $\log m$. In contrast, both the chain and the complete two-layer DAG have $\bar{a}_m$ of order m, leading to an estimation term of order mv_n/n.

G COMPARISON WITH PRIOR THEORY

Table 4 summarizes recent theoretical results most closely related to our work. Existing results primarily study distributional regression for Euclidean responses, and only Padilla et al. (2025) provides an explicit convergence rate for DNN-based distributional regression. Moreover, Padilla et al. (2025) does not provide monotonicity estimation with respect to set inclusion. Methods such as Cevid et al. (2022) and Shen & Meinshausen (2025) may preserve monotonicity from an algorithmic perspective; however, rather than addressing general set-indexed conditional event probabilities, their theoretical convergence guarantees apply to particular distributional functionals (e.g., CDFs and quantiles), and these guarantees do not provide explicit convergence rates in terms of n. In contrast, our framework estimates set-indexed conditional event probabilities on a general response space $\mathcal{Y}$ that allows non-Euclidean responses, and enforces monotonicity with respect to set inclusion. Specifically, we provide explicit convergence rates for both the uniform MSE of the preliminary estimator and the aggregated MSE of the constrained estimator.

Table 4: Comparison of recent theoretical results in distributional regression. Here, $o_p(1)$ denotes convergence to zero in probability.

	Method	Assumptions / Target	Monotonicity	Convergence Rate
Prior Work	Kohler & Langer (2021) DNN non-parametric regression (not distributional)	$Y \in \mathbb{R}$ $f^*(x) = \mathbb{E}(Y\|X = x)$ $f^*(\cdot) \in \mathcal{H}(l, \mathcal{P})$	No	ϕ_n (MSE)
	Cevid et al. (2022) Forest-weighted distributional regression	$Y \in \mathbb{R}^u$ $F_A^*(x) = \mathbb{P}(Y \in A \mid X = x)$ $F_A^*(\cdot) \in \mathcal{H}(1, \{1, d_0\})$	Yes	$o_p(1)$
	Shen & Meinshausen (2025) Generative DNN-based sampling	$Y \in \mathbb{R}$ for theory; extends to $Y \in \mathbb{R}^u$ methodologically $F_A^*(x) = \mathbb{P}(Y \in A \mid X = x)$ $F_A^*(\cdot) \in \mathcal{H}(1, \{1, d_0\})$	Yes	$o_p(1)$
	Padilla et al. (2025) DNN-based distributional regression for scalar CDF	$Y \in \mathbb{R}$ $F_t^*(x) = \mathbb{P}(Y \le t \mid X = x)$ $F_t^*(\cdot) \in \mathcal{H}(l, \mathcal{P})$	No	$\frac{\log n}{n} + \phi_n \log^4 n$ (CRPS (Gneiting & Raftery, 2007), Worst-case MSE)
This Work	Preliminary estimator	$Y \in \mathcal{Y}$: general, non-Euclidean Λ: totally ordered, potentially uncountable $F_A^*(x) = \mathbb{P}(Y \in A \mid X = x)$, $A \in \Lambda$ $F_A^*(\cdot) \in \mathcal{H}(l_A, \mathcal{P}_A)$	No	$\phi_n \log^4 n$ (Uniform MSE over Λ)
	Constrained estimator	$Y \in \mathcal{Y}$: general, non-Euclidean $\Lambda = \{A_j\}_{j=1}^m \subset \mathcal{Y}$ with inclusion DAG $F_{A_j}^*(x) = \mathbb{P}(Y \in A_j \mid X = x)$ $F_A^*(\cdot) \in \mathcal{H}(l_A, \mathcal{P}_A)$	Yes	$\phi_n \log^4 n$ (Average MSE over Λ with fixed sets)

H Compute resources

All experiments are conducted on a single laptop equipped with an Intel Core Ultra 9 275HX CPU, 32GB of RAM, and an NVIDIA RTX 5070 Ti laptop GPU with 12GB of VRAM.

I Detailed configuration and results for numerical experiments

In this section, we provide the detailed settings for all eight scenarios, including the data-generating mechanisms and the design of the sets (Appendix I.1), full experimental results (Appendix I.2), and the configurations of the estimators used (Appendix I.3).

I.1 Detailed design for the eight scenarios

Here we present detailed design for the three scenarios (S1, S2, S3) mentioned in the main text and the additional five scenarios (S4, S5, S6, S7, S8).

I.1.1 Scenario 1 (Swiss roll manifold).

We consider $m = 20$ sets for a response $Y \in \mathbb{R}^3$ lying on a Swiss roll manifold. Let $X \sim \mathrm{Unif}([0,2]^{50})$ and define $X_{\mathrm{noisy}} = X + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2(X) I_{50})$ with heteroscedastic variance $\sigma^2(X) = 1 + 0.25\|X\|^2/50$. We generate two latent coordinates from X_{noisy} as

$$t = 1.5\pi + 3\pi\tau(w_t^\top X_{\mathrm{noisy}} + b_t), \qquad h = \tau(w_h^\top X_{\mathrm{noisy}} + b_h),$$

where $\tau(u) = (1 + e^{-u})^{-1}$, the entries of $w_t, w_h \in \mathbb{R}^{50}$ are drawn i.i.d. from $\mathcal{N}(0, 0.25)$, and $b_t, b_h \sim \mathcal{N}(0, 0.01)$. Thus, $t \in [1.5\pi, 4.5\pi]$ and $h \in [0, 1]$. The response is then obtained through the Swiss roll parametrization

$$Y = (t\cos t,\ h,\ t\sin t).$$

The $m = 20$ sets are defined in the (t, h) parameter space using two families of quantile thresholds:

$$q_t^{(l_t)} = Q^{(t)}_{\alpha_t^{(l_t)}}, \quad \alpha_t \in \{15\%, 30\%, 50\%, 70\%, 85\%\}, \quad l_t = 1, \ldots, 5,$$

$$q_h^{(l_h)} = Q^{(h)}_{\alpha_h^{(l_h)}}, \quad \alpha_h \in \{20\%, 40\%, 60\%, 80\%\}, \quad l_h = 1, \dots, 4,$$

where $Q_\alpha^{(\cdot)}$ denotes the α-th percentile of the respective marginal distribution, estimated from an independent sample of size 10,000. The 20 sets form a full 5×4 grid:

- $A_{(l_t,l_h)} = \{(t,h) : t \le q_t^{(l_t)},\, h \le q_h^{(l_h)}\}, \quad l_t \in \{1,\dots,5\}, \quad l_h \in \{1,\dots,4\}.$

The DAG structure is induced by componentwise ordering: $A_{(l_t,l_h)} \subset A_{(l_t',l_h')}$ if and only if $l_t \le l_t'$ and $l_h \le l_h'$.

I.1.2 S2 (Random dot product graph).

Let $Y \in \{0,1\}^{N\times N}$ be the adjacency matrix with $N = 50$, where each node i carries a unit-length embedding vector $z_i \in \mathbb{S}^{d-1}$ with $d = 5$, and the probability of an edge between nodes i and j is $\frac{1}{2}(\langle z_i, z_j\rangle + 1)$. The RDPG is endowed with a community structure of $K = 3$ communities, where both the community assignment probabilities $\pi_k(X)$ and the unit-length community embedding centers $a_k(X) \in \mathbb{S}^{d-1}$ are non-linearly parametrized by the covariate $X \sim \mathrm{Unif}([0,2]^{50})$ for $k \in \{1,2,3\}$.

Specifically, we use a two-layer MLP to map X to a vector $Q(X) \in \mathbb{R}^{18}$ with $Q(X) := W^{(2)} \tanh\big(W^{(1)}X\big)$, where $W^{(1)} \in \mathbb{R}^{16\times 50}$ and $W^{(2)} \in \mathbb{R}^{18\times 16}$ have entries independently drawn from $\mathcal{N}(0, 2/50)$ and $\mathcal{N}(0, 2/16)$, respectively. The output $Q(X)$ is split into two parts indexed by community $k \in \{1,2,3\}$. The first 15 components are reshaped into 3 vectors $\{\tilde a_k(X)\}_{k=1}^3 \in \mathbb{R}^5$, each normalized to obtain the community embedding center $a_k(X) = \tilde a_k(X)/\|\tilde a_k(X)\| \in \mathbb{S}^{d-1}$. The remaining 3 components, denoted $\{\ell_k(X)\}_{k=1}^3$, are treated as logits and passed through a softmax to obtain the community assignment probability $\pi_k(X) = \exp(\ell_k(X))/\sum_{k'=1}^3 \exp(\ell_{k'}(X))$.

Each node i is then independently assigned $c_i \sim \mathrm{Cat}(\pi(X))$, with embedding $z_i = (a_{c_i}(X) + \varepsilon_i)/\|a_{c_i}(X) + \varepsilon_i\|$ where $\varepsilon_i \sim \mathcal{N}(0, 0.3^2 I_d)$, yielding the random graph Y.

We estimate the distribution of the network statistic transitivity. For a graph Y with adjacency matrix $A \in \{0,1\}^{N\times N}$, the transitivity is defined as $T(Y) = \frac{3\cdot \mathrm{tri}(Y)}{\mathrm{tp}(Y)}$, where $\mathrm{tri}(Y) = \sum_{i<j<k} A_{ij}A_{jk}A_{ik}$ counts the number of triangles in Y, and $\mathrm{tp}(Y) = \sum_i \binom{d_i}{2}$ with $d_i = \sum_{j\ne i} A_{ij}$ counts the number of length-two paths of node i.

The $m = 10$ sets form a chain $A_1 \subset A_2 \subset \cdots \subset A_{10}$, defined as sublevel sets of the transitivity:

$$A_j = \big\{Y : T(Y) \le q_{\alpha_j}\big\}, \qquad q_{\alpha_j} = Q_{\alpha_j}\big(T(Y)\big), \qquad j = 1, \dots, 10,$$

where Q_α denotes the α-th percentile of the marginal distribution of $T(Y)$, and $\alpha_j = 5 + 10(j-1)$ for $j = 1, \dots, 10$ (i.e. evenly spaced from 5% to 95%).

I.1.3 S3 (Pixel density estimation on single noisy natural image).

We consider $X \in [0,1]^2$ a pixel coordinate and $Y = (Y_R, Y_G, Y_B) \in [0,1]^3$ a noisy observation of the clean pixel value $J(X) = (J_R, J_G, J_B)$ at X, with the goal of estimating $Y \mid X$ for a single given noisy image. Since we use 96×96 images from STL-10 (see Figure I.1.3 for illustration), X takes values on the discrete grid $\{k/96 : k = 1, \dots, 96\}^2$. The noise follows a Poisson-Gaussian model: $Y_c = \mathrm{Poisson}(50 \cdot J_c(X))/50 + \varepsilon_c$ with $\varepsilon_c \sim \mathcal{N}(0, 0.05^2)$ i.i.d. across $c \in \{R, G, B\}$.

The $m = 13$ sets are upper-orthants in $\mathbb{R}^3$, where each channel admits three nested intervals:

$$L_1^{(c)} = (-\infty, Q_{30}^{(c)}], \quad L_2^{(c)} = (-\infty, Q_{60}^{(c)}], \quad L_3^{(c)} = (-\infty, Q_{90}^{(c)}],$$

with $Q_\alpha^{(c)}$ denoting the α-th percentile of the marginal distribution of Y_c, estimated from 50000 noisy samples. The 13 sets are organized as follows:

- $A_1 = L_1^{(R)} \times L_1^{(G)} \times L_1^{(B)}$,
- $A_2 = L_2^{(R)} \times L_1^{(G)} \times L_1^{(B)}, \quad A_3 = L_1^{(R)} \times L_2^{(G)} \times L_1^{(B)}, \quad A_4 = L_1^{(R)} \times L_1^{(G)} \times L_2^{(B)},$
- $A_5 = L_2^{(R)} \times L_2^{(G)} \times L_1^{(B)}, \quad A_6 = L_2^{(R)} \times L_1^{(G)} \times L_2^{(B)}, \quad A_7 = L_1^{(R)} \times L_2^{(G)} \times L_2^{(B)},$

- $A_8 = L_2^{(R)} \times L_2^{(G)} \times L_2^{(B)}$,
- $A_9 = L_3^{(R)} \times L_2^{(G)} \times L_2^{(B)}$, $A_{10} = L_2^{(R)} \times L_3^{(G)} \times L_2^{(B)}$, $A_{11} = L_2^{(R)} \times L_2^{(G)} \times L_3^{(B)}$,
- $A_{12} = L_3^{(R)} \times L_3^{(G)} \times L_2^{(B)}$,
- $A_{13} = L_3^{(R)} \times L_3^{(G)} \times L_3^{(B)}$.

The DAG structure is induced by componentwise inclusion: $A_i \subset A_j$ if and only if $L^{(c)}_{\ell_c^{(i)}} \subseteq L^{(c)}_{\ell_c^{(j)}}$ for all $c \in \{R, G, B\}$. Also, since the per-pixel data are abundant, we follow Foi et al. (2008) and use the heteroskedastic Gaussian approximation to obtain the true conditional probability in closed form:

$$\mathbb{P}(Y \in A_j \mid X) \approx \prod_{c \in \{R,G,B\}} \Phi\left(\frac{Q^{(c)}_{\alpha_c^{(j)}} - J_c(X)}{\sqrt{J_c(X)/\lambda + \sigma^2}}\right),$$

where $\alpha_c^{(j)} \in \{30, 60, 90\}$ is the percentile level associated with channel c in A_j.



Figure 1: Example noiseless images from the STL-10 dataset across 10 categories.

I.1.4 S4 (CHAIN DAG).

We consider $m = 10$ nested sets in a chain structure $A_1 \subset A_2 \subset \cdots \subset A_{10}$, representing the simplest DAG topology. Covariates are $X \sim \mathrm{Unif}([0, 2]^{10})$. The response $Y \in \mathbb{R}^2$ follows a heteroscedastic Gaussian model $Y \mid X \sim \mathcal{N}(\mu(X), \sigma^2(X) I_2)$, where $\mu(X) = \Theta_1 X + \Theta_2 X^{\odot 2}$ with Θ_1 and Θ_2 having entries drawn i.i.d. from $\mathrm{Unif}([-1, 1])$ and $\mathrm{Unif}([-0.5, 0.5])$, respectively, and $\sigma^2(X) = 1 + 0.25\|X\|^2$.

The $m = 10$ sets are Euclidean balls $A_j = \{y : \|y\|^2 \le r_j^2\}$ with radii given by the 10th through 95th percentiles of $\|Y\|^2$. Specifically, the sets are defined as

$$A_j = \{y \in \mathbb{R}^2 : \|y\|^2 \le r_j^2\}, \quad j = 1, \ldots, 10,$$

with radii $r_1 < r_2 < \cdots < r_{10}$ defined by

$$r_j = \sqrt{q_{\alpha_j}}, \quad q_{\alpha_j} = Q_{\alpha_j}(\|Y\|^2),$$

where Q_α denotes the α-th percentile of the marginal distribution of $\|Y\|^2$, and $\alpha_j = 10 + \frac{85}{9}(j - 1)$ for $j = 1, \ldots, 10$ (i.e. evenly spaced from 10% to 95%).

I.1.5 S5 (GENERAL DAG).

We use the same covariate and response distributions as Scenario 4, but consider $m = 7$ sets with a richer DAG structure involving multiple roots and shared descendants. Sets are axis-aligned rectangles $A_j = \{y : y_k \in [q_{k,l_j}, q_{k,u_j}],\ k = 1, 2\}$, where $q_{k,\alpha}$ are the α-quantile of Y_k.

The $m = 7$ sets are axis-aligned rectangles $A_j = I_1^{(j)} \times I_2^{(j)}$, where $Q_\alpha^{(k)}$ denotes the α-th percentile of the marginal distribution of Y_k. The 7 sets take:

- $A_1 = [Q_{30}^{(1)}, Q_{70}^{(1)}] \times [Q_{30}^{(2)}, Q_{70}^{(2)}]$,
- $A_2 = [Q_{10}^{(1)}, Q_{30}^{(1)}] \times [Q_{25}^{(2)}, Q_{75}^{(2)}]$,
- $A_3 = [Q_{25}^{(1)}, Q_{75}^{(1)}] \times [Q_{25}^{(2)}, Q_{75}^{(2)}]$,
- $A_4 = [Q_{70}^{(1)}, Q_{90}^{(1)}] \times [Q_{25}^{(2)}, Q_{75}^{(2)}]$,
- $A_5 = [Q_{10}^{(1)}, Q_{75}^{(1)}] \times [Q_{10}^{(2)}, Q_{90}^{(2)}]$,
- $A_6 = [Q_{25}^{(1)}, Q_{90}^{(1)}] \times [Q_{10}^{(2)}, Q_{90}^{(2)}]$,
- $A_7 = [Q_{5}^{(1)}, Q_{95}^{(1)}] \times [Q_{5}^{(2)}, Q_{95}^{(2)}]$.

The DAG structure follows from set inclusion and is illustrated in Figure I.1.5.

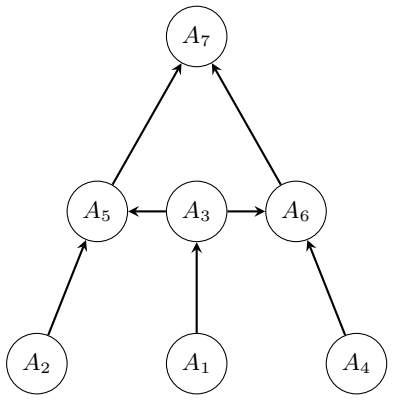


Figure 2: DAG structure for S5. An arrow from A_i to A_j indicates $A_i \subset A_j$.

I.1.6 S6 (GAUSSIAN MIXTURE).

Consider the covariates follow $X \sim \mathrm{Unif}([0,2]^{30})$. We generate $Y \in \mathbb{R}^5$ from a mixture $Y \mid X \sim \sum_{k=1}^{4} \pi_k(X)\mathcal{N}(\mu_k(X),\, \sigma_k^2(X)\, I_5)$, where component means follow $\mu_k(X) = \Theta_{k,1}X + \Theta_{k,2}X^{\odot 2}$ with $\Theta_{k,1} \in \mathbb{R}^{5\times 30}$ having entries from $\mathrm{Unif}([-1,1])$ and $\Theta_{k,2} \in \mathbb{R}^{5\times 30}$ from $\mathrm{Unif}([-0.5,0.5])$, and $X^{\odot 2}$ denotes element-wise square of X. The mixing weights are $\pi(X) = \mathrm{softmax}(W_\pi X + b_\pi)$ with $W_\pi \in \mathbb{R}^{4\times 30}$. The variance is heteroscedastic: $\sigma_k^2(X) = \sigma_{k,0}^2 + \sigma_{k,1}^2\|\mu_k(X)\|^2$, where $\sigma_{k,0} \sim \mathrm{Unif}([1,2])$ and $\sigma_{k,1} \sim \mathrm{Unif}([0.3,0.5])$.

The $m = 19$ sets are axis-aligned hyperrectangles $A_j = \prod_{k=1}^{5} I_k^{(j)}$, where each interval $I_k^{(j)}$ is drawn from four nested levels:

$$L_1^{(k)} = [Q_{20}^{(k)}, Q_{80}^{(k)}], \quad L_2^{(k)} = [Q_{15}^{(k)}, Q_{85}^{(k)}], \quad L_3^{(k)} = [Q_{10}^{(k)}, Q_{90}^{(k)}], \quad L_4^{(k)} = [Q_{5}^{(k)}, Q_{95}^{(k)}],$$

where $Q_\alpha^{(k)}$ denotes the α-th percentile of the marginal distribution of Y_k, estimated from an independent sample of size 10,000. The 19 sets are organized as follows:

- $A_1 = \prod_{k=1}^{5} L_1^{(k)}$,
- $A_{i+1} = L_2^{(i)} \times \prod_{k\neq i} L_1^{(k)}, \quad i = 1,\dots,5$,
- $A_{i+6} = L_3^{(i)} \times \prod_{k\neq i} L_1^{(k)}, \quad i = 1,\dots,5$,
- $A_{i+11} = L_3^{(i)} \times L_3^{(i \bmod 5+1)} \times \prod_{k\neq i, i \bmod 5+1} L_1^{(k)}, \quad i = 1,\dots,5$,
- $A_{17} = \prod_{k=1}^{5} L_2^{(k)}$,
- $A_{18} = \prod_{k=1}^{5} L_3^{(k)}$,
- $A_{19} = \prod_{k=1}^{5} L_4^{(k)}$.

The DAG structure follows from set inclusion and is illustrated in Figure 3.

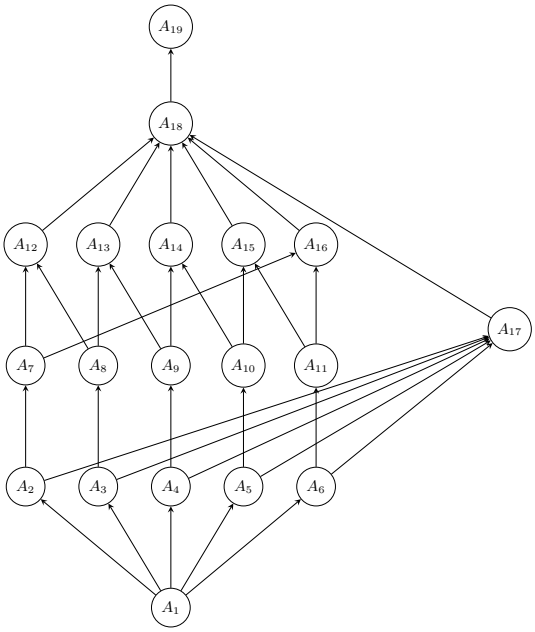


Figure 3: DAG structure for S1. An arrow from A_i to A_j indicates $A_i \subset A_j$.

I.1.7 Scenario 7 (Functional DAG).

We consider a functional-valued response $Y : [0,1] \to \mathbb{R}$ with covariates $X \sim \mathrm{Unif}([-1,1]^{10})$. The response is generated through the finite Fourier expansion

$$Y(t) = \sum_{k=1}^{6} \xi_k \varphi_k(t), \qquad t \in [0,1],$$

where $(\varphi_1, \ldots, \varphi_6) = \big(1, \sqrt{2}\sin(2\pi t), \sqrt{2}\cos(2\pi t), \sqrt{2}\sin(4\pi t), \sqrt{2}\cos(4\pi t), \sqrt{2}\sin(6\pi t)\big)$. Conditionally on $X = x$, the coefficient vector $\xi = (\xi_1, \ldots, \xi_6)^\top$ follows

$$\xi \mid X = x \sim \mathcal{N}_6\big(0, \Sigma(x)\big), \qquad \Sigma(x) = D(x) R D(x),$$

where $R_{k\ell} = 0.4^{|k-\ell|}$ and $D(x) = \mathrm{diag}\{s_1(x), \ldots, s_6(x)\}$. Let $b = (1, 0.9, 0.7, 0.6, 0.5, 0.4)$. The scale functions are defined by

$$u(x) = 1.2\sin(\pi x_1 x_2) + 0.8(x_3^2 - 1/3) + 0.6x_4,$$

and

$$s_k(x) = b_k\{0.4 + 0.8\tau(u(x))\}\{0.8 + 0.4\tau(1.5x_{k+4})\}, \qquad k = 1, \ldots, 6,$$

where $\tau(x) = 1/(1 + e^{-x})$. Thus, although the conditional mean of Y is zero, its conditional distribution varies nonlinearly with X through the covariance structure. To construct the target sets, we define two functionals

$$D_1(Y) = \int_0^1 |Y(t)|\, dt, \qquad D_\infty(Y) = \sup_{t \in [0,1]} |Y(t)|.$$

Let $Q_\alpha^{(D_1)}$ and $Q_\alpha^{(D_\infty)}$ denote the α-th percentiles of the marginal distributions of $D_1(Y)$ and $D_\infty(Y)$, respectively. We use the quantile levels

$$(\alpha_1^{(1)}, \alpha_1^{(2)}, \alpha_1^{(3)}) = (\alpha_\infty^{(1)}, \alpha_\infty^{(2)}, \alpha_\infty^{(3)}) = (40\%, 60\%, 80\%),$$

and define

$$q_1^{(l_1)} = Q^{(D_1)}_{\alpha_1^{(l_1)}}, \qquad q_\infty^{(l_\infty)} = Q^{(D_\infty)}_{\alpha_\infty^{(l_\infty)}}, \qquad l_1, l_\infty \in \{1,2,3\}.$$

We consider $m = 9$ sets,

$$A_{(l_1,l_\infty)} = \left\{ Y : D_1(Y) \le q_1^{(l_1)}, \quad D_\infty(Y) \le q_\infty^{(l_\infty)} \right\}, \qquad l_1, l_\infty \in \{1,2,3\}.$$

The DAG structure is then induced by componentwise ordering: $A_{(l_1,l_\infty)} \subseteq A_{(l_1',l_\infty')}$ if $l_1 \le l_1'$ and $l_\infty \le l_\infty'$. Thus, the DAG is a full 3×3 grid, induced by the componentwise ordering of the two quantile indices, with edges directed from each set to its immediate supersets.

I.1.8 Scenario 8 (Network balls).

We consider an additional experiment for network responses. Each response is the adjacency matrix $Y \in \{0,1\}^{50\times 50}$ of a 50-node RDPG generated using the same mechanism as in S2. For each replication, we generate a reference network $Y_0 \in \{0,1\}^{50\times 50}$ once from the same model with fixed input covariates $x_0 = (1,\dots,1) \in [0,2]^{50}$. Define the Frobenius distance

$$D(Y, Y_0) = \|Y - Y_0\|_F.$$

For aligned undirected binary networks, $D(Y,Y_0)^2$ is proportional to the fraction of edges on which Y and Y_0 disagree. We define ten nested network balls centered at Y_0:

$$A_j = \{Y : D(Y,Y_0) \le q_{\alpha_j}\}, \qquad j = 1,\dots,10,$$

where $q_{\alpha_j} = Q_{\alpha_j}(D(Y,Y_0))$ is the α_j-th percentile of the marginal distribution of $D(Y,Y_0)$, with $\alpha_j = 5\% + 10\%(j-1)$. We refer to this setting as network balls.

I.2 Full experiment results

We present the full experimental results for all scenarios in this section. Consistent with the conclusions in the main text, **Constrain-NN** achieves the best performance in most cases, while **Pre-NN** also converges and surpasses **DRF** and **Engression** as the sample n grows.

Table 5: Mean Squared Error across all simulation scenarios. Best result in each row is bolded.

Scenario	n_{train}	Constrain-NN	Pre-NN	Pre-NN-Clip	DRF	Engression
S1 (Swiss roll)	1000	**0.0096 (0.0007)**	0.0307 (0.0015)	0.0249 (0.0016)	0.0128 (0.0013)	0.0493 (0.0255)
	2000	**0.0072 (0.0004)**	0.0191 (0.0010)	0.0156 (0.0011)	0.0122 (0.0012)	0.0466 (0.0244)
	4000	**0.0057 (0.0004)**	0.0104 (0.0005)	0.0088 (0.0005)	0.0116 (0.0011)	0.0436 (0.0239)
S2 (RDPG)	1000	**0.0348 (0.0011)**	0.0464 (0.0010)	0.0439 (0.0010)	0.0713 (0.0146)	0.0420 (0.0135)
	2000	**0.0305 (0.0007)**	0.0372 (0.0007)	0.0355 (0.0007)	0.0694 (0.0143)	0.0395 (0.0134)
	4000	**0.0281 (0.0006)**	0.0309 (0.0006)	0.0299 (0.0006)	0.0630 (0.0121)	0.0383 (0.0132)
S4 (Chain)	1000	0.0061 (0.0009)	0.0095 (0.0011)	0.0091 (0.0011)	0.0128 (0.0046)	**0.0055 (0.0018)**
	2000	**0.0043 (0.0003)**	0.0056 (0.0004)	0.0055 (0.0004)	0.0108 (0.0038)	0.0049 (0.0015)
	4000	**0.0035 (0.0003)**	0.0038 (0.0002)	0.0037 (0.0002)	0.0089 (0.0031)	0.0047 (0.0015)
S5 (General DAG)	1000	**0.0055 (0.0009)**	0.0094 (0.0008)	0.0092 (0.0008)	0.0072 (0.0023)	**0.0055 (0.0019)**
	2000	**0.0036 (0.0004)**	0.0058 (0.0004)	0.0058 (0.0004)	0.0064 (0.0020)	0.0052 (0.0018)
	4000	**0.0026 (0.0004)**	0.0037 (0.0003)	0.0037 (0.0003)	0.0055 (0.0018)	0.0049 (0.0016)
S6 (Gaussian mixture)	1000	**0.0100 (0.0023)**	0.0145 (0.0025)	0.0134 (0.0023)	0.0118 (0.0028)	0.0114 (0.0037)
	2000	**0.0085 (0.0023)**	0.0105 (0.0022)	0.0101 (0.0022)	0.0114 (0.0027)	0.0108 (0.0037)
	4000	**0.0077 (0.0022)**	0.0082 (0.0022)	0.0081 (0.0022)	0.0109 (0.0027)	0.0102 (0.0036)
S7 (Functional DAG)	1000	0.0085 (0.0007)	0.0173 (0.0021)	0.0156 (0.0026)	**0.0075 (0.0006)**	0.0675 (0.0243)
	2000	**0.0053 (0.0004)**	0.0090 (0.0010)	0.0083 (0.0011)	0.0068 (0.0002)	0.0633 (0.0185)
	4000	**0.0047 (0.0005)**	0.0053 (0.0004)	0.0050 (0.0004)	0.0058 (0.0006)	0.0623 (0.0182)
S8 (Network balls)	1000	**0.0434 (0.0022)**	0.0645 (0.0018)	0.0612 (0.0025)	0.0691 (0.0011)	0.0463 (0.0009)
	2000	**0.0382 (0.0021)**	0.0519 (0.0018)	0.0492 (0.0017)	0.0668 (0.0013)	0.0438 (0.0011)
	4000	**0.0362 (0.0015)**	0.0417 (0.0015)	0.0402 (0.0013)	0.0644 (0.0009)	0.0428 (0.0010)

Table 6: Test MSE ($\times 10^2$) for S3 by object class. Standard deviations are shown in parentheses. Results are averaged over 30 images per class.

Class	Pre-NN	Constrain-NN	Pre-NN-Clip	DRF	Engression
Airplane	2.68 (0.95)	**2.62 (0.93)**	2.66 (0.95)	3.26 (1.07)	6.09 (2.21)
Bird	3.75 (1.34)	**3.52 (1.12)**	3.73 (1.32)	4.68 (1.53)	7.91 (2.29)
Car	3.70 (1.12)	**3.45 (1.05)**	3.69 (1.11)	4.24 (1.22)	8.91 (2.36)
Cat	3.70 (1.60)	**3.50 (1.62)**	3.69 (1.60)	4.41 (1.75)	7.47 (2.27)
Deer	4.17 (1.18)	**3.99 (1.14)**	4.15 (1.18)	4.87 (1.21)	7.97 (1.84)
Dog	3.72 (1.43)	**3.56 (1.32)**	3.71 (1.43)	4.55 (1.58)	7.92 (2.56)
Horse	4.50 (1.20)	**4.10 (1.09)**	4.49 (1.19)	5.37 (1.52)	9.23 (2.39)
Monkey	3.42 (1.11)	**3.18 (1.10)**	3.41 (1.12)	4.34 (1.25)	8.12 (2.06)
Ship	2.83 (0.81)	**2.35 (0.87)**	2.86 (0.82)	2.73 (0.98)	7.08 (1.69)
Truck	3.79 (1.12)	**3.38 (0.99)**	3.76 (1.10)	4.14 (1.36)	8.92 (2.44)
Overall	3.63 (1.29)	**3.37 (1.24)**	3.61 (1.29)	4.26 (1.53)	7.96 (2.37)

I.3 Configuration for estimators

In each simulation scenario except the special scenario using real images, each neural network for **Pre-NN** and **Constrain-NN** is a four-layer MLP with 64 ReLU units per layer, trained for 200 epochs using Adam with initial learning rate 10^{-3} and a cosine annealing schedule decaying to 1% of the initial learning rate. For the special scenario, in order to model the more complex data, we use a five-layer MLP with 256 ReLU units per layer, trained for 500 epochs with the same optimizer and schedule. For **Engression**, we use the same architecture as the corresponding **Pre-NN/Constrain-NN** and draw 500 samples per query point to estimate set probabilities with the `Engression` R package. For **DRF**, we fit 2000 trees using the default hyperparameters of the `drf` R package, with the same covariate inputs as the neural network methods.

J Real data experiments

In this section, we provide additional details and complete results for the two real-data experiments. Appendix J.1 covers the Beijing Multi-Site Air Quality experiment, including target set construction, input processing, estimator configurations, and per-station results, while Appendix J.2 reports the Chicago 2015 Crime experiment in full.

J.1 Real data: Beijing multi-Site air quality prediction

In this section, we provide the experimental details omitted from the main text. In Subsection J.1.1, we provide a summary of the response variables (the pollutants) and describe the construction of the target sets. In Subsection J.1.2, we list the variables contained in the dataset and describe how the model inputs are processed. In Subsection J.1.3, we provide the configurations of all estimators. In Subsection J.1.4, we provide full results and detailed results broken down by monitoring station.

J.1.1 Set construction

Here we describe the construction of the target sets. Specifically, an initial set of candidates is defined based on the regulatory document, and the collection is then pruned with several steps according to the empirical data distribution to retain only the meaningful sets.

Pollutant grading and empirical distribution. The Chinese Ambient Air Quality Standards (HJ 633-2012) regulate each pollutant under multiple threshold schemes. The starting point of our threshold design draws from this document, adopting the grading scheme that corresponds to the strictest available setting per pollutant; this choice is without loss of generality in comparing our experimental estimators in a distributional regression regime. Table 7 summarizes the original grade breakpoints we use and their empirical prevalence across all 420768 hourly observations in the dataset.

Grade merging. After merging sparse grades into the next lower grade from the top down (i.e., if the highest remaining grade has prevalence $< 2\%$, it is merged with the grade immediately below,

Table 7: AQI grade breakpoints we used and empirical prevalence (%) for each pollutant. All concentration units in μg/m^3. For each pollutant, the first row ("upper") gives the upper concentration boundary of each grade, and the second row ("%") gives the empirical prevalence of that grade in the dataset. The eight columns correspond to the eight pollution grades: Excellent, Good, Light pollution, Moderate pollution, Heavy pollution, Severe pollution, Hazardous, and Beyond Index.

	Excellent	Good	Light	Moderate	Heavy	Severe	Hazardous	Beyond
PM2.5 upper	35	75	115	150	250	350	500	∞
PM2.5 %	37.34	23.58	15.48	8.39	10.74	3.21	1.04	0.23
PM10 upper	50	150	250	350	420	500	600	∞
PM10 %	33.87	42.66	16.41	4.88	1.15	0.54	0.29	0.19
SO_2 upper	50	150	475	800	1600	2100	2620	∞
SO_2 %	92.81	6.99	0.20	0.00	0.00	0.00	0.00	0.00
NO_2 upper	40	80	180	280	565	750	940	∞
NO_2 %	46.99	34.07	18.59	0.35	0.00	0.00	0.00	0.00
CO upper	2000	4000	14000	24000	36000	48000	60000	∞
CO %	84.29	12.32	3.39	0.00	0.00	0.00	0.00	0.00
O_3 upper	160	200	300	400	800	1000	1200	∞
O_3 %	93.40	3.38	3.06	0.14	0.01	0.00	0.00	0.00

repeating until the top grade exceeds 2%), and removing the excellent/good boundary to focus on pollution events, we retain thresholds from the light pollution grade upward. For SO_2, although exceedances of the light pollution threshold ($SO_2 \geq 150$ μg/m^3) occur in only 0.20% of observations, we retain this pollutant to ensure all six dimensions are represented in the joint exceedance structure. The resulting thresholds used to define exceedance sets are summarized in Table 8.

Table 8: Exceedance thresholds after grade merging. Each value defines an exceedance event $\{\text{pollutant} \geq \text{value}\}$; the value 0 indicates the dimension is unconstrained.

Pollutant	Thresholds (μg/m^3)
PM2.5	0, 75, 115, 150, 250
PM10	0, 150, 250, 350
SO_2	0, 150
NO_2	0, 80
CO	0, 4,000
O_3	0, 200

Set construction and pruning. For each pollutant $k \in \{1, \ldots, 6\}$, let $\mathcal{T}_k = \{t_0^{(k)}, t_1^{(k)}, \ldots, t_{L_k}^{(k)}\}$ denote the set of available thresholds, with $t_0^{(k)} = 0$ indicating no constraint. Each exceedance set is then specified by a level tuple $(\ell_1, \ldots, \ell_6)$ with $\ell_k \in \{0, 1, \ldots, L_k\}$, defining

$$A_{(\ell_1,\ldots,\ell_6)} = \left\{ y \in \mathbb{R}^6 : y_k \geq t_{\ell_k}^{(k)} \text{ for all } k = 1, \ldots, 6 \right\}.$$

A level $\ell_k = 0$ imposes no constraint on pollutant k. Taking the full Cartesian product of all per-pollutant level sets yields $5 \times 4 \times 2 \times 2 \times 2 \times 2 = 320$ candidate sets, which we reduce in 2 steps:

1. **Prevalence filtering.** We retain only those sets whose empirical prevalence—the fraction of observations satisfying the joint exceedance condition—is at least 0.5%. This reduces the 320 candidates to 86.
2. **Gap pruning.** We construct DAG of the retained sets under the componentwise partial ordering: $A_{(\ell_1,\ldots,\ell_6)} \subseteq A_{(\ell'_1,\ldots,\ell'_6)}$ if and only if $\ell_k \geq \ell'_k$ for all k (higher levels correspond to smaller sets). Starting from the deepest level of the diagram, a set is removed if its empirical prevalence differs from that of its closest parent by less than 0.001 percentage points, indicating that the two sets are nearly indistinguishable. This step removes 9 sets.

Finally, we exclude the unconstrained full set ($\ell_k = 0$ for all k), whose probability is trivially 1, leaving 76 nested sets. Table 9 reports the ten sets with the highest and the ten sets with the lowest empirical prevalences.

Table 9: Exceedance sets with the ten highest and ten lowest empirical prevalences (after pruning). The full collection contains 76 sets.

Rank	Exceedance condition	Prevalence (%)
Highest prevalence		
1	$\mathrm{PM}_{2.5} \geq 75$	39.43
2	$\mathrm{PM}_{2.5} \geq 115$	23.80
3	$\mathrm{PM}_{10} \geq 150$	23.79
4	$\mathrm{PM}_{2.5} \geq 75 \ \wedge \ \mathrm{PM}_{10} \geq 150$	21.71
5	$\mathrm{NO}_2 \geq 80$	19.53
6	$\mathrm{PM}_{2.5} \geq 115 \ \wedge \ \mathrm{PM}_{10} \geq 150$	18.73
7	$\mathrm{PM}_{2.5} \geq 75 \ \wedge \ \mathrm{NO}_2 \geq 80$	16.16
8	$\mathrm{PM}_{2.5} \geq 150$	15.34
9	$\mathrm{PM}_{2.5} \geq 150 \ \wedge \ \mathrm{PM}_{10} \geq 150$	14.91
10	$\mathrm{PM}_{10} \geq 150 \ \wedge \ \mathrm{NO}_2 \geq 80$	13.07
Lowest prevalence		
67	$\mathrm{PM}_{2.5} \geq 250 \ \wedge \ \mathrm{PM}_{10} \geq 350$	1.74
68	$\mathrm{PM}_{10} \geq 350 \ \wedge \ \mathrm{NO}_2 \geq 80$	1.73
69	$\mathrm{PM}_{2.5} \geq 75 \ \wedge \ \mathrm{PM}_{10} \geq 350 \ \wedge \ \mathrm{NO}_2 \geq 80$	1.73
70	$\mathrm{PM}_{10} \geq 350 \ \wedge \ \mathrm{NO}_2 \geq 80 \ \wedge \ \mathrm{CO} \geq 4000$	1.15
71	$\mathrm{PM}_{2.5} \geq 250 \ \wedge \ \mathrm{PM}_{10} \geq 350 \ \wedge \ \mathrm{CO} \geq 4000$	1.14
72	$\mathrm{PM}_{2.5} \geq 250 \ \wedge \ \mathrm{PM}_{10} \geq 350 \ \wedge \ \mathrm{NO}_2 \geq 80 \ \wedge \ \mathrm{CO} \geq 4000$	1.13
73	$\mathrm{PM}_{2.5} \geq 115 \ \wedge \ \mathrm{O}_3 \geq 200$	0.92
74	$\mathrm{PM}_{10} \geq 150 \ \wedge \ \mathrm{O}_3 \geq 200$	0.91
75	$\mathrm{PM}_{2.5} \geq 75 \ \wedge \ \mathrm{PM}_{10} \geq 150 \ \wedge \ \mathrm{O}_3 \geq 200$	0.84
76	$\mathrm{PM}_{2.5} \geq 115 \ \wedge \ \mathrm{PM}_{10} \geq 150 \ \wedge \ \mathrm{O}_3 \geq 200$	0.62

J.1.2 Dataset variables and input processing

The Beijing Multi-Site Air Quality dataset contains hourly observations from 12 monitoring stations across Beijing, covering March 2013 to February 2017. For each station and each timestamp, the dataset records six pollutant concentrations ($\mathrm{PM}_{2.5}, \mathrm{PM}_{10}, \mathrm{SO}_2, \mathrm{NO}_2, \mathrm{CO}, \mathrm{O}_3$), five meteorological variables (temperature, pressure, dew point temperature, wind speed, precipitation), and the wind direction.

We construct the model inputs from these raw variables in three steps. First, the hour-of-day, day-of-week, and month-of-year are each mapped to a $(\sin, \cos)$ pair to preserve cyclic continuity, and the 16-direction wind label is similarly converted to its angle (e.g., N $\to 0°$, NNE $\to 22.5°$, etc.) and then to $(\sin, \cos)$. Second, given a forecast horizon h and an input window length W, we collect the measurements at lags $\{h, h+1, \ldots, h+W-1\}$ relative to the target time; for each lag, we include all five meteorological variables, the two wind-direction components, and the six pollutants, yielding 13 variables per lag. Third, rows containing any missing pollutant value at the target timestamp or any missing variable in the input window are removed.

For each target time s, the input is the concatenation of the cyclical time encodings at $s - h$ (the earliest end of the input window) with the W lagged variable blocks aggregated over the input window.

For each monitoring station, we construct the training and test sets separately within each year of the 4-year study period. In each year, 3 consecutive months are held out for testing, and all observations within the held out period are assigned to the test set, with no intervening observations included in the training set. The remaining 9 months are then used for training, yielding an overall 75/25 train/test split.

J.1.3 ESTIMATOR CONFIGURATIONS

Each neural network instance consists of a three-layer MLP with ReLU activations. Hidden layer width is set to 64 for all instances. All input features are standardized to zero mean and unit variance using training set statistics. Models are trained for 40 epochs with the Adam optimizer at a learning rate of 3×10^{-4} and a batch size of 100. For Distributional Random Forests, we fit 2000 trees per station using the default hyperparameters provided by the `drf` R package, with the same covariate inputs as the neural network methods.

J.1.4 DETAILED RESULTS

In this subsection, we report both the aggregated test MSE with standard deviation and the per-station test MSE for the five estimators: **Pre-NN**, **Constrain-NN**, **Pre-NN-Clip**, **DRF**, and **Engression**. Results are presented across all $3 \times 3 = 9$ configurations of the forecast horizon h and input window length W. Table 10 reports the aggregated results. Table 11 presents the results for all 12 monitoring stations, with each cell reporting the average MSE over the 76 target sets. Consistent with the results in the main text, **Constrain-NN** achieves the lowest MSE in the vast majority of cells across all stations, and the other findings reported in the main text also hold at the per-station level.

Table 10: Test MSE ($\times 10^2$) across 76 sets and aggregated over the 12 monitoring stations. Each entry reports mean, (std) over stations. The best performance is bolded.

Horizon h	Method	$W = 3$	$W = 6$	$W = 9$
$h = 3$	Pre-NN	3.43 (0.68)	3.41 (0.67)	3.35 (0.66)
	Constrain-NN	**2.28 (0.48)**	**2.24 (0.46)**	**2.23 (0.46)**
	Pre-NN-Clip	3.54 (0.71)	3.51 (0.70)	3.46 (0.69)
	DRF	2.29 (0.48)	2.33 (0.50)	2.36 (0.50)
	Engression	3.79 (0.79)	3.94 (0.87)	3.92 (0.87)
$h = 6$	Pre-NN	3.78 (0.74)	3.71 (0.74)	3.65 (0.74)
	Constrain-NN	**2.84 (0.59)**	**2.81 (0.57)**	**2.79 (0.57)**
	Pre-NN-Clip	3.88 (0.74)	3.79 (0.77)	3.73 (0.77)
	DRF	2.90 (0.62)	2.91 (0.62)	2.92 (0.62)
	Engression	4.12 (0.90)	4.24 (0.95)	4.18 (0.94)
$h = 9$	Pre-NN	3.97 (0.74)	3.87 (0.76)	3.79 (0.77)
	Constrain-NN	**3.15 (0.61)**	**3.10 (0.60)**	**3.09 (0.61)**
	Pre-NN-Clip	4.06 (0.75)	3.95 (0.76)	3.87 (0.79)
	DRF	3.21 (0.67)	3.20 (0.67)	3.17 (0.66)
	Engression	4.31 (0.94)	4.40 (0.99)	4.30 (0.96)

Table 11: Test MSE ($\times 10^2$) across 76 sets, reported per station and per (horizon h, input window W) configuration. Each cell reports results for **Pre-NN / Constrain-NN / Pre-NN-Clip / DRF / Engression**. The best performance in each cell is bolded.

Station	Horizon h	Input Window W (hours)		
		$W = 3$	$W = 6$	$W = 9$
Aotizhongxin	$h = 3$	3.86 / **2.72** / 3.99 / **2.72** / 4.36	3.90 / **2.73** / 4.02 / 2.83 / 4.69	3.86 / **2.71** / 3.99 / 2.88 / 4.70
	$h = 6$	4.29 / **3.30** / 4.38 / 3.44 / 4.78	4.27 / **3.37** / 4.37 / 3.48 / 5.08	4.25 / **3.39** / 4.36 / 3.52 / 5.06
	$h = 9$	4.46 / **3.62** / 4.54 / 3.75 / 5.01	4.44 / **3.65** / 4.52 / 3.79 / 5.27	4.38 / **3.75** / 4.47 / 3.78 / 5.16
Changping	$h = 3$	3.06 / **2.03** / 3.16 / 2.04 / 3.25	3.00 / **1.98** / 3.10 / 2.04 / 3.22	2.91 / **1.92** / 2.99 / 2.03 / 3.15
	$h = 6$	3.38 / **2.48** / 3.48 / 2.50 / 3.51	3.23 / **2.41** / 3.32 / 2.48 / 3.46	3.12 / **2.37** / 3.19 / 2.43 / 3.37
	$h = 9$	3.61 / 2.77 / 3.69 / **2.75** / 3.64	3.40 / **2.65** / 3.46 / 2.67 / 3.55	3.28 / **2.59** / 3.31 / 2.62 / 3.49
Dingling	$h = 3$	2.18 / 1.52 / 2.23 / **1.42** / 2.37	2.19 / 1.55 / 2.23 / **1.46** / 2.47	2.18 / 1.55 / 2.23 / **1.50** / 2.47
	$h = 6$	2.38 / 1.82 / 2.47 / **1.77** / 2.49	2.35 / 1.84 / 2.38 / **1.80** / 2.58	2.34 / 1.84 / 2.38 / **1.83** / 2.59
	$h = 9$	2.57 / 2.04 / 2.63 / **2.01** / 2.62	2.50 / **2.03** / 2.58 / **2.03** / 2.70	2.49 / 2.09 / 2.52 / **2.06** / 2.75
Dongsi	$h = 3$	3.56 / **2.43** / 3.70 / 2.50 / 3.88	3.54 / **2.39** / 3.69 / 2.52 / 4.13	3.44 / **2.40** / 3.62 / 2.53 / 4.02
	$h = 6$	3.94 / **3.08** / 4.04 / 3.14 / 4.27	3.88 / **2.99** / 4.00 / 3.13 / 4.43	3.77 / **2.89** / 3.91 / 3.09 / 4.25
	$h = 9$	4.12 / **3.35** / 4.21 / 3.44 / 4.49	4.01 / **3.17** / 4.12 / 3.40 / 4.57	3.92 / **3.09** / 4.05 / 3.34 / 4.38
Guanyuan	$h = 3$	3.53 / 2.49 / 3.65 / **2.42** / 3.88	3.53 / **2.40** / 3.65 / 2.47 / 4.13	3.47 / **2.37** / 3.61 / 2.50 / 4.13
	$h = 6$	3.94 / **3.06** / 4.04 / 3.10 / 4.32	3.90 / **3.04** / 4.00 / 3.12 / 4.52	3.83 / **3.00** / 3.94 / 3.10 / 4.44
	$h = 9$	4.11 / **3.33** / 4.19 / 3.44 / 4.54	4.03 / **3.30** / 4.12 / 3.40 / 4.65	3.94 / **3.25** / 4.04 / 3.34 / 4.50
Gucheng	$h = 3$	4.00 / **2.75** / 4.13 / 2.78 / 4.55	3.98 / **2.70** / 4.10 / 2.84 / 4.81	3.89 / **2.66** / 4.02 / 2.85 / 4.73
	$h = 6$	4.46 / **3.40** / 4.55 / 3.44 / 5.08	4.29 / **3.28** / 4.39 / 3.45 / 5.13	4.21 / **3.22** / 4.31 / 3.51 / 5.06
	$h = 9$	4.61 / **3.75** / 4.73 / 3.79 / 5.20	4.47 / **3.64** / 4.55 / 3.80 / 5.32	4.34 / **3.59** / 4.43 / 3.80 / 5.17
Huairou	$h = 3$	2.10 / **1.32** / 2.13 / 1.36 / 2.29	2.09 / **1.34** / 2.13 / 1.37 / 2.38	2.09 / **1.32** / 2.13 / 1.43 / 2.36
	$h = 6$	2.37 / **1.71** / 2.44 / 1.72 / 2.43	2.24 / **1.73** / 2.27 / 1.76 / 2.51	2.21 / **1.72** / 2.24 / 1.79 / 2.47
	$h = 9$	2.50 / **1.96** / 2.57 / **1.96** / 2.54	2.36 / **1.97** / 2.44 / **1.97** / 2.56	2.26 / 2.00 / 2.29 / **1.95** / 2.49
Nongzhanguan	$h = 3$	4.17 / **2.78** / 4.30 / 2.80 / 4.71	4.13 / **2.68** / 4.25 / 2.85 / 5.00	4.07 / **2.67** / 4.20 / 2.89 / 4.95
	$h = 6$	4.59 / **3.53** / 4.70 / 3.61 / 5.14	4.54 / **3.45** / 4.63 / 3.66 / 5.39	4.46 / **3.44** / 4.57 / 3.66 / 5.25
	$h = 9$	4.81 / **3.86** / 4.89 / 4.07 / 5.40	4.71 / **3.76** / 4.79 / 4.04 / 5.60	4.68 / **3.81** / 4.77 / 4.02 / 5.44
Shunyi	$h = 3$	3.39 / **1.91** / 3.47 / 2.09 / 3.56	3.24 / **1.87** / 3.33 / 2.12 / 3.54	3.16 / **1.81** / 3.19 / 2.11 / 3.46
	$h = 6$	3.72 / **2.55** / 3.78 / 2.72 / 3.95	3.62 / **2.48** / 3.68 / 2.71 / 3.93	3.44 / **2.53** / 3.52 / 2.73 / 3.79
	$h = 9$	3.95 / **3.00** / 4.05 / 3.08 / 4.19	3.75 / **2.93** / 3.81 / 3.05 / 4.12	3.58 / **2.84** / 3.63 / 2.97 / 3.90
Tiantan	$h = 3$	3.70 / 2.33 / 3.78 / **2.32** / 4.12	3.66 / **2.33** / 3.75 / 2.36 / 4.08	3.58 / **2.33** / 3.68 / 2.38 / 4.10
	$h = 6$	4.00 / **2.93** / 4.12 / 2.96 / 4.35	3.92 / **2.89** / 3.99 / 2.98 / 4.36	3.86 / **2.87** / 3.94 / 2.97 / 4.36
	$h = 9$	4.16 / **3.26** / 4.26 / 3.27 / 4.54	4.12 / **3.15** / 4.23 / 3.25 / 4.59	4.09 / **3.18** / 4.21 / 3.26 / 4.59
Wanliu	$h = 3$	3.65 / 2.53 / 3.77 / **2.46** / 4.05	3.66 / 2.51 / 3.78 / **2.50** / 4.29	3.68 / **2.55** / 3.82 / 2.56 / 4.42
	$h = 6$	4.00 / 3.06 / 4.09 / **3.04** / 4.39	4.04 / 3.12 / 4.14 / **3.11** / 4.65	4.06 / 3.18 / 4.16 / **3.15** / 4.73
	$h = 9$	4.24 / 3.41 / 4.31 / **3.39** / 4.64	4.23 / 3.46 / 4.32 / **3.41** / 4.88	4.20 / 3.45 / 4.29 / **3.39** / 4.87
Wanshouxigong	$h = 3$	4.01 / **2.56** / 4.13 / 2.59 / 4.41	3.96 / **2.47** / 4.08 / 2.61 / 4.58	3.90 / **2.44** / 4.03 / 2.63 / 4.53
	$h = 6$	4.33 / **3.21** / 4.43 / 3.32 / 4.76	4.25 / **3.10** / 4.34 / 3.28 / 4.85	4.20 / **3.08** / 4.30 / 3.27 / 4.76
	$h = 9$	4.54 / **3.50** / 4.62 / 3.62 / 4.97	4.41 / **3.47** / 4.49 / 3.57 / 5.03	4.38 / **3.53** / 4.47 / 3.56 / 4.91

J.2 Real data: Chicago 2015 crime distribution estimation

We further evaluate our framework on the 2015 Chicago Crime dataset (publicly available at `https://data.gov/open-gov/`), following an experimental setup similar to that of Tansey et al. (2018); Padilla et al. (2025) but extended to a multivariate response. Specifically, whereas Padilla et al. (2025) models the conditional CDF of total log-crime counts through univariate distributional regression, we model the joint distribution of crime composition across the most frequent categories. This category-level formulation is important for downstream decision making, since regions with similar total crime volume may differ substantially in their dominant crime types and therefore require different policing interventions.

Following Tansey et al. (2018), the city is discretized into a 100×100 grid based on latitude and longitude, and grid cells with no recorded crimes are excluded, yielding $n = 3844$ cells. The covariate X is two dimensional and records the longitude bin index and latitude bin index of each cell, so the regression is purely spatial in our setting. The response variable is four dimensional, consisting of the log-transformed crime counts of the four most frequent crime categories in the dataset, namely theft, battery, criminal damage, and narcotics.

We construct a finite collection of nested target sets indexed by quantile exceedances of the four crime categories. For each crime category $k \in \{1, \dots, 4\}$, let $\mathcal{T}_k = \{t_0^{(k)}, t_1^{(k)}, t_2^{(k)}, t_3^{(k)}\}$ denote the four marginal quantile thresholds at levels $20\%, 40\%, 60\%, 80\%$ of the empirical response within that category, where $t_\ell^{(k)}$ is the ℓ-th threshold for category k and increases in ℓ. Each candidate exceedance set is specified by a level tuple $(\ell_1, \dots, \ell_4)$ with $\ell_k \in \{0, 1, 2, 3\}$ encoding the threshold level applied to each category, defining $A_{(\ell_1,\dots,\ell_4)} = \left\{y \in \mathbb{R}^4 : y_k \geq t_{\ell_k}^{(k)} \text{ for all } k = 1, \dots, 4\right\}$. The full Cartesian product yields 256 candidate sets, which we reduce by the same pruning procedure as in Appendix J.1.1. Specifically, we retain only sets whose empirical prevalence is at least 1%, then construct the DAG of the retained sets under the partial ordering $A_{(\ell_1,\dots,\ell_4)} \subseteq A_{(\ell_1',\dots,\ell_4')}$ if and only if $\ell_k \geq \ell_k'$ for all k. Starting from the deepest level, we remove any set whose empirical prevalence differs from that of its closest parent by less than 1 percentage point. This step removes 174 sets, leaving $m = 82$ nested sets whose inclusion relations form a non-trivial DAG.

We use an experimental configuration similar to that in Padilla et al. (2025). Specifically, each neural network instance in **Pre-NN** and **Constrain-NN** is a five-layer MLP with ReLU activations and hidden width 256, trained for 1000 epochs with the Adam optimizer at a learning rate of 10^{-3} annealing to 10^{-6}. For **Engression**, we use the same model architecture and draw 500 samples per query point for distribution estimation. For **DRF**, we fit 2000 trees with the same covariate inputs as the neural network methods. We conduct 30 Monte Carlo rounds of the experiment, with each replication using an independent random $75/25$ train–test split.

Table 12: Test MSE across $m = 82$ sets, averaged over 30 random splits. Standard deviations in parentheses.

	Pre-NN	Constrain-NN	Pre-NN-Clip	DRF	Engression
MSE	0.1194 (0.0046)	**0.1186 (0.0044)**	0.1248 (0.0046)	0.1200 (0.0032)	0.1408 (0.0039)

Table 12 reports the test MSE averaged over the $m = 82$ sets and 30 replications. **Constrain-NN** achieves the best performance with the lowest MSE, followed by **Pre-NN** and DRF, which perform comparably, while **Pre-NN-Clip** performs worse than **Pre-NN**. These results demonstrate the importance of imposing the structural constraint, as done in **Constrain-NN**.

J.3 Practical guidance for constructing a set family

The target set family should be selected according to the conditional probabilities relevant to the scientific problem. We suggest the following general procedure.

1. **Use domain-relevant events when they are available.** If the application provides meaningful thresholds or downstream decisions, the sets can be defined directly from them. For example, the Beijing air-quality experiment uses joint pollutant exceedance events based on official grading thresholds.

2. **Use measurable summaries or metrics for complex responses.** A response object can be mapped to one or more interpretable scalar summaries, with quantiles of their marginal distributions used as thresholds. For a network response, examples include network statistics and distances to a reference network. For a functional response, examples include L^1, L^2, and L^∞ distances to a reference function. Sublevel sets of a single summary give a nested family, while intersections constructed from several summaries can produce a general DAG.
3. **Remove empirically redundant sets when necessary.** If an initial family is large, sets having nearly the same empirical prevalence as their closest supersets provide little additional distributional information and may be removed. After a node is removed, its adjacent inclusion relations can be reconnected and transitively redundant edges deleted, preserving the inclusion order among the retained sets.

The estimation procedure only requires membership indicators for the selected sets and the DAG encoding their inclusion relations. It therefore applies in the same way whether the sets arise from scientific knowledge, response summaries, or metrics.